TITLE: Human cellular homeostasis buffers *trans*-acting translational effects of heterologous gene expression with very different codon usage bias

SHORT-TITLE: Codon usage bias and trans-acting translational effects

AUTHORS: Arthur J Jallet[1], Antonin Demange[1], Fiona Leblay[1], Mathilde Decourcelle[2], Khadija El Koulali[2], Marion AL Picard[1], Ignacio G Bravo[1,*]

[1]Virostyle team, Laboratory MIVEGEC (CNRS, IRD, Univ Montpellier), Montpellier, France

[2]BioCampus Montpellier (Univ Montpellier, CNRS, INSERM), Montpellier, France

***Correspondence should be adressed to: ajallet@unistra.fr, ignacio.bravo@cnrs.fr**

Abstract

The frequency of synonymous codons in protein coding genes is non-random and varies both between species and between genes within species. Whether this codon usage bias (CUBias) reflects underlying neutral mutational processes or is instead shaped by selection remains an open debate, especially regarding the role of selection for enhanced protein production. Variation in CUBias of a gene (be it natural synonymous mutations or biotechnological synonymous recoding) can have an enormous impact on its expression by diverse *cis*- acting mechanisms. But expression of genes with extreme CUBias can also lead to strong phenotypic effects by altering the overall intracellular translation homeostasis *via* competition for ribosomal machinery or tRNA depletion. In this study, we expressed at high levels in human cells six different synonymous versions of a gene and used matched transcriptomic and proteomic data to evaluate the impact of CUBias of the heterologous gene on the translation of cellular transcripts. Our experimental design focused specifically on differences during translation elongation. Response to expression of the different synonymous sequences was assessed by various approaches, ranging from analyses performed on a per-gene basis to more integrated approaches of the cell as a whole. We observe that the transcriptome displayed substantial changes as a result of heterologous gene expression by triggering an intense antiviral and inflammatory response, but that changes in the proteomes were very modest. Most importantly we notice that changes in translation efficiency of cellular transcripts were not associated with the direction of the CUBias of the heterologous sequences, thereby providing only limited support for *trans*- acting effects of synonymous changes. We interpret that, in human cells in culture, changes in CUBias can lead to important *cis*- acting effects in gene expression, but that cellular homeostasis can buffer the phenotypic impact of overexpression of heterologous genes with extreme CUBias.

Keywords: codon usage bias, gene expression, competition for translational resources, cellular homeostasis, heterologous expression

Introduction

The genetic code allows to translate information stored in gene nucleotide sequences into protein sequences. This code is redundant because 61 codons encode the standard twenty amino acids. Codons encoding the same amino acid are called synonymous codons and amino acids either can be encoded by one, two, three, four or six synonymous codons. Moreover, these synonymous codons are usually not used at random. Accounting for the occurrence of a given amino acid in the coding part of a genome, some synonymous codons are indeed over-represented in the genome (i.e., more frequently used to encode the considered amino acid compared to a uniform situation) while some others are under-represented (i.e., less frequently used). This non-random usage of synonymous codons is referred to as codon usage bias (CUBias). CUBias widely varies across species (**Ikemura, 1982; Kanaya et al, 1999**; **Novoa et al, 2019**) as well as between genes within the same genome (**Gouy & Gautier, 1982**; **Sharp & Li, 1986**; **Duret, 2002**).

There are mainly two non-mutually exclusive hypotheses to explain the existence of CUBias and its intra and inter-specific variation (**Hershberg & Petrov, 2008 ; Plotkin & Kudla, 2011**). The first one is the mutational bias hypothesis, which is usually referred to as a neutral explanation as it does not involve any selective process shaping CUBias. According to this view, the CUBias of a coding sequence simply arises because of local biases in the mutation spectrum during DNA replication or repair and hence mirrors the nucleotide composition of non-coding sequences nearby, because neutral mutational biases are expected to be similar for coding and non-coding parts of the genome. This idea of CUBias being a side-effect of mutational processes is supported by several lines of evidence. For example, the GC composition of exons is usually similar to that of the introns in the same gene **Chamary, Parmley & Hurst, 2006**) or to non-coding sequence in its vicinity (**Chen et al, 2004**). The fact that for many genes exons and introns tend to display similar nucleotide composition illustrates that mapping or not onto a translated region is not the major determinant of the mutational bias. At broader scale, evidence of such co-variation of nucleotide composition between coding and non-coding regions is obvious in many vertebrate's genomes and manifests as the so-called isochores. Isochores are long chromosome stretches enriched in AT or in GC nucleotides, observable by means of cytochemical staining during metaphase (**Caspersson et al, 1968**). This strong compositional bias over large chromosome stretches is maintained over evolution, so that the physical mapping of a gene onto a given isochore is the most important determinant of nucleotide composition and therefore of CUBias (**Holmquist, 1989**; **Duret, 2002**; **Pouyet et al, 2017**). Finally, local CUBias in vertebrate genomes is further influenced by the distance to homologous recombination hotspots and could reflect the intensity of the GC-biased gene conversion around the considered loci (**Pouyet et al, 2017**).

In the second hypothesis, known as translational selection, synonymous mutations are considered subject to selection as they can provide phenotypic variation and ultimately differences in fitness by means of variation in protein amount and quality during gene expression, mostly during translation. According to this view, selection can thus discriminate synonymous codons either because of their different decoding speed, their different consequences on co-translational folding of the translated protein (an effect probably mediated by local decoding speed itself – **Pechmann & Fryman, 2013; Yu et al, 2015**) or the different accuracy for codon-anticodon decoding (**Stoletzki & Eyre-Walker, 2007; Drummond & Wilke, 2009; Walsh et al, 2020; Drummond & Wilke, 2008**). Experimental supports for all these mechanisms have been observed, altogether supporting that selection can, at least in part, contribute to shape the CUBias of a gene. In eukaryotes, synonymous mutations can be further associated to a phenotype without direct consequences on translation, for instance when modifying splicing regulatory elements located in exons (**Chamary, Parmley & Hurst, 2006**; **Savisaar & Hurst, 2018**). The proposed effect of CUBias on decoding speed and translational efficiency is strongly supported by the fact that genes preferentially using synonymous codons that are the most frequent genome-wide are usually those expressed at higher levels and decoded by more abundant transfer RNAs (tRNAs). This association between codon frequency and tRNA abundance has been extensively reported in multiple fast growing, unicellular organisms such as *E. coli* (**Ikemura, 1981; Sharp & Li, 1986; Tuller et al, 2010**), or *S. cerevisiae* (**Ikemura, 1982, Akashi, 2003**, **Tuller et al, 2010**), but also in multicellular organisms such as *C. elegans* (**Duret, 2000**) and *D. melanogaster* (**Duret & Mouchiroud, 1999**) and provide support of CUBias being shaped by selection to optimize protein production in these species. Moreover, technical advances such as ribosome profiling, either alone or in association with elongation inhibitors, seem to confirm that synonymous mutations play a role in how efficiently genes are translated, and that this differential efficiency is directly linked to tRNA availability (**Hia & Takeuchi, 2020**, but see **Charneski & Hurst, 2013**). In yeast for example, a positive correlation between translation efficiency and the match between codon frequency and tRNA availability has been reported (**Riba et al, 2019**). Other studies indeed showed that the decoding time of a codon during elongation is anti-correlated with the availability of its cognate tRNA (**Gardin et al, 2014**; **Weinberg et al, 2016**). While supported in the above-cited model species, translational selection in mammalian genomes is a much more debated topic (**Urrutia & Hurst, Genetics, 2001**) and if exists, its influence relatively to mutational processes seems to be much weaker (**Kanaya et al, 2001; Vogel et al, 2010**). In humans more specifically, small effective populations sizes render difficult selection on synonymous mutations that presumably display small selective coefficients in general (**Hershberg & Petrov, 2008**) and the importance of GC-biased gene conversion and isochore compartmentation on shaping the genome composition is probably strong enough to blur most of the signatures left within coding sequences by selection (**Pouyet et al, 2017**). Moreover, as for other multicellular organisms, most genes are

expressed at different levels across human tissues, thereby rendering the link between expression level and CUBias difficult to make. There is indeed an open debate in the field on whether tissue-specific gene expression patterns include or not a CUBias, which could be related or not to different tRNA availability in different tissues (**Plotkin, Robins & Levine, 2004**; **Dittmar, Goodenbour & Pan, 2006; Eraslan et al, 2019**). Complementary hypotheses posit that CUBias in eukaryotes could be the result of adaptation to varying tRNAs levels during the cell cycle (**Frenkel-Morgenstern, 2012**), or that CUBias could be constrained by the cellular processes in which the different genes are involved, such as proliferation and differentiation (**Gingold et al, 2014**), although this interpretation has been questioned regarding the direct role played by translational selection in shaping the distinct CUBias of genes involved in these two antagonistic processes (**Pouyet et al, 2017**). Finally, some studies revealed that synonymous sites do not evolve neutrally in mammals but instead have functional impacts, though not necessarily through their translational effects (**Newman et al, 2016; Chamary & Hurst, 2005; Kudla et al, 2006**).

The translational selection hypothesis proposes that the CUBias of a gene can modulate its own translation efficiency, by means of *cis*-effects. But it has been proposed that the CUBias of a gene can exert effects on the expression and translation efficiency of other genes, by means of *trans*-effects. Indeed, translation is the most expensive step of the gene expression process (**Lynch & Marinov, 2015**) and the availability of the translational machinery - itself also complex and costly - is the overall limitation for protein synthesis, as 95% of the cellular ribosomes are actively engaged in translation at any time point (**Princiotta et al, 2003**). The different cellular transcripts face thus a direct competition for finite translational resources (*e.g.* ATP, GTP, amino acids, ribosomes, tRNAs - **Li et al, 2014**). Selection has thus resulted in large differences between cellular transcripts in their ribosomal binding ability, so that not necessarily the most abundant mRNAs are those that are more active at recruiting ribosomes and starting translation, resulting in ribosomal sequestering and loss of translation opportunity for other mRNAs (**Callens et al, 2021**). Further, when abundant and actively ribosome-recruiting mRNA species perform poorly at translation elongation -because of a poor CUBias, for instance–, ribosomal pausing leads to accumulation of slowly proceeding ribosomes, amplifying ribosome sequestration and further reducing the pool of free ribosomes available for the translation of other mRNAs (**Pelechano, Wei & Steinmetz, 2015; Shah et al, 2013**). Since these *trans* effects can hamper translation of essential genes, it is conceivable that genes that require high expression levels have been selected to be encoded with a particular CUBias, to decrease the burden caused onto other genes arising from their high expression levels, or to avoid suffering from *trans* acting effects caused by high expression level of other genes. Two studies in *E. coli* strongly support this hypothesis of *trans*-acting effects mediated by competition between mRNAs. First, expressing different synonymous versions of the green fluorescent protein (GFP) surprisingly did not result in changes in its own

translation (*i.e.,* no *cis*- effects), but instead led to different *trans*- effects on global translation efficiency (**Kudla et al, 2009**). Second, it was shown that the global negative effects associated to translating highly expressed genes that use rare codons can be balanced if the availability of the cognate tRNAs was increased to meet the demand (**Frumkin et al, 2018**). Results also consistent with such *trans*- effects have been communicated for yeast (**Pop et al, 2014**). Finally, from a practical perspective, competition between mRNAs to access shared resources such as tRNAs can be considered under a perspective of cellular economy, using a supply-and-demand reasoning, where the pool of transcripts exerts a demand for being translated, requiring their codons to be decoded by the corresponding cognate anticodons, representing the "offer" (**Gingold & Pilpel, 2011; Gingold, Dahan & Pilpel, 2012**). A proxy for the overall demand in tRNAs could be computed using codon composition of mRNAs present in the cell and by accounting for their relative abundance in the cell transcriptome (**Gingold, Dahan & Pilpel, 2012; Schmitt et al, 2014**).

In this study we aim at studying *trans*-acting effects by analyzing the different cellular changes incurred as a function of CUBias of heterologous genes. We combine transcriptomic and proteomic data to identify patterns consistent with *trans*- effects on translation efficiency mediated by competition for the translation machinery availability, using human HEK293 cells as model system. We have analysed such effects by studying both individual gene translation efficiency and overall translation efficiency at the whole-cell scale. Globally our results show that heterologous gene overexpression triggers changes in the cellular transcriptome of large extent. In contrast, changes in the cellular proteome are more discrete and do not show any evident global trend with regards to CUBias of the heterologous genes. Our results suggest that, in our experimental conditions, cellular homeostasis can largely buffer the effects of gene overexpression, and provide only limited support for the hypothesis of *trans*-acting arising from directional competition for limited translational resources, associated to CUBias.

Results

Heterologous gene expression upon transfection leads to substantial changes in the cellular transcriptome, independently of the codon usage bias of the heterologous genes

A principal component analysis performed on our transcriptomic data (exclusively using cellular mRNAs, *i.e.*, heterologous mRNAs excluded) revealed that the main source of variation across our samples stems from transfection itself (Fig. S1). The first principal component captures 67% of the total transcriptomic variance, with Mock samples displaying values divergent from all other versions, while the second principal component captures 18% of the total transcriptomic variance and is strongly associated to differences between the three experimental batches (Fig. S1). Indeed, variation along the first component is largely explained by variation in the total amount of heterologous transcripts in the sample (Spearman correlation coefficient: 0.87, P = 4.8e-7, Fig. S2). Since heterologous expression is the largest determinant of transcriptomic changes, we focused on the identification of differentially expressed (DiffExp) cellular transcripts compared to the Mock for each transfected plasmid version. We state first that transfection with our "Empty" control, *i.e.* a plasmid that encodes for the *neoR* and for the *egfp* genes, leads to identification of 711 DiffExp transcripts. The number of DiffExp transcripts detected in the experimental conditions, *i.e.,* cells transfected with plasmids encoding for the *neoR* gene and for different synonymous versions of the *shble_egfp* gene, varies between 505 for Shble#6 to 1,312 for Shble#3 (**Fig. 1A)**. Interestingly, a vast majority (~ 90%) of DiffExp transcripts were up-regulated (Table S1). As noted above, the largest determinant of transcriptomic changes is the intensity of heterologous expression: variation in the number of DiffExp transcripts detected is largely explained by the variation in the amount of heterologous transcripts (Spearman correlation coefficient: 0.54; **Fig. 1A**). Considering the response to transfection and heterologous *egfp* expression alone, we observed that most of the 711 mRNAs identified as DiffExp in the Empty version were also identified as DiffExp in all six (63%) or in five out of the six (81%) *shble* synonymous versions (**Fig. 1B**). We hence interpret our sets of DiffExp genes (listed in Table S1) constitute a fundamental part of the cellular response to transfection and heterologous gene expression and not to a specific *shble* synonymous version. The precise overlap between DiffExp mRNAs in the Empty and in each of the six *shble* versions is given in Fig S3. Finally, and independently of their behaviour in the Empty condition, we observe that a vast majority of transcripts DiffExp compared to the Mock in each version are shared among *shble* synonymous versions, despite their very different CUBias (**Fig. 1C** and Fig. S4). For each set of DiffExp mRNAs with respect to the mock control condition we performed a functional enrichment analysis (see Methods). All seven DiffExp gene sets shared the same top three enriched categories, namely *Inflammatory response*, *Type I interferon signaling pathway* and *Response to virus*. Other categories

appeared always over-represented among the DiffExp in all gene sets: *Negative regulation of viral genome replication* and *Immune response* categories (Table S2). Of note, an overwhelming majority of mRNAs belonging to these five categories appeared to be up-regulated in comparison to the Mock, suggesting transfection triggered an inflammatory response in our HEK293 cells (Table S2).

Overall, the global trends for our transcriptomic results are that: i) transfected cells undergo substantial changes in the transcriptome, largely through gene upregulation; ii) the extent of these changes correlates with the extent of heterologous gene expression; iii) transcriptomic changes upon transfection largely overlap the cellular response to viral infection; and iv) the bulk of the transcriptomic response are shared among conditions, irrespective of the CUBias of the transfected *shble* version.

Lack of directional changes in the cellular proteome upon transfection and heterologous gene expression, irrespective of the codon usage bias of the heterologous genes

We next used proteomic data obtained on samples matching transcriptomic data to investigate whether we could find similar evidence for a shared response across versions or alternatively if a *shble* synonymous version-specific response is observed after transfection at the protein level compared to the transcript level. Variation along the first two axes of a principal component analysis (Fig. S5) succeed at capturing an important fraction of the global variation in the proteome (77%) among samples. However, in striking contrast to what we observed from transcriptomic data, the experimental conditions did not spread following any evident pattern. Especially, variation along the first component was not driven by variation in the total amount of heterologous transcripts expressed by the samples (Spearman correlation coefficient = 0.099; P = 0.67, Fig. S5). Differential expression analysis failed to identify differentially expressed proteins, either with regards to transfection itself or regarding the different *shble* synonymous versions in the transfected plasmids. We tried to refine these raw analyses by defining a set of proteins that could maximize the chances of detecting DiffExp proteins between conditions. We identified thus all proteins that were detected in all three replicates of a same condition and that were not detected in any of the three replicates of at least one other condition (see Methods and Table S3), reasoning that if an effect is to be found, working specifically on this subset of proteins would maximize variation across versions. We found 369 proteins that fulfilled this criterion, but again a principal component analysis failed at identifying systematic differences between conditions (Fig. S6). Further, we did not identify any enrichment in functional categories in this 369-protein set. Overall, we concluded that in our experimental setup, the cellular proteomic response to heterologous expression was much less important than the transcriptomic one (see discussion for potential explanations).

High translation level of heterologous sequences does not impair translation of cellular genes with similar codon usage bias

Our experimental setup of paired transcriptomic and proteomic data recovered from the same samples allowed us to estimate intra-sample covariation between mRNA and protein levels, measured in TPM and riBAQ, respectively. On average across samples, variation in mRNA levels accounts for around one-third of the variation in protein levels (median=0.31, min=0.29, max=0.35, Table S4). These values agree with previous measurements performed on mammalian cells, with studies reporting a R^2 of 0.41 for mouse cells (**Schawanhaüsser et al, 2011**) and usually between 0.30 and 0.40 in human cells (reviewed in **Vogel & Marcotte, 2012**). A representative example is provided in Fig. S7 for a sample in condition Shble#1. We concluded that our data fit well within the classical range of mRNA-protein covariation expected for human cells in culture.

For prokaryotic and unicellular eukaryotic systems overexpressing genes with extreme CUBias it has been proposed that overexpressing heterologous can affect the translation efficiency of other genes (**Kudla et al, 2009; Frumkin et al, 2018; Shah et al, 2013**). To test whether this is also the case in our human cultured cells system, we chose as a proxy for translation efficiency of a given gene the ratio protein-level-over-transcript-level (expressed as riBAQ/TPM ratio). We studied first the impact of heterologous expression on the translation efficiency of cellular genes by comparing the Empty condition to the Mock condition. Cells under the Empty condition were transfected with a plasmid encoding only the *egfp* gene, whose sequence, "enhanced" for expression in human cells, is strongly biased towards the use of the most frequent codons in the human genome. This first level of analysis should hence provide insights into potential consequences of over-expressing an "over-humanized" gene on translation efficiency of cellular genes. Changes in the riBAQ/TPM ratio in Empty *vs.* Mock samples calculated for 2,471 cellular genes and plotted as a function of the CUBias match between the corresponding gene and the average human genome are depicted in **Fig. 2A**. Our results show that the intensity of changes in riBAQ/TPM for cellular genes when EGFP is actively translated is not a function of gene's CUBias (slope = 0.00378, F-test P = 0.29). Specifically, we did not observe that genes enriched in the most-frequent codons in the human genome – *i.e.*, those used to encode *egfp*, with COUSIN values above one – suffer the most. This lack of directional impact in our system does not support the hypothesis that translation of highly expressed over-humanized genes impairs translation of cellular genes that also use these human frequent codons. We performed the same analysis comparing riBAQ/TPM ratios in the Mock to the ones for Shble#1 and Shble#2 conditions, as these two versions encode *shble* using an over-humanized CUBias, in the same direction than egfp (respective COUSIN values to the human genome for egfp, Shble#1 and Shble#2: 3.38, 3.47 and 3.42, respectively). We observed similar results than for the Empty condition (**Fig. 2B,** slope = 0.0048, F-test P = 0.18 and **Fig.**

2C, slope = -0.0059, F-test P = 0.11), further supporting the absence of competition exerted by translation of highly expressed "overhumanized" genes on the translation efficiency of host genes, irrespective of the CUBias of the cellular genes.

To extend the analysis of the impact of heterologous gene expression on the translation of cellular genes to the range of CUBias explored in our setup, we aimed at identifying cellular genes that presented consistent trends of changes in riBAQ/TPM ratio *across* conditions. As a proxy for translation level of the heterologous genes we used the total amount of EGFP and SHBLE detected in each sample, using the sum of the corresponding riBAQ values (Fig S8). We stratified the experimental conditions in two sets: those that expressed heterologous genes enriched in codons frequently used in the human genome (*i.e.,* Empty, Shble#1 and Shble#2) and those that expressed heterologous genes enriched in codons rarely used in the human genome (*i.e.,* Shble#3, Shble#4, Shble#5 and Shble#6). We focused first on the conditions using human-frequent codons. For a total of 2,550 cellular genes, we could perform linear regressions exploring correlation between variation in riBAQ/TPM ratios for each individual gene and the total amount of heterologous proteins in the corresponding sample. This analysis used twelve underlying data points per gene as explanatory variable: three values for the mock control samples and nine values for the transfected samples. Among these genes, 235 displayed a significant variation (assessed by the significance of the regression slope) in their riBAQ/TPM ratios as a function of heterologous protein expression: 109 genes showed a positive, significant co-variation with heterologous proteins levels and 126 showed a negative, significant co-variation (**Fig. 3A**, yellow and green sets, respectively). A representative example of a gene with a negative association (green set) is given in **Fig. 3B** (left, KIF11). We next compared these two sets of genes in terms of CUBias, reasoning that a competition for translational resources between cellular and heterologous transcripts should primarily negatively affect translation of genes enriched in most human-frequent codons (*i.e.,* in the same codons used by *egfp* and Shble#1 and Shble#2 versions). Instead, we found that the median COUSIN score for the two gene sets were not significantly different (0.58 for the set of 126 genes; 0.45 for the set of 109 genes; Wilcoxon Mann-Whitney test P = 0.58, **Fig. 3C,** left). An additional Anderson-Darling test failed to reject the null hypothesis that the COUSIN value distributions for each dataset could have been drawn from a same underlying distribution and were thus not significantly different (P=0.33, **Fig. 3D,** left). Furthermore, these two gene sets, with translation positively or negatively affected by overexpression of heterologous genes enriched in human-frequent codons, do not display different distribution of COUSIN values than the overall cellular genes (Fig. S9). Our results confirm the trend communicated above and show that cellular genes whose translation was positively or negatively impacted by overexpression of heterologous genes enriched in human-frequent codons do not display different CUBias. Hence, these results reinforce the view that in our experimental system we do not to recover the expected pattern consistent with competition for translational resources. Instead, high

expression heterologous genes enriched in human-frequent codons does not impose a substantial differential burden on host translation efficiency of cellular genes as a function of their CUBias.

It is conceivable that we failed to identify an impact of over-humanized protein expression on host translation due to a pool of tRNAs for the considered codons large enough to accommodate the demand exerted by translation of heterologous mRNAs. Alternatively, we can expect the pool of tRNAs decoding human rare codons to be more limiting and therefore perhaps more prone to be depleted by the demand imposed by the translation of heterologous mRNAs rich in these rare codons. We performed then a similar analysis focusing on the conditions expression heterologous genes enriched in codons that are underrepresented in the human genome. For 2,580 cellular genes we performed linear regressions of their riBAQ/TPM ratios as a function of the total amount of heterologous proteins. This analysis used 15 underlying data points per gene as explanatory variable: three values for the mock control samples and twelve values for the transfected samples (*i.e.* Shble#3, Shble#4, Shble#5 and Shble#6) (**Fig. S8**). Among these genes, 285 displayed a significant variation (assessed by the significance of the regression slope) in their riBAQ/TPM ratios as a function of heterologous protein expression: 130 genes showed a positive, significant co-variation with heterologous proteins levels and 155 showed a negative, significant co-variation (**Fig. 3A**, pink and blue sets, respectively). A representative example of a gene with a negative association (blue set) is given in **Fig. 3B** (right, SNRPG). We compared then the CUBias of the genes in each of these two sets and we observed that they differ in their match to the average human CUBias: cellular genes showing a negative association with overexpression of heterologous genes enriched in human-rare codons display higher COUSIN values than genes displaying positive association (respective median values 0.79 and 0.13; Wilcoxon Mann-Whitney test P = 0.010, **Fig. 3C**, right). The distribution of COUSIN values between the two gene datasets is also significantly different (Anderson-Darling test, P = 0.0010, **Fig. 3D**, right). Furthermore, cellular genes negatively affected by overexpression of heterologous genes enriched in human-rare codons do not display different COUSIN values distribution than the ensemble of the cellular genes, while genes positively affected do display a different CUBias distribution that is shifted towards lower COUSIN values (Fig. S10). This result is counter-intuitive because – assuming cellular mRNAs compete with heterologous mRNAs to access to rare-codon decoding tRNAs – we would instead have expected cellular genes negatively impacted to be the ones preferentially using these non-optimal codons that are required for translating heterologous transcripts and hence to be those presenting lower COUSIN values. We thus examined whether the 155 genes – which surprisingly are not enriched in rare codons despite being impaired by heterologous expression of under-humanized versions – were more expressed. This could indeed help to understand this counter-intuitive pattern, with the underlying assumption that highly translated transcripts are more prone to potential shortage of tRNAs due to heterologous translation. We nevertheless observed that these 155 genes were on average less

expressed (Fig. S11**,** Wilcoxon-Mann Whitney test P < 2.2e-16), in opposite to what this assumption predicts.

For the sake of completeness, we repeated all these analyses (*i.e.*, regressions performed using cellular genes from either conditions Empty, Shble#1 and Shble#2, or from conditions Shble#3, Shble#4, Shble#5 and Shble#6), after having excluded cellular genes that were significantly affected in the same direction in both datasets (*i.e.,* after having removed the 27 cellular genes that are positively affected in both regression analyses and the 34 cellular genes that are negatively affected in both regression analyses, as displayed in **Fig. 3A**). The results did not change with regards to those communicated above (details in Fig. S12 and Fig. S13).

Overall, the global trends for our experiment results for the show that: i) overexpressing heterologous genes with extreme CUBias leads to changes in the protein-over-mRNA levels for a limited number of cellular genes; and that ii) overexpression of heterologous genes enriched in human-frequent codons does not have a differential impact on cellular genes as a function of their CUBias. Our results do not support thus the hypothesis for resource competition among mRNAs for the rare cellular resources of the translation machinery, as overexpression of genes with a given CUBias does not hinder translation of other genes with similar CUBias.

Heterologous expression of genes with extreme CUBias does not lead to a global alteration of the translation efficiency of cellular transcripts

We finally used a more integrated approach of the cell to examine how global codon content that is present at the transcriptomic layer 'flows' to the proteomic layer during the process of translation. This approach follows a supply and demand reasoning: a demand for being decoded is exerted by codons present in the pool of cellular mRNAs whereas the codon composition ultimately inferred from the cell proteome informs us to what extend this demand has been met. For this purpose, we calculated for each sample the abundance of each codon as they are represented in the cellular transcriptome and proteome (see Methods, and Table S5 and Table S6 for values of these abundance values). By dividing proteome-wide relative synonymous codon frequency (RSCF) of each codon by its transcriptome-wide RSCF counterpart we obtained for each sample an integrated view of how effectively each synonymous codon had been translated. We called this variable Prot-to-RNA RSCF (see Methods and Fig. S14). We validated this Prot-to-RNA RSCF variable as a proxy of whole-cell translation efficiency of synonymous codons by showing that it modestly but significantly correlates (correlation coefficient = 0.31, P = 0.045) with anticodon content, using previously published tRNA quantification data obtained on HEK293 cells (**Mattijssen et al, 2017** and see Table S7 for a detailed map of codon-

anticodon pairing and the associated counts of tRNA). We indeed observed a correlation between our Prot-to-RNA RSCF variable and the relative synonymous proportion of the total tRNAs decoding each codon encoding a similar amino acid (**Fig. 4A**). Further confirming this positive relationship between Prot-to-mRNA RSCF and tRNAs availability, we observed that when assessed individually on a per amino acid basis, a positive trend was found for 10 out of the 12 amino acids included in our analysis (the six amino acids Asn, Asp, Cys, His, Phe and Tyr have their two synonymous codons decoded by a single anticodon, and so were excluded by definition – see Methods) (**Fig. 4A**).

For each 59 amino acids-encoding codons we next tested if the version of the transfected construct had a significant effect onto the Prot-to-RNA RSCF of the considered codon (Table S8). We found a significant effect for 23 codons, representing a total of 11 amino acids (Leu, Arg, Ser, Val, Ala, Gly, Thr, Ile, Phe, Glu, Gln - see Table S9). Considering only these 23 codons, we interestingly observed in **Fig. 4B** that, compared to other samples, those expressing Shble#3 – an AT-rich version – seem to « favor » (respectively « avoid ») translation of GC-ending (respectively AT-ending) codons (Fig. S15 and Table S9). This intriguing pattern led us to consider the first two axes of variation in Prot-to-RNA RSCF across samples, in order to see whether samples clustered along these axes depending on the AT richness of the *shble* version they express. Results are shown in **Fig. 4C**. We first checked that correlation between heterologous transcripts expression and the first component (61% of variance) of variation of Prot-to-RNA RSCF across samples was non-significant, ruling out expression level of the different synonymous versions as a confounding factor (Fig. S16). Regarding the second component (14% of variance explained), we indeed observed (**Fig. 4C**) that samples expressing AT-rich versions (Shble#3, Shble#4, Shble#6) tended to cluster together while GC-rich versions (Shble#1, Shble#2, Shble#5 + Empty) tended to cluster together in opposite direction: samples expressing AT-rich versions display high values projected onto PC2 (median = 0.027) and those expressing GC-rich versions display low values (median = -0.019). Superimposing to this pattern the loadings onto the 2nd axis of the 59 variables (*i.e.,* 59 codons), the following association emerges: most samples expressing AT-rich versions (Shble#3, Shble#4 and Shble#6) seem to have preferentially translated cellular mRNAs rich in GC-ending codons compared to samples expressing GC-rich versions. This pattern is shown in **Fig. 4D**. Hence, although not the primary difference across samples (14% of inter-sample variance), this result suggests a differential effect in terms of codon third base-dependent translation efficiency in host cells driven by base composition of heterologous they express. This observation is not in total agreement to what we would have expected under our initial assumption, which stated that translation of cellular transcripts should be impacted depending on the overall match between their CUBias and the one of the heterologous genes expressed by the host cell, which was not the case here using COUSIN values of our synonymous versions, as displayed in **Fig. 4D**. As an example, versions Shble#1 and Shble#5 have clear distinct CUBias but the corresponding samples have very close Prot-to-RNA RSCF profiles. Instead

these two synonymous versions are very similar regarding their GC composition, supporting that base composition of heterologous genes may have a stronger translational consequences than their overall CUBias match with the human genome in our experimental setup.

Discussion

We present here a systematic analysis of the changes in the cellular transcriptomic and proteomic profiles upon experimental transfection, using a number of synonymous versions of heterologous genes with divergent CUBias. Our results show that transfection and heterologous gene expression elicited substantial changes in the cellular transcriptome, while changes in the proteome were of a lesser extent. Transcriptomic changes triggered by plasmid DNA transfection in human cells in culture mainly involved the activation of genes related to inflammatory response and to antiviral immunity. This cellular response was shared across all experimental conditions used (*i.e.* mock transfections, EGFP expression alone or SHBLE-EGFP expression) pointing towards a common response to the presence of plasmid dsDNA in the cytoplasm of the transfected cells. Indeed, DNA in eukaryotic cells is restricted to enveloped subcellular structures (nucleus, mitochondria, chloroplasts), and the presence of naked DNA in the cytoplasm is most often related to viral infections. In vertebrates several cytoplasmic DNA sensors exist such as the cyclic GMP-AMP synthase (cGAS) and the stimulator of interferon genes (STING) (**Wu & Chen, 2014**). Consistent with the activation of such DNA-sensing pathways and their known downstream effects on gene expression (**Stetson & Medzhitov, 2006; Wu & Chen, 2014**), we detect in our transfected cells many activated genes related to Type I interferon signaling pathway. This was the case of upregulation in all our experimental conditions of interferon stimulated genes ISG15 and ISG20, interferon induced proteins with tetratricopeptide repeats IFIT1, IFIT2 and IFIT3, or of the interferon alpha-inducible protein 6 IFI6 genes. Though our study did not aim at characterizing the HEK293 response to transfection, our results suggest that it induces an inflammatory response that could be dependent on STING activation, as recently shown by a study conducted on this cell line (**Khiar et al, 2017**). Finally, we also observed that transfection led to up-regulation of genes known to have viral RNA sensor roles, such as the four members of the OAS gene family (2'-5' oligoadenylate synthetases, OAS1, OAS2, OAS3, OASL), all of them enzymes regulated by interferon signaling and displaying a RNAse activity (**Kristiansen et al, 2011**; **Zhu et al, 2014**). Note that this observation is not necessarily in contradiction with the fact that we transfected DNA and not RNA, knowing that crosstalk exists between antiviral sensors of DNA and double-stranded RNA (**Cheng et al., 2007**). In conclusion transfection of plasmid DNA into HEK293 resulted in the induction of a

common anti-viral response through the overexpression of mRNAs involved in inflammatory processes.

Expressing plasmid-encoded heterologous genes in human cells triggered large transcriptomic changes. For each of the different synonymous construct we identify at least 500 cellular transcripts to be differentially expressed after transfection. In striking contrast to these transcriptomic changes we observed no differences in the proteome of the transfected cells. A technical explanation for this differential response between transcriptomic and proteomic responses could be related to the differential power of the technical approaches quantifying transcripts and proteins: on the one hand, RNASeq identifies by sequencing the presence of cDNA molecules, which are assembled into transcripts and eventually mapped onto the reference transcriptome; on the other hand, label-free mass spectrometry identifies the presence of peptides of a given mass and charge, compatible with several amino acid combinations, and for which the sequence of the putative peptide is identified after systematic comparison against the universe of possible peptides in the reference database. The essential difference in nature is thus that RNASeq can "discover" the presence of molecules, for instance by *de novo* transcriptome assembly, while label-free proteomics can only evaluate the presence of peptides originated from the proteins present in the reference database. Beyond these intrinsic difference in their technical limitations, transcriptomic and proteomic analyses present different sensitivity. In our case we have detected the presence of 13,737 gene transcripts (median value) per sample while we could only detect 2,969 (median value) proteins per sample. Furthermore, the detected proteins were biased towards those corresponding to transcripts displaying relatively little variation of expression after transfection, which are also the more expressed ones. Despite this technical gap, our results contribute to the growing evidence obtained by combining transcriptomics and proteomics (or Ribo-Seq) approaches, showing that proteomes tend to be more stable, and/or to display a larger inertia to change, than transcriptomes and so that many changes observed at the mRNA level are post-transcriptionally buffered and not immediately detected at the protein level. Such a trend has been also communicated for two yeasts species (**McManus et al., 2014**), for the molecular responses of other fungi either to stressors or environmental cues (**Brancini et al, 2019; Blevins et al, 2019**), as well as across primates (**Khan et al, 2013**). Regarding cultured human cells, it has been demonstrated that overall protein levels in aneuploidic cell lines - such as the HEK293 cells used here - are overall close to those of the normal diploid state, than overall transcript levels, further suggesting a buffering effect at the translation level (**Stingele et al, 2012**). In this regard, it has been shown that human tissues involved in similar broad functions display similar proteomes, even if their transcriptomes are divergent (**Wang , 2019**). It is also interesting to note that this more similar pattern of protein expression compared to mRNA expression across different tissues is conserved across evolutionary timescale, as recently demonstrated for mammals (**ZY Wang et al, 2020**). Finally, a

genome-wide association study illustrated that expression-QTL detected from different human cell lines usually have a more reduced effect on protein levels than on mRNA levels, and that this holds true whatever the noise level of the transcriptomics and proteomics data, excluding statistical power biases as a source to explain these differences between layers of expression (**Battle et al, 2015**). Additionally, differences in degradation rates of mRNAs and proteins can further enhance this gap between the transcriptomic and the proteomic responsiveness: in mammalian cells in culture, proteins display in average five times larger half-lifes than mRNAs (46 h *vs.* 9h, respectively), and there is no correlation between the stability of mRNA and the corresponding protein (**Schwanhaüsser et al., 2011**). It is thus conceivable that changes at the proteome level appear buffered because highly expressed and long-lived proteins might still be detectable present at their initial expression level (*i.e.* at the time transfection occurred) and had not sufficient time to achieve a complete turnover. Nevertheless, even if such effect exists, we anticipate its impact on our experimental data to be mitigated because we collected our cells for the transcriptomic and proteomic determination 48h after transfection. In summary, although we cannot rule out a technical effect in terms of sensitivity and number of features detected between RNA and protein measurements, and although we cannot dismiss the possible effects of differential degradation kinetics, we interpret that the observed differences in the extent of changes triggered by heterologous expression between the transcriptomic and the proteomic molecular layers possibly reflect a phenomenon of post-transcriptional buffering of biological significance.

It is well known that CUBias strongly influences the expression level of a gene (either directly or indirectly via linked variables such as GC composition, dinucleotide composition or mRNA folding energy – **Boël et al, 2016**; **Cambray, Guimaraes & Arkin, 2018**). Such *cis*- acting effects of CUBias on gene expression have been thoroughly documented for our *shble_egfp* experimental system **(Picard et al, in preparation)**. The purpose of the present study was mainly to determine whether and to what extent CUBias may display *trans*- acting effects on overall translation, specifically: does strong expression of a heterologous gene lead to a differential impact on the translation of cellular genes as a function of the match between their individual CUBias and the CUBias of the overexpressed heterologous gene? In mammalian cells, virtually all ribosomes are engaged in translation at any given time (**Princiotta et al, 2003**), and translation is the most expensive step in the biological information flow process (**Lynch & Marinov, 2015**), consuming around 45% of all cellular energy (**Princiotta et al, 2003**). It is thus conceivable that high expression levels of certain genes may come at the expense of a limited expression of other genes, due to competition for non-finite pool of translational resources shared by mRNAs present in the cell. This may be especially the case for highly abundant mRNA species that are not efficiently translated (**Shah et al, 2013**) – for example due to a poor CUBias. The existence of such *trans*- effects has been verified in *E. coli* using high-throughput approaches by Frumkin and

coworkers with respect to tRNAs availability (**Frumkin et al, 2018**). Another study on *S. cerevisiae* led to the same conclusion that CUBias of highly expressed genes is key in maintaining the overall translation efficiency of the rest of mRNAs present in the cell (**Qian et al, 2012**). Our study was precisely conceived to test on a mammalian model the hypothesis that CUBias of highly expressed genes impacts translation efficiency via *trans*- acting effects, linked to competition between heterologous transcripts and cellular transcripts to access tRNAs. We found no evidence of such effects in our experimental system, despite having constructed synonymous shble versions that are expressed at remarkably high levels. Indeed, heterologous *shble-egfp* transcripts represent more than 1% of transcripts in most samples (*i.e.*, above 10,000 Transcripts Per Million) and the SHBLE and EGFP proteins together represent more than 0.6% of the total protein abundance detected for all versions except Shble#6. Results presented in **Fig.2** and **Fig. 3** show no support for the codon usage-specific decrease in translation efficiency of cellular mRNAs expected under the hypothesis that high expression of heterologous genes leads to *trans*- acting effects through tRNA shortage. In the case of overexpression of codon-humanized genes, one could suppose that the pool of tRNAs corresponding to these most used codons in human genes is sufficiently large to buffer the strong additional demand exerted by the translation of heterologous mRNAs. It is thus perhaps not surprising that genes displaying increased or decreased translation efficiency when humanized heterologous genes are overexpressed do not differ in their CUBias preferences. However, regarding the use of rare codons that are decoded by tRNAs probably more prone to shortage, the competition hypothesis predicts that overexpression of heterologous genes enriched in rare codons should result in a negative impact on the translation of cellular genes enriched in these same rare codons (**Frumkin et al, 2018; Yona et al, 2013**). Our results however do not support this hypothesis. In contradiction to this expectation, we instead observe a decreased translation efficiency in cellular genes enriched in common codons when we force the cells to overexpress heterologous genes rich in rare codons. We ruled out that genes with decreased translation efficiency (negatively correlated with the amount of proteins expressed from versions rich in rare codons) are more expressed, which could have render them more prone to potential tRNA shortage independently of their CUBias. We must admit that we have not found yet a satisfying explanation of this counter-intuitive pattern.

The synonymous versions of the shble coding sequence used in this study were designed to present distinct CUBias, except for the first seven codons, which were identical across versions and that correspond to the amino acids in the AU1 epitope (MDTYRYI) used for western-blot protein detection. Thus, the chemical and coding environments immediately surrounding the translation starting point are identical for all synonymous versions. Notwithstanding, our postulate is that our gene recoding strategy allows to largely tease apart initiation- from elongation-driven effects. Indeed, local mRNA structures present around the start codon can hinder ribosomal binding and early

progression and hence translation initiation, but as a result of our design such local structures should be equally present in all our synonymous versions. Molecular modelling suggests indeed that translation is mainly limited by initiation rate rather than by elongation rate (**Riba et al, 2019**; **Shah et al, 2013**). These predictions are supported by numerous experimental studies reporting a link between mRNA secondary structures immediately upstream the translation start site and protein synthesis: the stronger the secondary structures the stronger the hamper for translation initiation (**Gu, Zhou & Wilke, 2010**). The inverse positive relationship between mRNAs with less structured 5' UTR and higher protein production seems to be conserved throughout evolution as has been identified in *E. coli* (**Kudla et al, 2009**), unicellular eukaryotic organisms (**Shah et al, 2013; Weinberg et al, 2016; SE Wang et al, 2020**) and mammals including humans (**Gandin et al, 2008; Mauger et al, 2019** but see **De Sousa Abreu et al, 2009** that described no effect of the initiation rate on translation efficiency in human transcripts). It also appears that, in addition to initiation, early elongation events (typically the decoding of the first 5 amino acids) can play a critical role in determining overall translation efficiency, either because ribosome arresting at certain early codons leads to abortive translation (**Verma et al, 2019**) or because early elongation steps interfere with initiation (**Chu et al, 2014**). Our design for synonymous version recoding of the shble gene focused thus on the elongation steps of protein synthesis, allowing us to evaluate the cellular impact of the increased demand for translation of either rare or common codons, rather than on the impact of codon recoding on translation initiation. This design was made on purpose because we were explicitly interested in studying how the elongation steps of a highly expressed gene imposes a burden onto the host's translation machinery and eventually leads to -trans effects. It remains nevertheless true that long-range interactions within the same mRNA molecule exist, so that the translation start can be differentially involved in global secondary structures for different shble gene versions (**Shah et al, 2013; Chu et al, 2014**).

Results presented in **Fig. 4** suggest that differences in the GC composition between our synonymous versions impacted host translation efficiency in opposite ways: cultures transfected with AT-rich versions (Shble#3, Shble#4, Shble#6) presented an enhanced translation efficiency of cellular mRNAs rich in GC-ending codons whereas cultures transfected with GC-rich versions (Shble#1, Shble#2, Shble#5) presented an enhanced translation efficiency of cellular mRNAs rich in AT-ending codons. Considering the hypothesis of *trans*- effects exerted on cellular genes through translation of heterologous genes, we expected our synonymous versions to cluster not according to their GC-content but rather according to their overall match with the CUBias of the host cell (proxied here by their COUSIN score). It is difficult to explain this effect of GC composition of heterologous genes on host translation. As a first line of thought regarding cellular mechanisms, we did not observe the expected pattern of AT-rich versions triggering the expression of *schlafen11*. In mammals, this gene is known to be a part of the innate immune response against viral infections resulting in an altered

translation in a CU-dependent manner, specifically impairing AT-rich transcript translation (**Li et al, 2012**). As a second line of thought regarding transcript stability, mRNA nucleotide composition regulates its own abundance independently of translational effects driven by CUBias (**Vogel et al, 2010**). Indeed, in mammals, effects of transcript GC composition on its own half-life are mediated either by changes in mRNA secondary structures (**Chamary & Hurst, 2005**) or changes decay pathways (**Courel et al, 2019**). Further, higher amounts of the corresponding protein were observed not because the novel CUBias arising from GC content manipulation led to an increased translation efficiency but rather simply because of the increased mRNA steady levels it enables (**Kudla et al, 2006; Mauger et al, 2019**). As a third line of thought, a highly expressed heterologous gene could globally modify the cellular translation dynamics because of the overload it imposes on the translation machinery, via ribosomal sequestration, for instance (**Princiotta et al, 2003**). It has been indeed shown that adenine-rich mRNAs promoted ribosomal binding to these transcripts (**Castillo-Méndez et al, 2012**).

Our results suggest that translation efficiency of human cellular transcripts in cells expressing a transgene is not dependent on the match between their CUBias and the one of the heterologous gene. This suggests that forcing human cells to express important quantity of heterologous proteins from underlying sequences with over-humanized CUBias – as usually done in biotechnology – is not detrimental to cells. Beyond this conclusion we would like to draw attention to the fact that modifying CUBias so that the designed sequences are either enriched or depleted in the most frequent synonymous codons found in a genome has not always straightforward consequences on translation efficiency. Albeit probably true that in most cases such strategy of codon “optimization” (or “de-optimization”) gives the expected results in terms of protein production (see **Quax et al, 2015** for a review; **Welch et al, 2009**; **Schmitz & Zhang, 2021**), there are several instances that gave surprising results, not only in humans but also for micro-organisms (**Pop et al, 2014**). In bacteria for example some studies reported that varying CUBias did not change protein expression of the target, was it a transgene or a gene in the genome of a recoded strain (**Kudla et al, 2009; Frumkin et al, 2018**) or worse reported that encoding a protein with putatively most optimal codons ultimately led to a decrease in expression (**Agashe et al, 2013**) or activity (**Zucchelli et al, 2017**). Regarding mammalian cells, the link between codon manipulation and protein expression is even more difficult to study as we do not know yet formally whether rare codon limit protein production. A systematic examination of sequence features that impact protein concentration performed on a human cell line revealed for example that CUBias only has a minor impact compared to other features (**Vogel et al, 2010**). It is also difficult for humans to establish a link between CUBias and protein expression because level of tRNAs is usually found to vary across tissues (**Dittmar, Goodenbour & Pan, 2006** but see **Pinkard et al, 2020**). Yet since tRNAs make the link between mRNA composition in codons and their decoding into amino acids we cannot establish a single rule for codon optimization that would be valid for every tissue. It is probable

that this problem of tissue-specific codon "optimization" is more salient in the context of biotechnology and heterologous gene expression than under physiological conditions (**Eraslan et al, 2019**). As a consequence of this difficulty, the relevance of the current adopted strategy of codon "optimization" in humans has been debated (**Mauro & Chappell, 2014**) and some studies found only very slight effects of the use of human most frequent codons in heterologous protein expression within human cell (**Ngumbela et al, 2008**).

Materials and Methods

Design of the synonymous six synonymous plasmidic constructs

The original sequence of the shble gene found in *Streptoalloteichus hindustans* was obtained from the GenBank database (X52869.1, https://www.ncbi.nlm.nih.gov/nuccore/X52869.1?report=genbank). Six synonymous coding sequences (CDS) of this gene were designed according to the 'one codon by amino acid' rule. Codon usage of these six synonymous versions was designed as follows: version Shble#1 uses exclusively the human most frequent codon; versions Shble#2 and Shble#3 use codons with the highest GC or the highest AT contents among the two most common codons, respectively; version Shble#4 uses exclusively the human least frequent codon; versions Shble#5 and Shble#6 use codons with the highest GC or the highest AT contents among the two least common codons, respectively. An invariable AU1 tag (MDTYRYI) was added at the start of each synonymous version, resulting in the same first seven codons for all synonymous versions. A PCA illustrating how these synonymous versions differ in terms of their CUBias among one another is given in Fig. S17. Each of these synonymous versions of the shble CDS was linked to the CDS of the Enhanced Green Fluorescent Protein (EGFP) via a P2A self-cleaving peptide sequence, so that a bicistronic mRNA is produced and translated into two proteins after cleavage of the P2A peptide. Synonymous versions were synthesized (Genescript) and cloned into the XhoI site of the pcDNA3.1-P2A-EGFP-C plasmid, that also contains in its backbone a Neomycin resistance gene (*3'-glycosyl phosphotransferase*). In addition to the six constructs bearing different synonymous versions of shble fused to egfp, a control vector missing the shble gene upstream the P2A-egfp region was designed, named the Empty version. This Empty version serves as a control to account for EGFP expression. Note that the sequence encoding EGFP is similar for all constructs and that its CUBias is strongly biased towards human's most favored codons, resembling to the CUBias of versions Shble#1 and Shble#2 (Fig. S17).

Cell transfection and sampling design

HEK293 cells (Human Embryonic Kidney cells, CRL-1573, ATCC) were cultured at 37°C and 5% CO2, in Minimum Essential Medium (Earle, M1MEM10K, Eurobio), with 10% FBS (Fetal Bovine Serum, CVFSVF0001, EuroBio) and 1% Penicillin-Streptomycin (15140122, Fisher scientific). Transfection was done in six-well plates, with $1.17x10^5$ cells/mL. The next day, medium was replaced by MEM 2% FBS. Each synonymous plasmid mixed with turbofect reagent (12331863, Fisher scientific) was added in each well and cells were sampled at day 2 (Trypsin-EDTA, CEZTDA000U, Eurobio) for further processing. In total, three independent series of transfection experiments – each with the eight conditions in duplicates – were performed. This resulted in 48 samples, balanced as follows: six

Shble#1 (three series of two replicates), six Shble#2 (three series of two replicates), six Shble#3 (three series of two replicates), six Shble#4 (three series of two replicates), six Shble#5 (three series of two replicates), six Shble#6 (three series of two replicates), as well as six Empty samples (three series of two replicates) and six Mock samples (three series of two replicates).

Sequencing, RNAseq data analyses and quantification

RNA extraction was performed using the Monarch Total RNA miniprep kit (T2010S, NEB), following the manufacturer's recommendations. Total RNAs were sent to Genewiz NGS laboratory (New Jersey, USA), where they performed polyA selection, strand-specific RNA library preparation and 2x150bp sequencing on an Illumina HiSeq4000 system. After demultiplexing, we received raw data of the 48 samples. Quality checks of raw reads were performed with FastQC (available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Reads were passed to Cutadapt (v1.10, **Martin, 2011**) to remove universal Illumina adapters then trimmed with Trimmomatic (v0.38, **Bolger, Lohse & Usadel, 2014**) using the following options: PE HEADCROP:13 SLIDINGWINDOW:4:30 MINLEN:85. Processed reads were pseudo-mapped onto the human transcriptome (//ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.39_GRCh38.p13). To allow reads coming from expression of the genes contained in the plasmids to be aligned, the transcriptome was appended with the Neomycin resistance transcript sequence and with the shbleX-p2a-egfp transcript sequence (where X stands for 1 to 6). For each sample transfected with Shble#1 to Shble#6, the corresponding synonymous sequence of shble was used to allow proper mapping. For Empty, only the egfp transcript sequence was added. For Shble#4 and Shble#6 samples, spliced forms of the shble-p2a-egfp transcript identified (**Picard et al, in preparation**) were also added to allows reads spanning the junctions to be aligned. Pseudo-mapping onto the transcriptome was performed with Kallisto (v0.46.0, **Bray et al, 2016**) with default options except for the number of bootstrap samples and the number of threads, that were respectively set as follows: -b 100 -t 16. Quantification at the transcripts level obtained from Kallisto as Transcripts Per Million (TPM) were retrieved for further analyses, both from cellular and heterologous transcripts. For each of the three independent experiments of transfection per construct we averaged TPM obtained from the two repetitions of each experiment, ultimately obtaining data for 24 samples (eight conditions, three experiments) – (Fig. S18).

Differential expression of cellular transcripts after 2-days post transfection by different synonymous constructs

For each condition, we defined the set of expressed mRNAs as those above 1 TPM in at least two out the three replicates that correspond to the condition. On average across all eight conditions 13,717

mRNA were detected as being expressed, ranging from 13,296 in the Mock condition to 13,951 in the Shble#5 condition. Working on the union of these sets of expressed mRNAs (14,226 mRNA), fold-changes relative to the Mock were calculated for each seven conditions (Shble#1 to Shble#6 + Empty). Genes were considered differentially expressed relatively to the Mock in a condition if the median fold-change (across the three samples corresponding to this condition) was above 2 or below 0.5 for this gene. In parallel, we also used DESeq2 (**Love, Huber & Anders, 2014**) on the total set of 19,812 mRNAs to similarly identify genes differentially expressed in comparison to the Mock after transfection by our different constructs. For this, we used counts estimated by Kallisto, that we normalized using the estimateSizeFactors() function of DESeq2. This second approach presents the advantage of returning q-values and takes into account the lack of power of differential expression for genes with low counts. Overlaps of genes identified as differentially expressed by these two distinct approaches were considered for the analysis of differential expression we describe in the corresponding part of the Results section of the manuscript.

Content specificity of sets of differentially expressed genes using combinatorial configurations

We used a combinatorial approach to estimate to what extent the content of the six sets (Shble#1 to Shble#6) of DiffExp cellular transcripts compared to the Mock was version-specific or redundant across versions. For this, for a given number N of sets included (N from 1 to 6), we calculated the number of unique genes present in the union of these N sets. For each collection of size N, we reported the median number of unique genes obtained from all possible arrangements. The number of possible arrangements of size N drawn among six sets is 6, 15, 20, 15, 6, 1, for N ranging from 1 to 6, respectively.

Protein extraction and label-free proteomic analysis

For protein extraction, the two replicates sampled from the same transfection experiment were pooled, thereby reducing the number of samples from 48 to 24. Solubilized proteins were resuspended in Laemmli buffer and 20-30 µg of proteins were stacked on a SDS-PAGE gel. Proteins were in-gel digested using Trypsin (Trypsin Gold, Promega), as previous described in **Shevchenko et al, 2006**. Proteomic data were collected in data dependent acquisition mode using a Q Exactive HF mass spectrometer coupled with Ultimate 3000 RSLC (Thermo Fischer Scientific). The software MaxQuant (**Cox & Mann, 2008**) was used to analyze tandem mass spectrometry data. All m/Z spectra were searched using standard settings with the search engine Andromeda (**Cox et al, 2011**) against a target decoy database to deliver false-positive estimations. The database contains entries from the *H. sapiens* Reference Proteome (UP000005640, release 2019_02, https://www.uniprot.org) and a list of potential

contaminants. Sequences of the two heterologous proteins of interest (SHBLE and EGFP) were added into this database to allow for their identification. Search parameters were let at their default values, oxidation (Met) and acetylation (Nt) were applied as variable modifications and carbamidomethyl (Cys) as a fixed modification. FDRs of peptides and proteins identification were let at their default values (both at 1%). Proteins groups were automatically constructed by MaxQuant. A representative ID in each protein group was automatically selected using an in-house bioinformatics tool (Leading_v3.2) After excluding usual contaminants, 4539 human proteins were identified in at least one sample (out of 24 samples). Protein quantifications in intensity based absolute quantification (iBAQ, **Schwanhausser et al, 2011**) were retrieved and normalized by the total sum of iBAQ within each sample (heterologous proteins excluded) to obtain relative iBAQ (riBAQ). riBAQ quantification has been shown to accurately report the mole fraction of each protein within sample (**Shin et al, 2013**). Protein quantification by Label Free Quantification (LFQ) was also retrieved for the analysis of differentially expressed proteins (**Cox et al., 2014**).

Analysis of differentially expressed proteins

The software Perseus (**Tyanova et al, 2016**) was used to identify human proteins which level of expression vary depending on the version of the plasmid that was transfected into our cells. Adding the Mock condition in addition to the six synonymous versions and the Empty construct, this resulted to a total of eight groups, each supported by n=3 biological replicates. We used LFQ for this analysis of differentially expressed proteins as this metrics is tailored for comparisons of the expression level of a given protein between different samples. LFQ values were log2 transformed and the following rationale was applied: proteins that were not present in at least 2 out of 3 replicates in a least one condition were filtered out. Only the remaining 1,989 proteins were considered for the rest of the analysis as they probably represent proteins identified with sufficient level of confidence. Note that this number is low due to the inherent calculation of LFQ values, which returns zero when there is not enough information for a peptide to be detected (**Cox et al, 2014**).

Analysis of proteins with a qualitative expression pattern according to the different conditions

Complementary to the analysis of differentially expressed proteins, we selected proteins that displayed qualitative differences (*i.e.,* ON/OFF) across our eight groups (Shble#1 to Shble#6, Empty, Mock). To do so, we selected – based on their riBAQ – proteins that were detected in all three replicates of at least one group but not detected in any of the three replicates in at least another group. This stringent definition of qualitative expression pattern yielded 369 proteins displaying such behavior (see Table S3). Within each of the 24 samples, we found that those expressed in the considered sample

among this set of 369 proteins systematically displayed lower riBAQ values than the rest of the proteins expressed in that sample (Fig. S19).

Calculation of Protein to mRNA ratios for samples expressing over-humanized versions compared to ratios in the Mock condition

We tested for competition between translation of mRNAs from plasmidic genes that use codons close to the human average CUBias and translation of host mRNAs. To do so, we calculated separately for Shble#1, Shble#2 and Empty conditions the ratio between gene's Protein to mRNA ratios (riBAQ/TPM) in the considered condition and in the Mock condition. For each condition, ratios compared to the Mock were calculated from the n=3 samples of the condition and the n=3 Mock samples. After excluding genes that were not detected in all the three Mock samples, we retained 2471 genes for which we calculated these $[riBAQ/TPM]_{Condition}$ / $[riBAQ/TPM]_{Mock}$ ratios. These ratios enable to compare the Protein to mRNA ratio of cellular genes in transfected cells in regard to the same measure in the absence of heterologous expression.

Changes in individual Protein to mRNA ratios under different conditions and levels of heterologous protein expression

Transfected samples were separated depending on the CUBias of the heterologous genes they encoded: nine samples expressing heterologous proteins with an over-humanized CUBias (n=3 Empty samples, n=3 Shble#1 samples, and n=3 Shble#2 samples) and twelve samples expressing EGFP plus a version of SHBLE rich in rare codons (n=3 Shble#3 samples, n=3 Shble#4 samples, n=3 Shble#5 samples, and n=3 Shble#6 samples). Adding the three samples of the Mock condition, this led to two sets of samples (n=12 and n=15) that do not overlap except for the three Mock samples. Note that Mock samples were included to serve as «anchors» for our linear regressions, representing what happens in the absence of heterologous protein expression. Independently for these two sets of samples, we used linear regressions to model how riBAQ/TPM ratios of a given gene varied with expression levels of over-humanized (n=9 + the three Mock) or rare codons rich (n=12 + the three Mock) heterologous proteins (EGFP+ SHBLE). The range of the x axis for these two sets of regression is provided in Fig. S8. Before running regressions, we filtered genes based on their values of riBAQ/TPM ratio in order to calculate a slope only when considered sufficiently meaningful. We used the following rationale: when the regression was performed based on the n=12 samples, we fixed a threshold of at least 8 non-null values of ratio and when the regression was performed based on the n=15 samples we fixed a threshold of at least 9 non-null values of ratio. This results in 2,554 and 2,585 genes to analyze for regressions based on n=12 and n=15 samples respectively, that later became 2,550 and 2,580 after having removed the very few genes with more than three (respectively four) infinite values of ratio. In each case, genes for

which the F-test testing the significance of the regression slope had a p-value smaller than alpha = 5% were considered for further analysis.

Measurement of codon usage bias

COUSIN scores (Codon Usage Similarity Index, **Bourret, Alizon & Bravo, 2019**) were used to compare the (overall) codon usage of both human and plasmidic genes to the average *Homo sapiens* codon usage. The COUSIN score reflects the extent to which codon usage of a query sequence matches the one of a reference. In our study, the reference is the whole codon composition of the coding part of the human genome. Briefly values above 1 reflect an overmatch compared to the codon usage preferences of the reference while values below zero reflect preferences opposite to those of the reference. Note that by design, all six synonymous versions of the shble CDS have a different COUSIN values, with some being over-humanized (higher COUSIN; Shble#1 and Shble#2) while others being enriched in infrequent codons compared to the average human codon usage (lower COUSIN; Shble#3 to Shble#6). Ordered by decreasing COUSIN scores, the COUSIN score of the six different synonymous versions based on their CDS are as followed: 3.473 (Shble#1), 3.421 (Shble#2), 0.192 (Shble#5), -0.484 (Shble#3), -1.323 (Shble#6) and -2.533 (Shble#4). The COUSIN score of the CDS encoding EGFP is high (3.379), close to the one of Shble#1 and Shble#2 versions (Fig. S17).

Comparison of COUSIN scores distribution between two different gene sets

Sets of genes with either a significantly positive or negative slope of riBAQ/TPM ratio with increasing heterologous protein expression across samples were compared regarding their CU. We compared the distribution of the COUSIN scores between the two sets of genes using Anderson-Darling tests, with the use of the *ad_test()* function implemented in the R package "twosamples" (https://github.com/cdowd/twosamples).

Synonymous codon content in sample's proteomes compared to transcriptomes

For every annotated gene in the human genome the number of occurrences of each 59 amino acid encoding codons (Tryptophane and Methionine excluded) was retrieved using the CDS of its longest predicted transcript. Then using expression levels of either all detected cellular mRNAs or all detected cellular proteins (measured in TPM and riBAQ, respectively), we quantified for each codon the amount of its "expression" in the transcriptome or in the proteome of our 24 samples as follows :$\sum_{g=Gene} Codoncount_g \cdot Expression_g$. Codon content in each sample was hence represented by two vectors of length 59, one derived from its transcriptome-wide profile of expression and another derived from its proteome-wide profile of expression. From these vectors we obtained vectors of

transcriptome-wide and proteome-wide relative synonymous codon frequency (RSCF) by dividing values of each codons encoding a similar amino acid by the total sum of values of codons encoding the considered amino acid. Ultimately, to quantify how the CUBias 'flows' from transcriptome to proteome we divided the proteome-wide RSCF of each sample by its transcriptome-wide RCSF counterpart. We called this variable Prot-to-RNA RSCF and considered it as a proxy for whole-cell translation efficiency of synonymous codons. Note that our codon counts vectors were computed without considering heterologous features in the cell transcriptome and proteome.

Analysis of Prot-to-RNA RSCF variation across conditions

For each 59 amino acid encoding codons we tested the effect of the transfected construct onto the whole-cell translation efficiency of the considered codon through linear models, with Prot-to-RNA RSCF as a response variable. Before running our models, we checked assumptions of residuals normality and homoscedasticity using Shapiro and Levene tests, respectively. Only 3 out of the 59 codons – Thr_ACG, Gly_GGA and Leu_UUA (Table S8) – deviated from normality but we still decided to perform parametric tests for all 59 codons. For each individual codon a one-way ANOVA was performed with the eight conditions corresponding to our design (Shble#1 to Shble#6, Empty, Mock), each supported by n=3 measurements of Prot-to-RNA RSCF. The p-value threshold of significance for the ANOVA F-test was set at 0.05. When this test was significant for a codon, we performed multiple pairwise comparisons across constructions on the codon Prot-to-RNA RSCF applying an FDR correction (**Benjamini & Hochberg, 1995**).

Link between Prot-to-RNA RSCF and tRNA availability

We checked that our Prot-to-RNA RSCF constructed variable could serve as a proxy for whole-cell translation efficiency of synonymous codons by leveraging tRNA count data obtained on HEK293 cells (**Mattijssen et al, 2017 –** Table S7). The purpose was to combine our Prot-to-RNA RSCF values obtained for synonymous codons with a measure that reflects tRNA-anticodon availability with respect to the considered amino acid. For this, we used relative fractions of the total tRNA counts corresponding to a given amino acid that decode each synonymous codon. For the six two-fold degenerated amino acids that have their synonymous codons decoded by a single anticodon (Asn, Asp, Cys, His, Phe, Tyr), these relative fractions are necessarily of values 1 for each of the two synonymous codons. In cases where several synonymous codons can be decoded by the same tRNA-anticodon species, the same relative fraction of the total tRNAs was attributed to these codons. Considering the case of Alanine, the four codons are only decoded by three tRNA-anticodon species: GCA by the UGC anticodon, GCG by the CGC anticodon, and both GCU and GCC by the anticodon AGC. Hence, in such case, fractions of tRNAs

that are devoted to decode GCU and GCC (relatively to the total tRNAs counts) would be the same for both codons and be equal to:

tRNA-AGC / (tRNA-AGC + tRNA-UGC + tRNA CGC + tRNA-AGC). In complex cases where a single codon can be read by several distinct tRNA-anticodon species - namely Thr-ACC decoded by UGU and AGU anticodons, Thr-ACG that can be decoded by UGU and CGU anticodons, Pro-CCU that can be decoded by AGG and UGG anticodons, and Pro-CCG that can be decoded by CGG and UGG anticodons (Table S7) we decided to not attribute to Thr-ACC, Thr-ACG, Pro-CCU and Pro-CCG a fraction of total tRNAs because we could not have a one to one mapping. This explains why in **Fig. 4A** only two Threonine codons (Thr-ACU and Thr-ACA) and two Proline codons (Pro-CCC and Pro-CCA) are represented.

Functional enrichment analysis

The Functional Annotation Chart module of the DAVID tool (https://david.ncifcrf.gov/, **Dennis et al, 2003**) was used to detect functional categories over-represented in gene sets. Functional categories displaying a Fold-Enrichment above 2 and a FDR-corrected p-value below 0.05 were considered as significantly over-represented.

Statistical analyses

All statistical tests were performed with R (version 3.6.3). Principal components analyses were performed using the *prcomp()* function of the R package "stats". Principal components (the eigenvectors of the covariance matrix) were visualized and superimposed to PCA graphs using prcomp()$rotation. The heatmap presented in **Fig. 4** was constructed using the *heatmap.3()* function (https://raw.githubusercontent.com/obigriffith/biostar-tutorials/master/Heatmaps/heatmap.3.R), with Z-normalization performed intra-row (*i.e.* on a codon by codon basis). Except when explicitly noted, all reported correlation coefficients correspond to Pearson correlation coefficients.

Acknowledgments

The authors acknowledge the IRD itrop HPC (South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper.

Mass spectrometry experiments were carried out using facilities of the Functional Proteomics Platform (FPP) of the Proteomics Pole of Montpellier (PPM, Montpellier, France).

Author contributions

IGB conceived the work and obtained funding; MALP, FL, MD, AD, KEK performed experiments and analysed primary data; AJJ analysed the data; AJJ and IGB conceptualized the results and wrote the manuscript.

Funding

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

References

Agashe, D., Martinez-Gomez, N. C., Drummond, D. A., & Marx, C. J. (2013). Good codons, bad transcript: Large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Molecular Biology and Evolution*, *30*(3), 549–560. https://doi.org/10.1093/molbev/mss273

Akashi, H. (2003). Translational selection and yeast proteome evolution. *Genetics*, *164*(4), 1291–1303. https://doi.org/10.1093/genetics/164.4.1291

Battle, A., Khan, Z., Wang, SH., Mitrano, A., Ford, MJ., Pritchard, JK., Gilad, Y. (2015). Impact of regulatory variation from RNA to protein. *Science*, *347*(6228), 1362–1367. 10.1126/science.1260793

Benjamini, Y. & Hochberg, Y. (1995). *J. R. Stat. Soc. Ser. B Stat. Methodol.* https://doi.org/10.2307/2346101

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., & Blüthgen, N. (2013). Efficient translation initiation dictates codon usage at gene start. *Molecular Systems Biology*, *9*. https://doi.org/10.1038/msb.2013.32

Blevins, W. R., Tavella, T., Moro, S. G., Blasco-Moreno, B., Closa-Mosquera, A., Díez, J., … Albà, M. M. (2019). Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker's yeast. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-47424-w

Boël, G., Letso, R., Neely, H., Price, W. N., Wong, K. H., Su, M., Luff, J. D., Valecha, M., Everett, J. K., Acton, T. B., Xiao, R., Montelione, G. T., Aalberts, D. P., & Hunt, J. F. (2016). Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature*, *529*(7586), 358–363. https://doi.org/10.1038/nature16509

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bourret, J., Alizon, S., & Bravo, I. G. (2019). COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences. *Genome Biology and Evolution*, *11*(12), 3523–3528. https://doi.org/10.1093/gbe/evz262

Brancini, G. T. P., Ferreira, M. E. S., Rangel, D. E. N., & Braga, G. Ú. L. (2019). Combining Transcriptomics and Proteomics Reveals Potential Post-transcriptional Control of Gene Expression after Light Exposure in Metarhizium acridum. *G3: Genes, Genomes, Genetics*, *9*(9), 2951–2961. https://doi.org/10.1534/g3.119.400430

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Callens, M., Pradier, L., Finnegan, M., Rose, C., & Bedhomme, S. (2021). Read between the Lines: Diversity of Nontranslational Selection Pressures on Local Codon Usage. *Genome Biology and Evolution*, *13*(9), 1–14. https://doi.org/10.1093/gbe/evab097

Cambray, G., Guimaraes, J. C., & Arkin, A. P. (2018). Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in escherichia coli. *Nature Biotechnology*, *36*(10), 1005. https://doi.org/10.1038/nbt.4238

Caspersson, T., Farber, S., Foley, G. E., Kudynowski, J., Modest, E. J., Simonsson, E., Wagh, U., & Zech, L. (1968). Chemical differentiation along metaphase chromosomes. *Experimental Cell Research*, *49*(1), 219–222. https://doi.org/10.1016/0014-4827(68)90538-7

Castillo-Méndez, M. A., Jacinto-Loeza, E., Olivares-Trejo, J. J., Guarneros-Peña, G., & Hernández-Sánchez, J. (2012). Adenine-containing codons enhance protein synthesis by promoting mRNA binding to ribosomal 30S subunits provided that specific tRNAs are not exhausted. *Biochimie*, *94*(3), 662–672. https://doi.org/10.1016/j.biochi.2011.09.019

Chamary, J. V., & Hurst, L. D. (2005). Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biology*, *6*(9). https://doi.org/10.1186/gb-2005-6-9-r75

Chamary, J. V., Parmley, J. L., & Hurst, L. D. (2006, February). Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg1770

Chaney, J. L., & Clark, P. L. (2015). Roles for Synonymous Codon Usage in Protein Biogenesis. *Annual Review of Biophysics*, *44*(1), 143–166. https://doi.org/10.1146/annurev-biophys-060414-034333

Charneski, C. A., & Hurst, L. D. (2013). Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. *PLoS Biology*, *11*(3). https://doi.org/10.1371/journal.pbio.1001508

Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L., & McAdams, H. H. (2004). Codon usage between genomes is constrained genome-wide mutational processes. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(10), 3480–3485. https://doi.org/10.1073/pnas.0307827100

Cheng, G., Zhong, J., Chung, J., & Chisari, F. V. (2007). Double-stranded DNA and double-stranded RNA induce a common antiviral signaling pathway in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(21), 9035–9040. https://doi.org/10.1073/pnas.0703285104

Chu, D., Kazana, E., Bellanger, N., Singh, T., Tuite, M. F., & Von Der Haar, T. (2014). Translation elongation can control translation initiation on eukaryotic mRNAs. *EMBO Journal*, *33*(1), 21–34. https://doi.org/10.1002/embj.201385651

Courel, M., Clément, Y., Bossevain, C., Foretek, D., Cruchez, O. V., Yi, Z., Bénard, M., Benassy, M. N., Kress, M., Vindry, C., Ernoult-Lange, M., Antoniewski, C., Morillon, A., Brest, P., Hubstenberger, A., Crollius, H. R., Standart, N., & Weil, D. (2019). Gc content shapes mRNA storage and decay in human cells. *ELife*, *8*, 1–32. https://doi.org/10.7554/eLife.49708

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, *26*(12), 1367–1372. https://doi.org/10.1038/nbt.1511

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., & Mann, M. (2011). Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*, *10*(4), 1794–1805. https://doi.org/10.1021/pr101065j

Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., & Mann, M. (2014). Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ* □ S Technological Innovation and Resources. *Molecular & Cellular Proteomics*, *13*, 2513–2526. https://doi.org/10.1074/mcp

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, *4*(5), 3–4. https://doi.org/10.1186/gb-2003-4-9-r60

De Sousa Abreu, R., Penalva, L. O., Marcotte, E. M., & Vogel, C. (2009). Global signatures of protein and mRNA expression levels. *Molecular BioSystems*. https://doi.org/10.1039/b908315d

Dittmar, K. A., Goodenbour, J. M., & Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. *PLoS Genetics*, *2*(12), 2107–2115. https://doi.org/10.1371/journal.pgen.0020221

Drummond, D. A., & Wilke, C. O. (2008). Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell*, *134*(2), 341–352. https://doi.org/10.1016/j.cell.2008.05.042

Drummond, D. A. & Wilke, C. O. (2009). The evolutionary consequences of erroneous protein synthesis. In *Nature Reviews Genetics* (Vol. 10, Issue 10, pp. 715–724). https://doi.org/10.1038/nrg2662

Duret, L., & Mouchiroud, D. (1999). *Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis* (Vol. 96). Retrieved from www.pnas.org

Duret, L. (2000). tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends in Genetics*, *16*(7), 287–289. https://doi.org/10.1016/S0168-9525(00)02041-2

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetic & Development*, 12(6):640-649*6*. 10.1016/s0959-437x(02)00353-2

Eraslan, B., Wang, D., Gusic, M., Prokisch, H., Hallström, B. M., Uhlén, M., … Gagneur, J. (2019). Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. *Molecular Systems Biology*, *15*(2), 1–25. https://doi.org/10.15252/msb.20188513

Frenkel-Morgenstern, M., Danon, T., Christian, T., Igarashi, T., Cohen, L., Hou, Y. M., & Jensen, L. J. (2012). Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. *Molecular Systems Biology*, *8*(572), 1–10. https://doi.org/10.1038/msb.2012.3

Frumkin, I., Lajoie, M. J., Gregg, C. J., Hornung, G., Church, G. M., & Pilpel, Y. (2018). Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(21), E4940–E4949. https://doi.org/10.1073/pnas.1719375115

Gandin, V., Miluzio, A., Barbieri, A. M., Beugnet, A., Kiyokawa, H., Marchisio, P. C., & Biffo, S. (2008). Eukaryotic initiation factor 6 is rate-limiting in translation, growth and transformation. *Nature*, *455*(7213), 684–688. https://doi.org/10.1038/nature07267

Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., & Futcher, B. (2014). Measurement of average decoding rates of the 61 sense codons in vivo. *ELife*, *3*, 1–20. https://doi.org/10.7554/eLife.03735

Gingold, H., & Pilpel, Y. (2011). Determinants of translation efficiency and accuracy. *Molecular Systems Biology*, *7*(481), 1–13. https://doi.org/10.1038/msb.2011.14

Gingold, H., Dahan, O., & Pilpel, Y. (2012). Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome. *Nucleic Acids Research*, *40*(20), 10053–10063. https://doi.org/10.1093/nar/gks772

Gingold, H., Tehler, D., Christoffersen, N. R., Nielsen, M. M., Asmar, F., Kooistra, S. M., Christophersen, N. S., Christensen, L. L., Borre, M., Sørensen, K. D., Andersen, L. D., Andersen, C. L., Hulleman, E., Wurdinger, T., Ralfkiær, E., Helin, K., Grønbæk, K., Orntoft, T., Waszak, S. M., … Pilpel, Y. (2014). A dual program for translation regulation in cellular proliferation and differentiation. *Cell*, *158*(6), 1281–1292. https://doi.org/10.1016/j.cell.2014.08.011

Gouy, M., & Gautier, C. (1982). Codon usage in bacteria: Correlation with gene expressivity. *Nucleic Acids Research*, *10*(22), 7055–7074. https://doi.org/10.1093/nar/10.22.7055

Gu, W., Zhou, T., & Wilke, C. O. (2010). A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Computational Biology*, *6*(2), 1–8. https://doi.org/10.1371/journal.pcbi.1000664

Hershberg, R., & Petrov, D. A. (2008). Selection on Codon Bias. *Annual Review of Genetics*, *42*(1), 287–299. https://doi.org/10.1146/annurev.genet.42.110807.091442

Hershberg, R., & Petrov, D. A. (2009). General rules for optimal codon choice. *PLoS Genetics*, *5*(7). https://doi.org/10.1371/journal.pgen.1000556

Hia, F., & Takeuchi, O. (2021). The effects of codon bias and optimality on mRNA and protein regulation. *Cellular and Molecular Life Sciences*, *78*(5), 1909–1928. https://doi.org/10.1007/s00018-020-03685-7

Holmquist, G. P. (1989). Evolution of chromosome bands: Molecular ecology of noncoding DNA. *Journal of Molecular Evolution*, *28*(6), 469–486. https://doi.org/10.1007/BF02602928

Ikemura, T. (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of Molecular Biology*, *151*(3), 389–409. https://doi.org/10.1016/0022-2836(81)90003-6

Ikemura, T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *Journal of Molecular Biology*, *158*(4), 573–597. https://doi.org/10.1016/0022-2836(82)90250-9

Kanaya, S., Yamada, Y., Kudo, Y., & Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: Gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, *238*(1), 143–155. https://doi.org/10.1016/S0378-1119(99)00225-5

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., & Ikemura, T. (2001). Codon usage and tRNA genes in eukaryotes: Correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *Journal of Molecular Evolution*, *53*(4–5), 290–298. https://doi.org/10.1007/s002390010219

Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., & Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. *Nature*, *467*(7311), 103–107. https://doi.org/10.1038/nature09322

Khan, Z., Michael J. Ford, Darren A. Cusanovich, Amy Mitrano, Jonathan K. Pritchard, Yoav Gilad (2013). Primate Transcript and Protein Expression Levels Evolve Under Compensatory Selection Pressures, *Science*, *342*(6162), 1100-1104. https://doi.org/10.1126/science.1242379

Khiar, S., Lucas-Hourani, M., Nisole, S., Smith, N., Helynck, O., Bourgine, M., Ruffié, C., Herbeuval, J. P., Munier-Lehmann, H., Tangy, F., & Vidalain, P. O. (2017). Identification of a small molecule that primes the type i interferon response to cytosolic DNA. *Scientific Reports*, *7*(1), 1–15. https://doi.org/10.1038/s41598-017-02776-z

Kristiansen, H., Gad, H. H., Eskildsen-Larsen, S., Despres, P., & Hartmann, R. (2011). The oligoadenylate synthetase family: An ancient protein family with multiple antiviral activities. *Journal of Interferon and Cytokine Research*, *31*(1), 41–47. https://doi.org/10.1089/jir.2010.0107

Kudla, G., Lipinski, L., Caffin, F., Helwak, A., & Zylicz, M. (2006). High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biology*, *4*(6), 0933–0942. https://doi.org/10.1371/journal.pbio.0040180

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of expression in escherichia coli. *Science*, *324*(5924), 255–258. https://doi.org/10.1126/science.1170160

Li, M., Kao, E., Gao, X., Sandig, H., Limmer, K., Pavon-Eternod, M., … David, M. (2012). Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. *Nature*, *491*(7422), 125–128. https://doi.org/10.1038/nature11433

Li, G. W., Burkhardt, D., Gross, C., & Weissman, J. S. (2014). Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, *157*(3), 624–635. https://doi.org/10.1016/j.cell.2014.02.033

Liu, Y., Beyer, A., & Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, *165*(3), 535–550. https://doi.org/10.1016/j.cell.2016.03.014

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. https://doi.org/10.1186/s13059-014-0550-8

Lynch, M., & Marinov, G. K. (2015). The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(51), 15690–15695. https://doi.org/10.1073/pnas.1514974112

Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal**,** [S.l.], v. 17, n. 1, p. pp. 10-12, may 2011. ISSN 2226-6089. https://journal.embnet.org/index.php/embnetjournal/article/view/200/479

Mattijssen, S., Arimbasseri, A. G., Iben, J. R., Gaidamakov, S., Lee, J., Hafner, M., & Maraia, R. J. (2017). LARP4 mRNA codon-tRNA match contributes to LARP4 activity for ribosomal protein mRNA poly(A) tail length protection. *ELife*, *6*, 1–33. https://doi.org/10.7554/eLife.28889

Mauger, D. M., Joseph Cabral, B., Presnyak, V., Su, S. V., Reid, D. W., Goodman, B., … McFadyen, I. J. (2019). mRNA structure regulates protein expression through changes in functional half-life. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(48), 24075–24083. https://doi.org/10.1073/pnas.1908052116

Mauro, V. P., & Chappell, S. A. (2014). A critical analysis of codon optimization in human therapeutics Optimizing codon usage for increased protein expression. *Trends in Molecular Medicine*, *20*(11), 604–613. https://doi.org/10.1016/j.molmed.2014.09.003.A

McManus, C. J., May, G. E., Spealman, P., & Shteyman, A. (2014). Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Research*, *24*(3), 422–430. https://doi.org/10.1101/gr.164996.113

Newman, Z. R., Young, J. M., Ingolia, N. T., & Barton, G. M. (2016). Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(10), E1362–E1371. https://doi.org/10.1073/pnas.1518976113

Ngumbela, K. C., Ryan, K. P., Sivamurthy, R., Brockman, M. A., Gandhi, R. T., Bhardwaj, N., & Kavanagh, D. G. (2008). Quantitative effect of suboptimal codon usage on translational efficiency of mRNA encoding HIV-1 Gag in intact T cells. *PLoS ONE*, *3*(6), 1–5. https://doi.org/10.1371/journal.pone.0002356

Novoa, E. M., Jungreis, I., Jaillon, O., Kellis, M., & Leitner, T. (2019). Elucidation of Codon Usage Signatures across the Domains of Life. *Molecular Biology and Evolution*, *36*(10), 2328–2339. https://doi.org/10.1093/molbev/msz124

Pechmann, S., & Frydman, J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature Structural and Molecular Biology*, *20*(2), 237–243. https://doi.org/10.1038/nsmb.2466

Pelechano, V., Wei, W., & Steinmetz, L. M. (2015). Widespread co-translational RNA decay reveals ribosome dynamics. *Cell*, *161*(6), 1400–1412. https://doi.org/10.1016/j.cell.2015.05.008

Perl, K., Ushakov, K., Pozniak, Y., Yizhar-Barnea, O., Bhonker, Y., Shivatzki, S., … Shamir, R. (2017). Reduced changes in protein compared to mRNA levels across non-proliferating tissues. *BMC Genomics*, *18*(1), 1–14. https://doi.org/10.1186/s12864-017-3683-9

Pinkard, O., McFarland, S., Sweet, T., & Coller, J. (2020). Quantitative tRNA-sequencing uncovers metazoan tissue-specific tRNA regulation. *Nature Communications*, *11*(1), 1–15. https://doi.org/10.1038/s41467-020-17879-x

Plotkin, J. B., Robins, H., & Levine, A. J. (2004). Tissue-specific codon usage and the expression of human genes. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(34), 12588–12591. https://doi.org/10.1073/pnas.0404957101

Plotkin, J. B., & Kudla, G. (2011, January). Synonymous but not the same: The causes and consequences of codon bias. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2899

Pop, C., Rouskin, S., Ingolia, N. T., Han, L., Phizicky, E. M., Weissman, J. S., & Koller, D. (2014). Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation . *Molecular Systems Biology*, *10*(12), 770. https://doi.org/10.15252/msb.20145524

Pouyet, F., Mouchiroud, D., Duret, L., & Sémon, M. (2017). Recombination, meiotic expression and human codon usage. *ELife*, *6*, 1–19. https://doi.org/10.7554/eLife.27344

Princiotta, M. F., Finzi, D., Qian, S. B., Gibbs, J., Schuchmann, S., Buttgereit, F., … Yewdell, J. W. (2003). Quantitating protein synthesis, degradation, and endogenous antigen processing. *Immunity*, *18*(3), 343–354. https://doi.org/10.1016/S1074-7613(03)00051-7

Qian, W., Yang, J. R., Pearson, N. M., Maclean, C., & Zhang, J. (2012). Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genetics*, *8*(3). https://doi.org/10.1371/journal.pgen.1002603

Quax, T. E. F., Claassens, N. J., Söll, D., & van der Oost, J. (2015, July 16). Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell*. Cell Press. https://doi.org/10.1016/j.molcel.2015.05.035

Riba, A., Nanni, N. Di, Mittal, N., Arhné, E., Schmidt, A., & Zavolan, M. (2019). Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(30), 15023–15032. https://doi.org/10.1073/pnas.1817299116

Savisaar, R., & Hurst, L. D. (2018). Exonic splice regulation imposes strong selection at synonymous sites. *Genome Research*, *28*(10), 1442–1454. https://doi.org/10.1101/gr.233999.117

Schmitt, B. M., Rudolph, K. L. M., Karagianni, P., Fonseca, N. A., White, R. J., Talianidis, I., Odom, D. T., Marioni, J. C., & Kutter, C. (2014). High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA-tRNA interface. *Genome Research*, *24*(11), 1797–1807. https://doi.org/10.1101/gr.176784.114

Schmitz, A., & Zhang, F. (2021). Massively parallel gene expression variation measurement of a synonymous codon library. *BMC Genomics*, *22*(1), 1–12. https://doi.org/10.1186/s12864-021-07462-z

Schwanhüusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., … Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, *473*(7347), 337–342. https://doi.org/10.1038/nature10098

Shah, P., Ding, Y., Niemczyk, M., Kudla, G., & Plotkin, J. B. (2013). Rate-limiting steps in yeast protein translation. *Cell*, *153*(7), 1589. https://doi.org/10.1016/j.cell.2013.05.049

Sharp, P. M., & Li, W. H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution*, *24*(1–2), 28–38. https://doi.org/10.1007/BF02099948

Shevchenko, A., Tomas, H., Havliš, J., Olsen, J. V., & Mann, M. (2007). In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nature Protocols*, *1*(6), 2856–2860. https://doi.org/10.1038/nprot.2006.468

Shin, J. B., Krey, J. F., Hassan, A., Metlagel, Z., Tauscher, A. N., Pagana, J. M., … Barr-Gillespie, P. G. (2013). Molecular architecture of the chick vestibular hair bundle. *Nature Neuroscience*, *16*(3), 365–374. https://doi.org/10.1038/nn.3312

Stetson, D. B., & Medzhitov, R. (2006). Recognition of cytosolic DNA activates an IRF3-dependent innate immune response. *Immunity*, *24*(1), 93–103. https://doi.org/10.1016/j.immuni.2005.12.003

Stingele, S., Stoehr, G., Peplowska, K., Cox, J., Mann, M., & Storchova, Z. (2012). Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Molecular Systems Biology*, *8*(608). https://doi.org/10.1038/msb.2012.40

Stoletzki, N., & Eyre-Walker, A. (2007). Synonymous codon usage in Escherichia coli: Selection for translational accuracy. *Molecular Biology and Evolution*, *24*(2), 374–381. https://doi.org/10.1093/molbev/msl166

Tuller, T., Waldman, Y. Y., Kupiec, M., & Ruppin, E. (2010). Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(8), 3645–3650. https://doi.org/10.1073/pnas.0909910107

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., & Cox, J. (2016). The Perseus computational platform for comprehensive analysis of proteomics data. *Nature Methods*, *13*(9), 731–740. https://doi.org/10.1038/nmeth.3901

Urrutia, A. O., & Hurst, L. D. (2001). Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics*, *159*(3), 1191–1199. https://doi.org/10.1093/genetics/159.3.1191

Verma, M., Choi, J., Cottrell, K. A., Lavagnino, Z., Thomas, E. N., Pavlovic-Djuranovic, S., … Djuranovic, S. (2019). A short translational ramp determines the efficiency of protein synthesis. *Nature Communications*, *10*(1), 1–15. https://doi.org/10.1038/s41467-019-13810-1

Vogel, C., & Marcotte, E. M. (2008). Calculating absolute and relative protein abundance from mass spectrometry-based protein expression data. *Nature Protocols*, *3*(9), 1444–1451. https://doi.org/10.1038/nport.2008.132

Vogel, C., De Sousa Abreu, R., Ko, D., Le, S. Y., Shapiro, B. A., Burns, S. C., … Penalva, L. O. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular Systems Biology*, *6*(400), 1–9. https://doi.org/10.1038/msb.2010.59

Vogel, C., & Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, *13*(4), 227–232. https://doi.org/10.1038/nrg3185

Walsh, I. M., Bowman, M. A., Soto Santarriaga, I. F., Rodriguez, A., & Clark, P. L. (2020). Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(7), 3528–3534. https://doi.org/10.1073/pnas.1907126117

Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D. P., … Kuster, B. (2019). A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular Systems Biology*, *15*(2), 1–16. https://doi.org/10.15252/msb.20188503

Wang, S. E., Brooks, A. E. S., Poole, A. M., & Simoes-Barbosa, A. (2020). Determinants of translation efficiency in the evolutionarily-divergent protist Trichomonas vaginalis. *BMC Molecular and Cell Biology*, *21*(1), 1–13. https://doi.org/10.1186/s12860-020-00297-8

Wang, Z. Y., Leushkin, E., Liechti, A., Ovchinnikova, S., Mößinger, K., Brüning, T., … Kaessmann, H. (2020). Transcriptome and translatome co-evolution in mammals. *Nature*, *588*(7839), 642–647. https://doi.org/10.1038/s41586-020-2899-z

Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B., & Bartel, D. P. (2016). Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Reports*, *14*(7), 1787–1799. https://doi.org/10.1016/j.celrep.2016.01.043

Welch, M., Govindarajan, S., Ness, J. E., Villalobos, A., Gurney, A., Minshull, J., & Gustafsson, C. (2009). Design parameters to control synthetic gene expression in Escherichia coli. *PLoS ONE*, *4*(9). https://doi.org/10.1371/journal.pone.0007002

Wu, J., & Chen, Z. J. (2014). Innate immune sensing and signaling of cytosolic nucleic acids. *Annual Review of Immunology*, *32*(March), 461–488. https://doi.org/10.1146/annurev-immunol-032713-120156

Yona, A. H., Bloom-Ackermann, Z., Frumkin, I., Hanson-Smith, V., Charpak-Amikam, Y., Feng, Q., Boeke, J. D., Dahan, O., & Pilpel, Y. (2013). Trna genes rapidly change in evolution to meet novel translational demands. *ELife*, *2013*(2), 1–17. https://doi.org/10.7554/eLife.01339

Yu, C. H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M. S., & Liu, Y. (2015). Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Molecular Cell*, *59*(5), 744–754. https://doi.org/10.1016/j.molcel.2015.07.018

Zhu, J., Zhang, Y., Ghosh, A., Cuevas, R. A., Forero, A., Dhar, J., Ibsen, M. S., Schmid-Burgk, J. L., Schmidt, T., Ganapathiraju, M. K., Fujita, T., Hartmann, R., Barik, S., Hornung, V., Coyne, C. B., & Sarkar, S. N. (2014). Antiviral Activity of Human OASL Protein Is Mediated by Enhancing Signaling of the RIG-I RNA Sensor. *Immunity*, *40*(6), 936–948. https://doi.org/10.1016/j.immuni.2014.05.007

Zucchelli, E., Pema, M., Stornaiuolo, A., Piovan, C., Scavullo, C., Giuliani, E., … Bovolenta, C. (2017). Codon Optimization Leads to Functional Impairment of RD114-TR Envelope Glycoprotein.

Molecular Therapy - Methods and Clinical Development, *4*(March), 102–114. https://doi.org/10.1016/j.omtm.2017.01.002

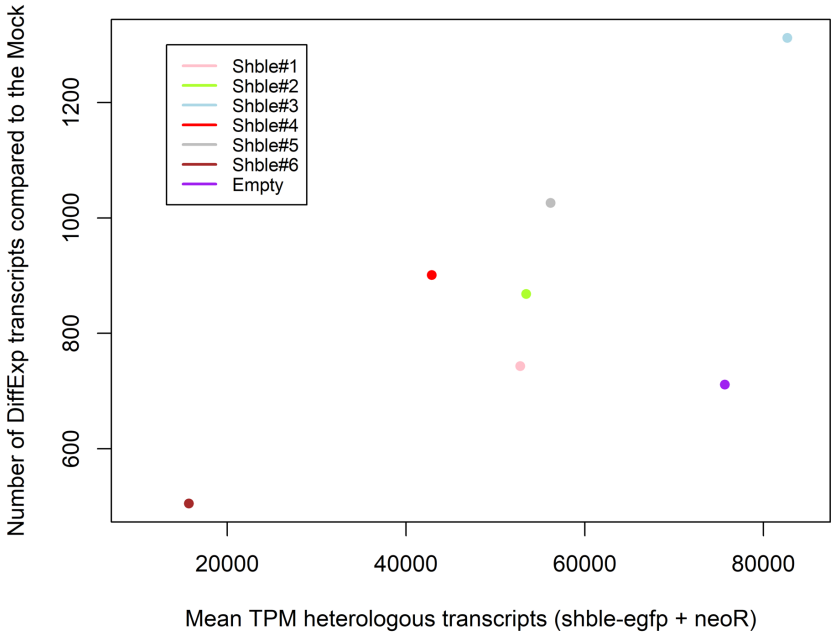

Figure 1. Panel A. Link between the amount of differentially expressed cellular transcripts relatively to the Mock and the mean quantity of heterologous transcripts produced. For each condition the mean expression of heterologous mRNAs (*shble-egfp* transcript + *neoR* transcript) is the average obtained from the n=3 transfection replicates. Expression levels are given as Transcripts Per Million (TPM). Samples are color-coded as follows: Shble#1 in pink, Shble#2 in green, Shble#3 in light blue, Shble#4 in red, Shble#5 in grey, Shble#6 in brown and Empty in purple.

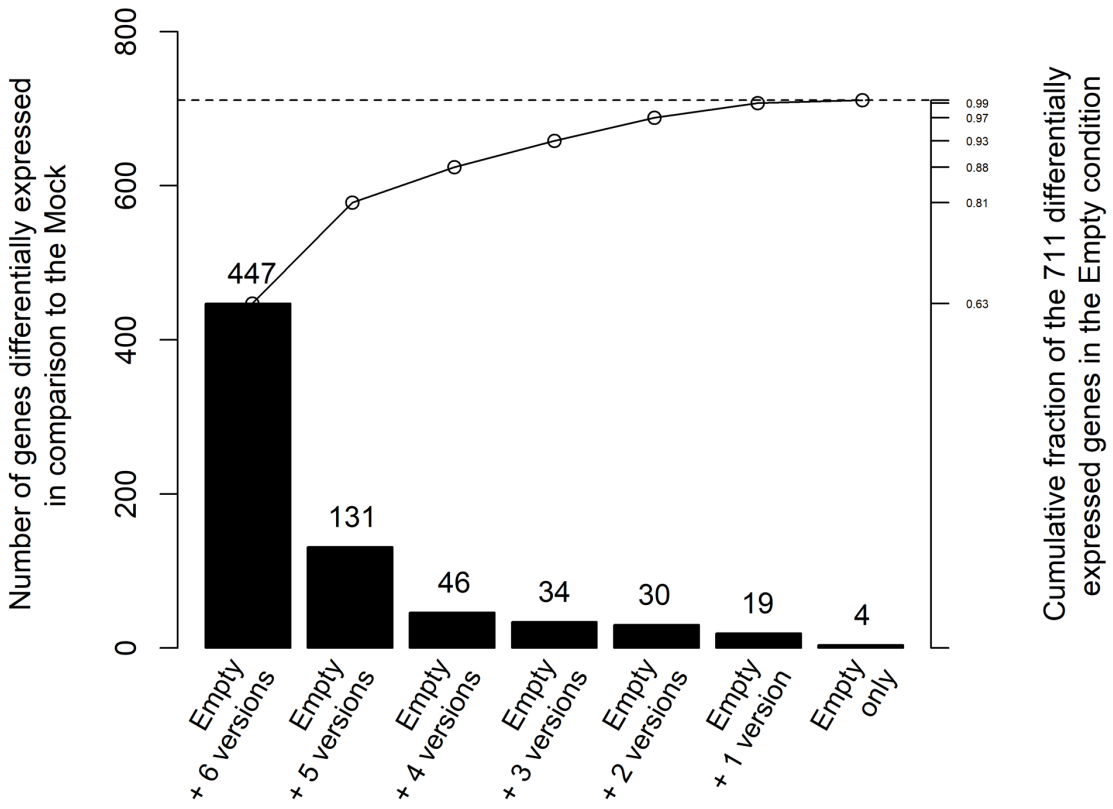

Figure 1. Panel B. Match between differentially expressed (DiffExp) transcripts in the Empty control and DiffExp transcripts in the six *shble_egfp* synonymous versions. Each of the 711 DiffExp transcripts in the Empty condition were assigned to a category depending on the number of other conditions this gene was found to be DiffExp. Bar heights correspond to the number of genes identified in each category (left axis). In all cases the transcripts are labelled as DiffExp with respect to the mock control condition. The secondary axis on the right reports the cumulative fraction of the 711 DiffExp transcripts in the Empty control condition as each category is added to the previous one. Data should be read as follows, using "Emtpy + 5 versions" as an example: among the 711 transcripts DiffExp between the Empty version and the mock, a total of 447+131 transcripts (*i.e.,* 81% of the 711 transcripts) are also DiffExp in five or more of the *Shble* synonymous conditions

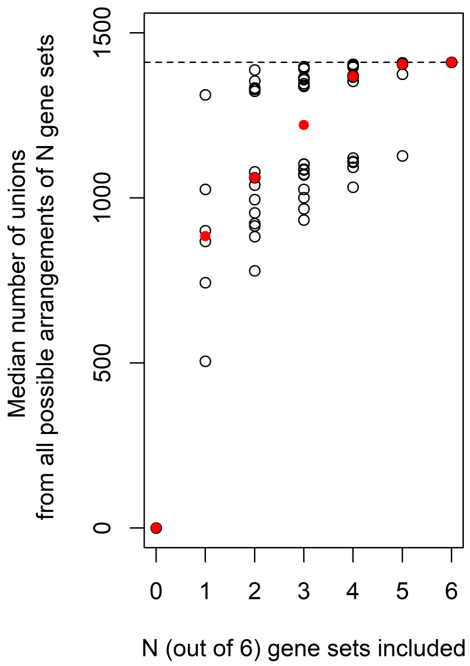

Figure 1. Panel C. Number of differentially expressed (DiffExp) transcripts identified in cells transfected with different *shble_egfp* synonymous versions. Values along the x-axis represent how many out of the six sets of DiffExp transcripts (corresponding to conditions Shble#1 to Shble#6) were included, and values along the y-axis represent the number of unique DiffExp genes obtained across these sets. The graph should be read as follows, using the value x=3: for each of the twenty different combinations of three data sets sampled among the six *shble_egfp* conditions, the number of unique DiffExp transcripts are shown in open black circles, while the median of these values is plotted as a red dot (median value of y=1220.5 across the 20 combinations in this example). The horizontal dash line corresponds to the full universe of 1,411 DiffExp genes identified when all six sets of DiffExp genes are included and represents the upper limit of genes that can be identified with sub-samples of the six genes sets. Hence, by combining only half of the six conditions (x=3), a large proportion (86%) of unique genes detected when all six conditions are included is already recapitulated.

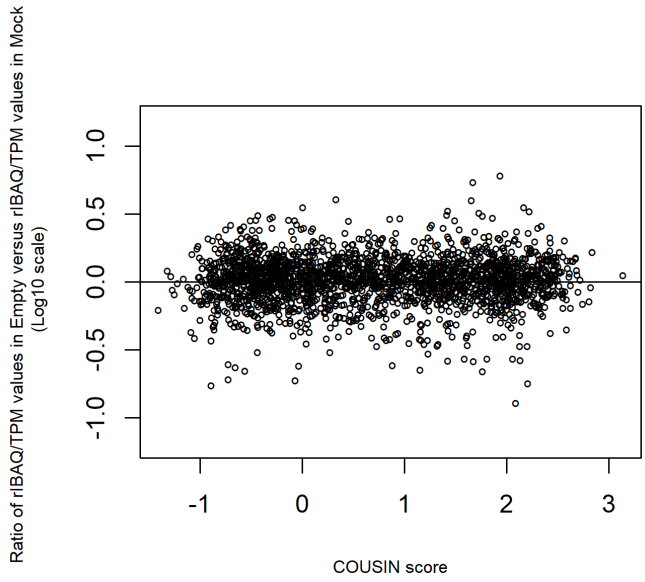

Figure 2. Panel A. Protein-to-RNA ratios in the Empty condition compared to the Mock control condition as a function of individual gene's codon usage bias (CUBias). Transcript levels were estimated as Transcripts Per Million (TPM), while protein levels were estimated as relative intensity-Based Absolute Quantification (riBAQ) values. Each dot represents, for one gene, the averaged $[riBAQ/TPM]_{Empty}$ normalized by the $[riBAQ/TPM]_{Mock}$, calculated from the n=3 transfection replicates. This ratio of ratios reflects the extent to which cellular transfection and heterologous expression of a gene with an over-humanized sequence (*egfp*, encoded in the Empty condition) affects translation efficiency of cellular transcripts, and it has been calculated for the 2,471 genes for which both proteomic and transcriptomic values were available. For visual purposes, the horizontal line centered on the y=0 value (log scale) is shown. The x-axis displays the match between the CUBias of each individual gene to that of the human genome average, calculated using the COUSIN index: negative scores correspond to genes with CUBias opposite to the human average while scores above one correspond to genes with CUBias similar in direction but of stronger intensity than the human average. No link was found between gene's $[riBAQ/TPM]_{Empty}$ normalized by the $[riBAQ/TPM]_{Mock}$ and gene's CUBias (linear regression slope value = 0.00378, F-test P = 0.29).

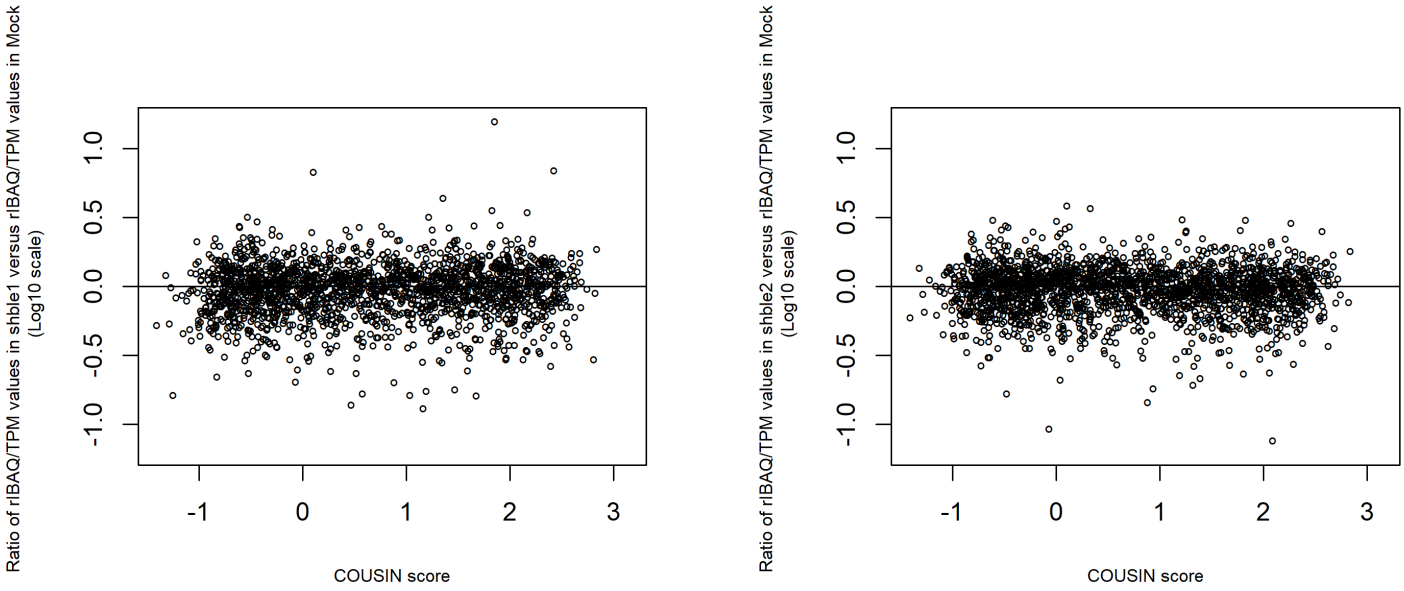

Figure 2. Panel B. Similar as panel A, but for the Shble#1 condition compared to the mock control condition. The *shble* gene in the Shble#1 version had been synonymously recoded using systematically for each amino acid the most used among the synonymous codons in the human genome average. It corresponds thus to an overhumanized heterologous gene (COUSIN value=3.47). No link was found between gene's $[riBAQ/TPM]_{Shble\#1}$ normalized by the $[riBAQ/TPM]_{Mock}$ and gene's CUBias (linear regression slope value = 0.0048, F-test P = 0.18).

Figure 2. Panel C. Similar as panel A, but for the Shble#2 condition compared to the mock control condition. The *shble* gene in the Shble#2 version had been synonymously recoded using systematically for each amino acid the GC-richest among the two most used synonymous codons in the human genome average (COUSIN value=3.42). It differs only in eight codons from the Shble#1 version. No link was found between gene's $[riBAQ/TPM]_{Shble\#2}$ normalized by the $[riBAQ/TPM]_{Mock}$ and gene's CUBias (linear regression slope value = -0.0059, F-test P = 0.11).

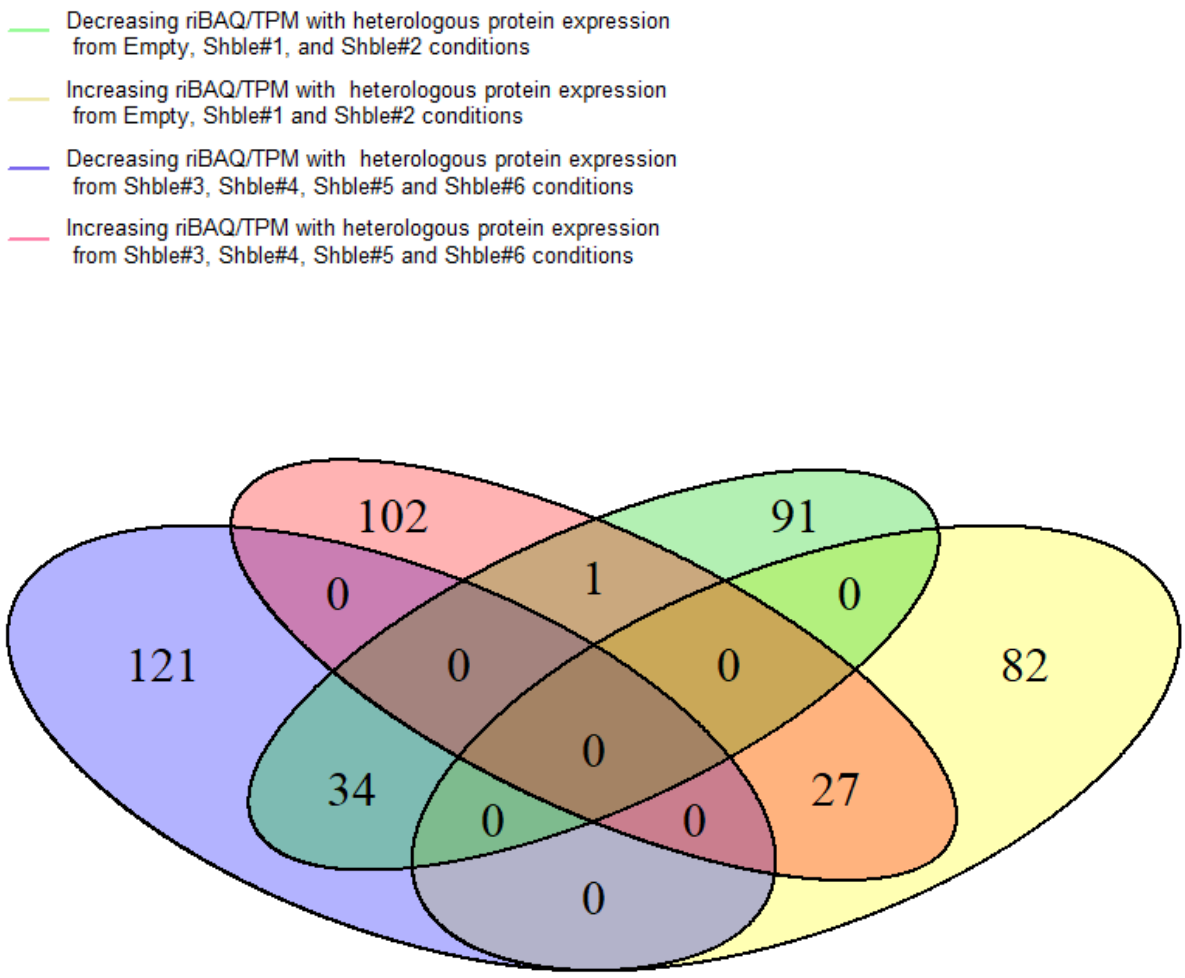

Figure 3. Panel A. Categories of genes according to their variation in riBAQ/TPM ratios with varying heterologous protein expression across conditions. The ratio riBAQ/TPM was taken as a proxy for the translation efficiency of a given transcript. The green set corresponds to cellular genes displaying a decreasing translation efficiency as heterologous protein expression under the conditions Empty, Shble#1 and Shble#2 increases (*i.e.*, when overexpressing humanized heterologous genes), while the yellow set corresponds to cellular genes displaying an increasing translation efficiency under these same conditions. The blue set corresponds to cellular genes displaying a decreasing translation efficiency as heterologous protein expression under the conditions Shble#3, Shble#4, Shble#5 and Shble#6 increases (*i.e.,* when overexpressing non-humanized heterologous genes), while the pink set corresponds to cellular genes displaying an increasing translation efficiency under these same conditions. Note that, by construction, neither the green and yellow sets nor the blue and pink sets overlap.

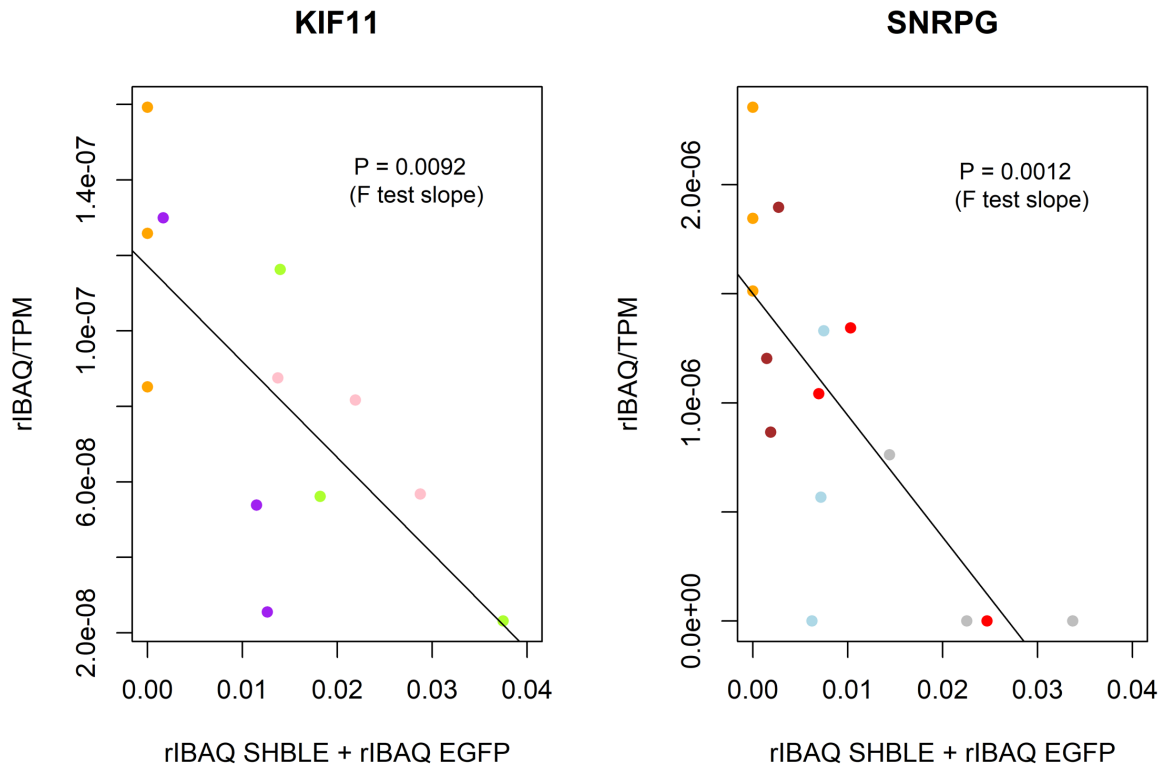

Figure 3. Panel B. Example of two cellular genes displaying significant variation in their riBAQ/TPM ratio as a function of heterologous protein expression levels across samples. The ratio riBAQ/TPM was taken as a proxy for the translation efficiency of a given transcript. KIF11 belongs to the green set shown in panel A (negative association with increasing amount of protein expressed from over-humanized heterologous genes) and SNRPG belongs to the blue set (negative association with increasing amount of protein expressed from non-humanized heterologous genes). Samples are color-coded according to the transfected construct: Shble#1 in pink, Shble#2 in green, Shble#3 in light blue, Shble#4 in red, Shble#5 in grey, Shble#6 in brown, Empty in purple and Mock in orange. P-value of the F-test testing the significance of the regression is indicated.

Fig. 3C

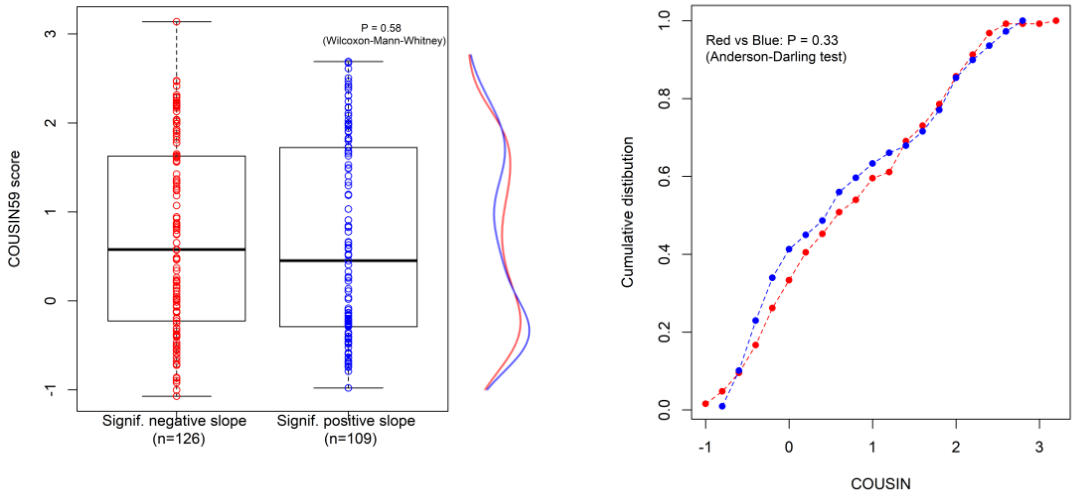

Fig. 3D

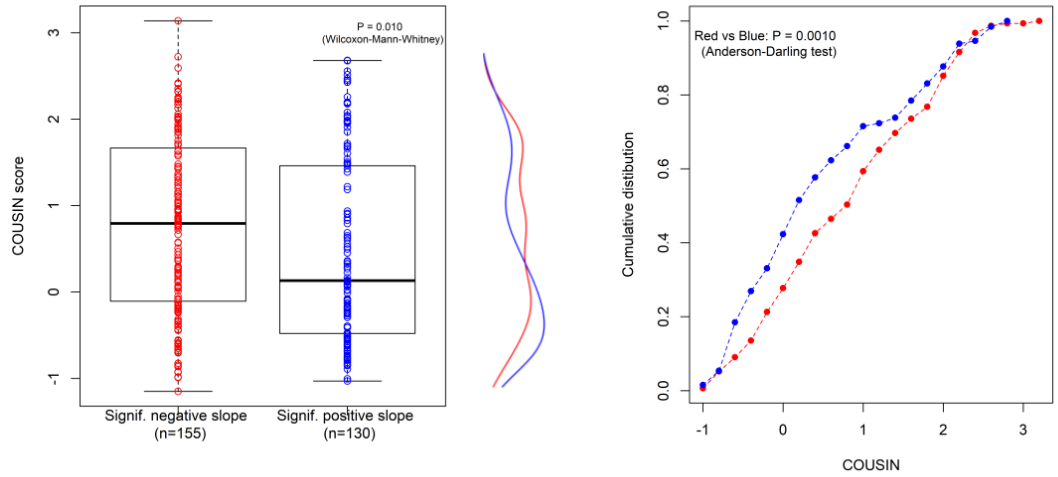

Figure 3. Panel C. Left: COUSIN scores for cellular genes displaying either negative (red) or positive (blue) changes in the riBAQ/TPM ratio values under increasing expression of over-humanized genes (Empty, Shble#1 and Shble#2). Below each boxplot the underlying number of genes included is shown. Right: Cumulative distribution of COUSIN values for the two sets of genes described in the left. P-values for an Anderson Darling test contrasting the two distributions is indicated.

Figure 3. Panel D. Left: COUSIN scores for cellular genes displaying either negative (red) or positive (blue) changes in the riBAQ/TPM ratio values under increasing expression of non-humanized genes (Shble#3, Shble#4, Shble#5 and Shble#6). Below each boxplot the underlying number of genes included is shown. Right: Cumulative distribution of COUSIN values for the two sets of genes described in the left. P-values for an Anderson Darling test contrasting the two distributions is indicated.

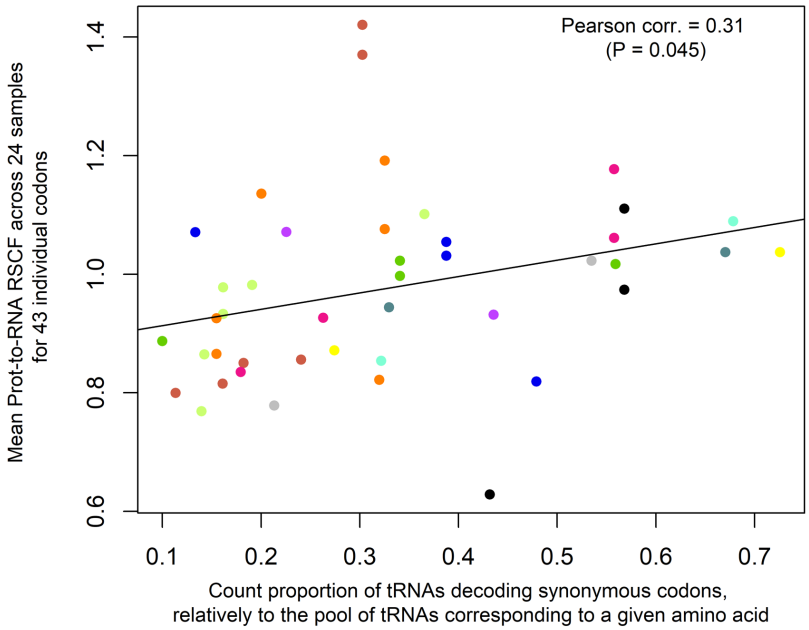

Figure 4. Panel A. Covariation between Prot-to-RNA relative synonymous codon frequency (RSCF) and relative tRNA content (per amino acid). Values on the x-axis were calculated on a per amino acid basis and represent the count of tRNAs bearing a given anticodon, normalized by the total count of all tRNAs decoding synonymous codons that encode the considered amino acid. tRNAs counts were obtained from **Mattijssen et al, 2017**. Values on the y-axis have been calculated by averaging Prot-to-RNA RSCF values across all 24 samples. 43 codons-anticodons pairs, corresponding to 12 amino acids, are included in this analysis (Leu, Arg, Ser, Val, Ala, Gly, Ile, Lys, Glu, Gln, Thr, Pro). These amino acids are coloured as follows: light green for Leu, brown for Arg, orange for Ser, dark green for Val, blue for Ala, deep pink for Gly, black for Ile, aquamarine for Lys, blue green for Glu, yellow for Gln, Grey for Thr and purple for Pro. The overall Pearson correlation between averaged Prot-to-RNA RSCF values and relative tRNA content is slightly significant and the regression line is shown. For ten out of the twelve amino acids for which it was possible to conduct this analysis, a positive trend between codons Prot-to-RNA RSCF values and the relative amount of tRNA that decode them was found (*i.e.,* for all twelve amino acids except Ala – blue – and Pro – purple). Note that by construction it is expected that some dots drawn with the same colour display similar x-axis values: this is the case when several synonymous codons are decoded by the same tRNA-anticodon (see Methods).

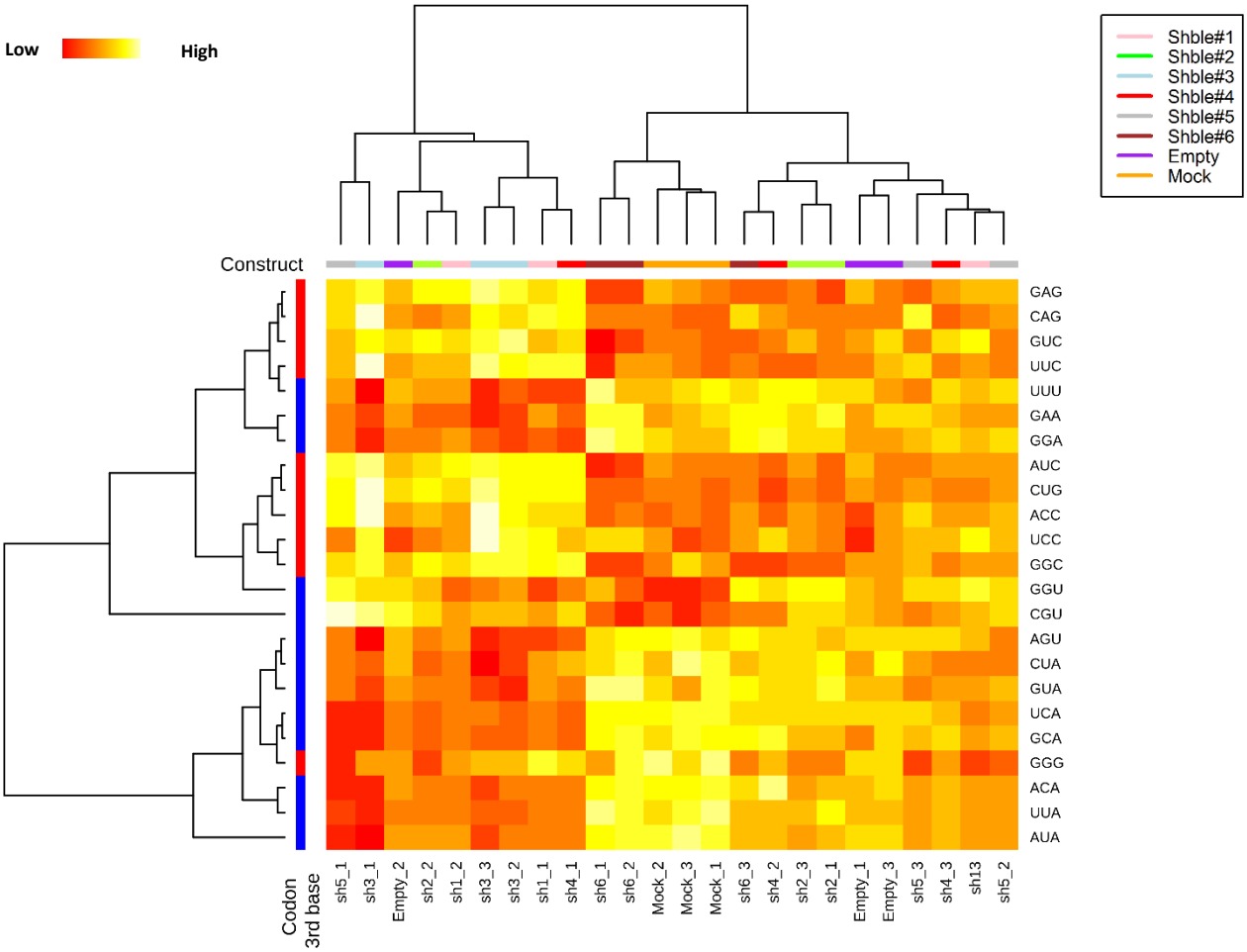

Figure 4. Panel B. Heatmap of Prot-to-RNA relative synonymous frequency (RSCF) profiles based on codons displaying a significant construct effect on their translation efficiency. The 23 codons for which one-way ANOVAs yielded a significant construct effect (eight modalities: Shble#1 to Shble#6 plus the Empty and Mock control conditions) on Prot-to-RNA RSCF values are included and displayed in rows. The vertical color bar on the left indicates whether codons are A or U-ending (blue) or G or C-ending (red). All 24 samples are shown in columns, with the horizontal color bar indicating the corresponding experimental conditions (Shble#1 to Shble#6, Empty and Mock). Samples are color-coded according to the transfected construct: Shble#1 in pink, Shble#2 in green, Shble#3 in light blue, Shble#4 in red, Shble#5 in grey, Shble#6 in brown, Empty in purple and Mock in orange. Heatmap color intensity corresponds to the Z-score, after a per codon (row) Z-normalization. Note that samples of the Shble#3 condition display high Prot-to-RNA RSCF values for GC-ending codons (red) and low values for AT-ending codons (blue).

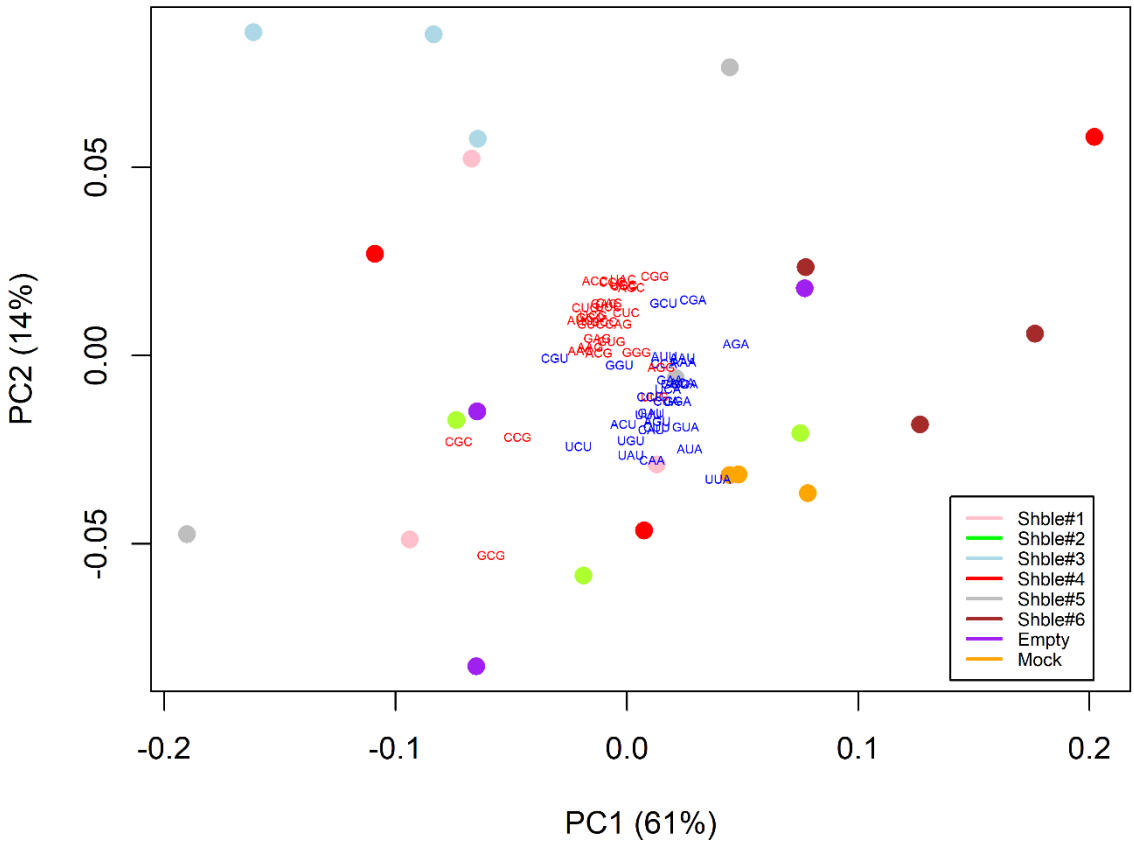

Figure 4. Panel C. Inter-sample variation of Prot-to-RNA relative synonymous codon frequency (RSCF) profiles considering all 59 amino acid encoding codons. Contrary to Panel B, the principal component analysis – of which PC1 and PC2 are shown – has been constructed using Prot-to-RNA RSCF data from all the 59 amino acids. Eigenvectors loads of the covariance matrix are superimposed to the PCA graph and are colored according to the nucleotide composition at the codon 3rd position (blue for A or U-ending codons, red for G or C-ending codons). The 24 samples are color-coded according to the transfected construct: Shble#1 in pink, Shble#2 in green, Shble#3 in light blue, Shble#4 in red, Shble#5 in grey, Shble#6 in brown, Empty in Purple and Mock in orange. Values in parentheses represent the fraction of the total variance captured by the corresponding axis.

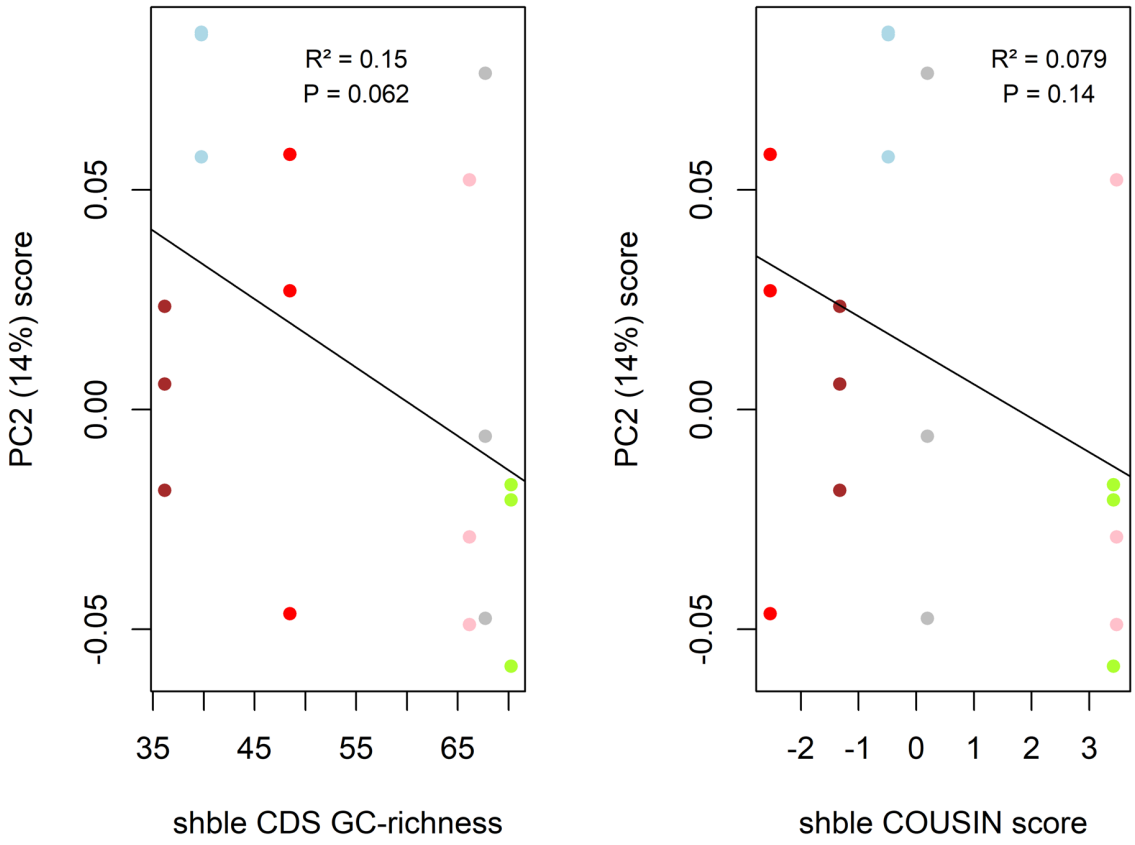

Figure 4. Panel D. Link between base composition or CUBias of the *shble* gene with preferentially translated codons. The y-axis corresponds in both cases to sample scores projected onto the PC2 axis of the PCA of Panel C (capturing 14% of the total variance in RSCF). Only the 18 experimental synonymous *shble_egfp* conditions are included. Samples are color-coded according to the transfected construct: Shble#1 in pink, Shble#2 in green, Shble#3 in light blue, Shble#4 in red, Shble#5 in grey and Shble#6 in brown. On the left, the x-axis indicates the overall GC-richness of each synonymous version and on the right it indicates the value of the COUSIN index for the different versions. Negative scores COSUIN values correspond to genes with CUBias opposite to the human average while values above one correspond to genes with CUBias similar in direction but of stronger intensity than the human average. Regression lines are plotted, as well as the P values of the F test for the slope significance and the coefficient of determination (R^2).

List of supplementary tables

Table_S1.xlsx

Table_S2.xlsx

Table_S3.xlsx

Table_S4.xlsx

Table_S5.xlsx

Table_S6.xlsx

Table_S7.xlsx

Table_S8.xlsx

Table_S9.xlsx

List of supplementary figures

All supplementary figures are grouped in a single pdf file