Accessibility of the Shine–Dalgarno sequence dictates N-terminal codon bias in *E. coli*

Sanchari Bhattacharyya[1,*], William M. Jacobs[1,*], Bharat V. Adkar[1], Jin Yan[1,2], Wenli Zhang[1,3], & Eugene I. Shakhnovich[1,#]

[1]Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford St., Cambridge MA, USA

[2]College of Chemical Engineering, Sichuan University, Chengdu, Sichuan, China

[3]State Key Laboratory of Food Science and Technology, Jiangnan University, Wuxi, China

*These authors contributed equally to the work

[#]Correspondence should be addressed to E.S. (shakhnovich@chemistry.harvard.edu)

Abstract

Despite considerable effort, no physical mechanism has been experimentally shown to explain the N-terminal codon bias in prokaryotic genomes. Using a systematic study of synonymous substitutions in two endogenous *E. coli* genes, we show that interactions between the coding region and the upstream Shine–Dalgarno sequence modulate both the rate of mRNA degradation and the efficiency of translation-initiation via a 'ribosome-protection' mechanism, which ultimately affects cellular fitness. We further demonstrate that occlusion of the Shine–Dalgarno sequence depends on the formation of non-equilibrium secondary structures, which may include long-range contacts, highlighting the role of co-transcriptional mRNA folding and the influence of the 5'-untranslated region in determining the optimal coding sequence. Finally, a statistical analysis of the *E. coli* genome also specifically implicates avoidance of intra-molecular base-pairing with the Shine–Dalgarno sequence. Our results therefore provide general physical insights into the coding-level sequence features that optimize protein expression in prokaryotes.

Introduction

The genetic code is degenerate, such that almost all amino acids are coded by more than one codon. However, on a genome-wide scale, not all synonymous codons are used with the same frequency. This codon-usage bias varies across different organisms (Grantham et al., 1980a; Grantham et al., 1980b) and is reflected in the overall GC content of a genome (Lynn et al., 2002; Muto and Osawa, 1987). Yet within a particular genome, synonymous codons are not distributed randomly. For example, codons that are under-represented in highly expressed genes tend to be translated more slowly and with slightly less accuracy (Bulmer, 1991; Eyre-Walker and Bulmer, 1993; Sharp and Li, 1987), suggesting that essential genes have been optimized for translational efficiency. Even within a gene, rare codons often appear in conserved clusters (Chaney and Clark, 2015; Chaney et al., 2017; Clarke and Clark, 2010; Jacobs and Shakhnovich, 2017), indicating that, in some cases, rare codons might also be under positive selection.

In most prokaryotic and eukaryotic genes, the first ~30 amino acids are more likely to be encoded by rare codons (Eyre-Walker and Bulmer, 1993; Tuller et al., 2010). The origin of this 'N-terminal codon bias' is debatable, with various mechanisms put forward in different studies. One hypothesis is that rare codons have been selected to reduce the speed of translation at the beginning of genes (Tuller et al., 2010). Such a 'ramp' in elongation might reduce the likelihood of ribosomal traffic jams along a transcript, preventing premature termination. However, other studies have implicated RNA secondary structure near the 5'-end of mRNA transcripts. Based on synthetic libraries of 154 synonymous variants of GFP (Kudla et al., 2009) and more than 14,000 gene fragments fused to GFP (Goodman et al., 2013), the folding free energy of truncated mRNA transcripts was found to be the best predictor of the rate of protein production. In these studies, codon rarity did not appear to have a measurable effect on protein expression. This idea has also been supported by an informatics analysis of several bacterial genomes (Bentele et al., 2013), which showed that rare codons correlate with the formation of weaker mRNA secondary structures in GC-rich genomes. However, a contradictory study of synonymous variants of two human genes expressed in *E. coli* (Welch et al., 2009) found no statistically significant correlation between protein expression and mRNA folding stability.

A key limitation of these previous experimental studies is their reliance on non-endogenous genes. For example, biases in the selection of promoters and N-terminal regions in the synthetic libraries might give rise to spurious correlations while concealing more fundamental physical reasons for the variability of expression among synonymous variants. The use of randomized libraries, with substitutions confined to the 5'-end, also hinders our ability to distinguish between competing proposals. For these reasons, a hypothesis-driven experimental study is needed to understand the role of N-terminal codon bias.

In this paper, we report a systematic study of synonymous substitutions in two endogenous *E. coli* genes: *folA*, which codes for the low-copy essential protein Dihydrofolate Reductase (DHFR; ~40 copies/cell), and *adk*, which codes for the more abundant essential protein Adenylate Kinase (ADK; ~700 copies/cell) (Taniguchi et al., 2010). We find that codon selection near the N-terminus is the primary determinant of not only protein expression, but also the steady-state mRNA levels. In particular, the first rare codon appears to play a crucial role in both *folA* and *adk*. However, the influence of synonymous substitutions at this position is not explained by either codon rarity or GC content. We instead show that occlusion of the upstream Shine–Dalgarno sequence, which may form secondary

structures with bases throughout the coding region of a transcript, simultaneously reduces protein expression and increases the rate of mRNA degradation. In general, this effect cannot be predicted with equilibrium mRNA folding calculations, but rather requires attention to non-equilibrium secondary structures that often involve long-range contacts. A complementary statistical analysis of the *E. coli* genome reveals that genes have evolved to minimize binding between coding regions and their respective Shine–Dalgarno sequences. Consequently, the mechanism that we propose appears to be a universal and essential determinant of synonymous codon bias in *E. coli*.

Results

Synonymous mutations affect both mRNA and protein levels in *E. coli*

We initially introduced synonymous substitutions directly on the chromosomal copy of *folA* in *E.coli* in order to observe any fitness and proteomic effects in their endogenous context. First, we designed a fully optimized synonymous variant ('MutAll') of the *folA* gene, in which all codons were replaced by the most common synonymous codon in the wild-type (WT) sequence (Figure 1A). Contrary to our expectations, we observed a significant decrease in the exponential-phase growth rate of the *E. coli* population and an increase in the lag time, defined as the time required to achieve the maximum growth rate (Adkar et al., 2017) (Figure 1B). We subsequently determined that these mutations cause a 20-fold decrease in the amount of intra-cellular soluble protein, measured by quantitative Western blot, and a 4-fold drop in intra-cellular mRNA levels, measured via quantitative PCR (qPCR) (Figure 1C). These measurements indicate that, for this gene, codon 'optimization' reduces cellular fitness by lowering DHFR expression.

Next, we determined that the decreases in soluble-protein and mRNA levels can be attributed almost entirely to the optimization of the 15 'rare' codons, i.e., those codons that occur substantially less frequently than their synonymous alternatives in the most abundant *E. coli* genes, in the *folA* gene (`MutRare'; Figure 1A,C). This observation is surprising in light of several recent studies, which found that, in general, rare codons have either deleterious or neutral effects on mRNA abundances in *E. coli* (Boel et al., 2016), yeast (Presnyak et al., 2015) and a filamentous fungus (Tokuoka et al., 2008). Nevertheless, the significantly reduced fitness of the MutRare construct in terms of both growth rate and lag times (Figure 1B) strongly suggests that the presence of rare codons in the WT *folA* gene is not simply a result of neutral drift.

Codon usage does not affect transcription

Suspecting that variations in the mRNA levels are integral to changes in protein expression, we sought to determine the source of mRNA-level variations in our initial *folA* variants. The effect of synonymous codons on mRNA levels can arise in two ways, either via differential production (i.e., during transcription) or via differential degradation. A recent study on *Neurospora crassa* concluded that the presence of optimal codons increases mRNA levels mainly through effects on transcription (Zhou et al., 2016). To investigate this possibility, we carried out *in vitro* transcription with T7 RNA polymerase by fusing the coding region of 'MutAll' and 'MutRare' sequences to the bacteriophage T7 promoter. Although the mRNA levels obtained for the partially and fully optimized constructs were found to be marginally lower than the WT *folA* coding sequence on this promoter (Figure 2), the observed

differences do not account for the ~80% drop in mRNA levels observed *in vivo*. Consequently, we reasoned that the *in vivo* mRNA-level variations are more likely to be a result of differential mRNA degradation.

Systematic codon replacements point to a crucial role of the first rare codon

We next set out to dissect the effects of individual codons on the mRNA levels of *folA* and *adk*. In order to perform a systematic investigation, each gene was cloned under an arabinose-inducible pBAD promoter, which allows modulation of expression from 10 to 1000-fold over endogenous levels (Bhattacharyya et al., 2016). Using a fixed arabinose concentration of 0.05%, we observed 13-fold and 9-fold reductions in the mRNA abundances of the MutAll and MutRare *folA* constructs on the plasmid, respectively (Figure 3A). In an attempt to further narrow down the search, we tried optimizing only the eight rare codons that are also present in the *folA* gene of the bacterium *Citrobacter koseri*, which shares 96% sequence similarity with the *E.coli folA* gene ('MutSelect'). MutSelect also showed an 8-fold drop in mRNA levels, suggesting that this subset of rare codons is most critical. The majority of this effect could, in turn, be attributed to substitutions of the first four rare codons ('MutN-term') at positions 3, 8, 28, and 36. Finally, we determined that optimization of the first rare codon at position 3 ('3AGC') has the single greatest contribution to the decrease in mRNA levels. Maintaining only the first four rare codons while optimizing the remaining rare positions ('MutC-term') kept the mRNA abundance close to the WT level.

To determine if these trends are generalizable beyond *folA*, we turned to the *E. coli* gene *adk*. Again, optimization of all six WT rare codons (MutRare) profoundly lowered the mRNA levels, with the dominant contribution coming from the first rare codon at position 6 of *adk* (Figure 3B). Consistent with our findings in *folA*, mutations in the last three rare codons of *adk*, at positions 83, 175 and 213, did not contribute to the decrease in mRNA levels. Nevertheless, while the N-terminal rare codons were found to play the most crucial role in both genes, far-downstream mutations did result in measurable effects. For example, variations in mRNA levels due to the optimization of select downstream rare codons in *folA* are shown in Figure 3C.

The effect of the first rare codon is not due to rarity

Having observed that the greatest mRNA-level effects arise from the first rare codon, we investigated whether the rarity of this codon plays any role. To this end, we mutated position 3 of *folA* to all six Serine codons and position 6 of *adk* to all six Leucine codons (Figure 4A,B). We found that, in the case of *folA*, only one codon (TCA) maintained the WT mRNA levels, while in the case of *adk*, two codons (CTA and CTC) actually increased the mRNA abundance above the WT level. All other choices resulted in significant reductions in the mRNA levels. Notably, these effects do not correlate with codon rarity. We also determined that re-introduction of the WT codons at these key positions on the MutRare backgrounds ('MutRare+3AGT' and `MutRare+6CTT') resulted in mRNA levels that differ significantly from the WT (Figure 4C,D). Meanwhile, the effects of optimizing these key codons were significantly altered by synonymous substitutions at adjacent positions (the second and fourth codons of *folA* and the fifth codon of *adk*). These experiments emphasize that the effects of synonymous substitutions are dependent on their local sequence contexts.

Variations in mRNA abundances correlate with mRNA stability

We next sought to determine whether changes in the measured mRNA levels correlate with the predicted stabilities of the mRNA transcripts. We quickly ruled out variations in the GC content as a plausible explanation (Figure 4A,B). Although there is no unique definition of mRNA folding stability, one common approach for examining sequence-specific effects near the 5'-end is to compute the folding free energy of the first L bases of the transcript, which we refer to as $\Delta G_{1,L}$ (Bentele et al., 2013; Goodman et al., 2013; Kudla et al., 2009). Physically, this quantity represents the free-energy difference between a random coil and the ensemble of all possible base-paired configurations, assuming that the 5'-end does not interact with bases beyond the Lth position (Figure 5A). In agreement with the work of Goodman et al. (Goodman et al., 2013) and Kudla et al. (Kudla et al., 2009), we find that $\Delta\Delta G_{1,L} \equiv \Delta G_{1,L} - \Delta G_{1,L}^{\mathrm{WT}}$ correlates well with the log-ratio of the mRNA levels (Figure 5B; see Methods). For these comparisons we have used our complete data set of optimized variants, augmented by constructs designed to vary $\Delta G_{1,L}$ for both *folA* and *adk* ('VarE' sequences in Table S1).

These calculations suggest that stable mRNA secondary structures correlate with lower steady-state mRNA levels. However, multiple uncertainties associated with this approach make a clear mechanistic explanation difficult. The portion of a transcript that should be considered in the stability definition is unclear, because the correlation that we obtain is nearly independent of the number of bases included in the $\Delta G_{1,L}$ calculation beyond the fifth codon (Figure 5B). At the same time, when we restrict our attention to constructs with substitutions in the first six codons, the maximum correlation is only obtained with $L \gtrsim 20$ codons (Figure S1A). Thus, while mutations at the N-terminus tend to have the largest effects, interactions with far-downstream portions of the coding region cannot be ignored.

Local interactions between the Shine–Dalgarno and coding sequences reduce mRNA levels

To better understand the role of mRNA folding stability in regulating mRNA degradation, we first focused on our synonymous variants with substitutions closest to 5'-end. Following previous studies (Goodman et al., 2013; Voges et al., 2004), we calculated the probability that a given base is not paired in a transcript's equilibrium secondary structure, p_{unbound}. For these calculations, we analyzed the full-length transcripts, although the results are robust with respect to truncation. We then examined the correlation between the log-ratio of p_{unbound}, relative to the WT sequence, and the log-ratio of the mRNA levels. Using a bootstrap analysis to account for outliers (see Methods), we observed a multi-modal distribution of p-values within ~20 bases of the start codon (Figure 5C; see Methods). Interestingly, the dominant peak in this distribution that does not correspond to a coding-region substitution (i.e., away from positions 3 and 6) occurs in the middle of the Shine–Dalgarno (SD) sequence, which is part of the 5'-untranslated region (5'-UTR) of the pBAD promoter (near position -10). Here we find a moderate, yet significant, correlation (Pearson $R = 0.50$, $p = 0.03$) between the measured mRNA levels and p_{unbound} for both *folA* and *adk* (Figure 5D, red and orange points). To further test this idea, we engineered a synonymous variant of *adk* that incorporates an exact anti-SD sequence (CCTCCT) at codon positions 4 and 5, while retaining all native rare codons, and observed a three-fold reduction in the mRNA abundance (Figure 5D, green point).

This analysis suggests that, for synonymous variants with substitutions at the 5'-end, interactions with the SD sequence are primarily responsible for the correlation between the mRNA levels and $\Delta\Delta G_{1,L}$. Nevertheless, this approach fails for constructs with substitutions beyond the first six codons, where p_{unbound} is predicted to be nearly constant but the mRNA levels vary significantly (Figure 5D, blue and gray points). In fact, there is no position in the sequence for which p_{unbound} correlates with the mRNA-level variations of sequences with downstream mutations. It is thus apparent that, while equilibrium secondary-structure calculations provide useful information regarding local interactions with the 5'-UTR, a different approach is needed to understand far-downstream substitutions.

Long-range interactions point to non-equilibrium mRNA folding effects

We therefore introduce an alternative folding free energy to capture the effects of mutations throughout the length of a transcript. To account for all interactions involving the 5'-end, we define

$$\Delta G_{\text{unfold}} \equiv \Delta G_{1,N} - \Delta G_{L,N}, \qquad (1)$$

which eliminates the contributions of helices that do not involve the first L bases and thus comprise $\Delta G_{L,N}$. Equivalently, ΔG_{unfold} is equal to the sum of $\Delta G_{1,L}$ and $\Delta G_{\text{non-local}}$, the excess free-energy due to non-local contacts between the 5' and 3'-ends, like those shown schematically in Figure 5A. This approach is justified by the observation that the equilibrium secondary structures of mRNA transcripts are dominated by contacts between bases that are less than ~30 nucleotides apart in the sequence (Figure S2). Consequently, as shown in Figure 5D, *equilibrium* RNA secondary structure calculations fail to predict any effects of mutations beyond the tenth codon on the intra-molecular base-pairing probability of the SD sequence.

In contrast, $\Delta G_{\text{non-local}}$ captures long-range contacts that may not be part of the minimum-free-energy secondary structure. Because an RNA molecule reaches its equilibrium structure at a rate proportional to L^{-b}, where $b > 1$ (Galzitskaya and Finkelstein, 1998), the folding time can far exceed the transcription time, which scales linearly with the transcript length. Transient contacts between the 5'-UTR and the coding region are likely to occur while a portion of the molecule is out of equilibrium and sampling higher-free-energy secondary structures. Such contacts may therefore sequester the SD sequence on a timescale that is relevant to mRNA degradation. Furthermore, during the process of co-transcriptional folding, the 5'-UTR has opportunities to interact with downstream bases whose binding partners in the minimum-free-energy structure have not yet been synthesized. Mutations that increase the stability of non-local interactions between the 3' and 5'-ends, regardless of whether they contribute to the lowest-free-energy secondary structure, result in a more negative value of $\Delta G_{\text{non-local}}$. Consequently, even if the equilibrium p_{unbound} of the SD region is unaffected, $\Delta\Delta G_{\text{non-local}}$ correlates with the roughness of the 5'-ends's folding landscape and indicates whether *non-equilibrium* secondary structures involving the SD sequence are likely to form.

Consistent with these expectations, we find that the correlation between $\Delta\Delta G_{\text{unfold}}$ and the log-ratio mRNA levels attains a statistically significant maximum at $L \sim 5$ codons (Figure 5B). With this definition, L is about 30 bases downstream from the pBAD SD sequence, and the average magnitude of ΔG_{unfold} is a physically reasonable $10kT$. For illustration, the correlation at $L = 6$ codons is shown in Figure 5E (Pearson $R = 0.65$, $p = 3.3 \times 10^{-8}$), and this approach works equally well for the subset of constructs with N-terminal mutations only (Figure S1A). Our

focus on the 5'-end is further justified by the lack of a significant relationship between the log-ratio mRNA levels and $\Delta\Delta G_{L,N}$ (Figure S1B). However, long-range interactions between the 5' and 3'-ends are clearly important, since the residual variations in the mRNA levels that cannot be explained by $\Delta\Delta G_{1,L}$ correlate significantly (Pearson $R = 0.54$, $p = 1.2 \times 10^{-5}$) with $\Delta\Delta G_{\text{non-local}}$ (Figure 5F). We therefore attribute the ability of $\Delta\Delta G_{\text{unfold}}$ to predict the effects of far-downstream substitutions to the transient sampling of non-equilibrium secondary structures during mRNA transcription and folding.

Chromosomal incorporation reveals a strong dependence on the promoter sequence

Our analysis based on ΔG_{unfold} implies that mutations in the 5'-UTR may significantly change the effects of synonymous substitutions in the coding region. To test this prediction, we introduced select synonymous substitutions in the chromosomal copies of *folA* and *adk* using genome editing. As expected, use of the endogenous chromosomal promoter, as opposed to the pBAD system, alters $\Delta\Delta G_{\text{unfold}}$ and thus the relative mRNA levels (Figure 5G). Here we have normalized the mRNA levels with respect to the WT coding sequence on the same promoter to control for the ribosome binding site strength and to isolate the effects of the synonymous substitutions. With the endogenous promoters, we find that the log-ratio mRNA levels maintain a strong correlation with $\Delta\Delta G_{\text{unfold}}$ (Pearson $R = 0.86$, $p = 0.0026$). Furthermore, as predicted by $\Delta\Delta G_{\text{unfold}}$, single synonymous substitutions in two *adk* constructs (6TTG and 6TTA) that have deleterious effects on the mRNA levels when expressed on the pBAD promoter turn out to have negligible, or even beneficial, effects when chromosomally incorporated.

A ribosome-protection mechanism explains differential mRNA degradation

Numerous studies have shown that ribosome occupancy near the SD sequence can directly modulate the rate at which mRNAs decay, suggesting that ribosomes pose a steric hindrance against the action of endonucleases (Deana and Belasco, 2005). We therefore propose a 'ribosome protection' mechanism to explain sequence-specific effects on mRNA degradation. We hypothesize that synonymous mutations throughout the *folA* and *adk* genes affect the folding of their mRNA transcripts such that all or part of the SD sequence is occluded (Figure 6A), resulting in inefficient translation initiation (de Smit and Van Duin, 1990; Hall et al., 1982; Li et al., 2012), and, in the absence of translating ribosomes, degradation by endonucleases such as RNaseE (Mackie, 2013; Steege, 2000). This hypothesis implies that codon rarity and the average GC content are not essential, but rather that deleterious interactions with the 5'-end need to be avoided.

We can thus formulate a minimal kinetic model to describe the coupling of the mRNA and protein abundances:

$$\frac{dM}{dt} = k_{\text{M}+} - k_{\text{M}-}(M - M_{\text{R}}) \quad (2)$$

$$\frac{dM_{\text{R}}}{dt} = k_{\text{R}+}(R - M_{\text{R}})(M - M_{\text{R}}) - k_{\text{R}-}M_{\text{R}} \quad (3)$$

$$\frac{dP}{dt} = k_{\text{R}-}M_{\text{R}}. \quad (4)$$

Here, M is the total mRNA concentration, M_{R} is the concentration of mRNA bound by ribosomes, R is the concentration of ribosomes, and dP/dt is the rate of protein production. The mRNA transcription and degradation rates are $k_{\mathrm{M+}}$ and $k_{\mathrm{M-}}$, respectively, and are assumed to be constant for all constructs. The rate of ribosome attachment, $k_{\mathrm{R+}} \sim \exp(-\Delta G_{\mathrm{unfold}}/kT)$, depends on the accessibility of the 5'-UTR. Meanwhile, comparison with titrated WT DHFR on the plasmid (Figure S3) suggests that the ribosome detachment rate $k_{\mathrm{R-}}$ is relatively constant at low to moderate over-expression levels, but decreases with M_{R} due to resource limitations at high over-expression ($\gtrsim 1000$-fold over the chromosomal level).

This model makes two key predictions. First, the steady-state mRNA concentration

$$M_{\mathrm{ss}} = \left(\frac{k_{\mathrm{M+}}}{k_{\mathrm{M-}}}\right)\left[1 + R\left(\frac{k_{\mathrm{R-}}}{k_{\mathrm{R+}}} + \frac{k_{\mathrm{M+}}}{k_{\mathrm{M-}}}\right)^{-1}\right] \quad (5)$$

is a linear function of $k_{\mathrm{R+}}$ in the regime where translation-initiation is rate-limiting. As shown in Figure 5E, mRNA levels do indeed follow a linear relationship with $\Delta\Delta G_{\mathrm{unfold}}$. Second, the rate of protein production is proportional to the steady-state mRNA concentration,

$$\left(\frac{dP}{dt}\right)_{ss} = k_{R-}\left(M_{ss} - \frac{k_{M+}}{k_{M-}}\right), \quad (6)$$

and is consequently also a linear function of $k_{\mathrm{R+}}$ when translation-initiation is rate-limiting and $k_{\mathrm{R-}}$ is constant. Figure 6B,C shows that these relationships hold for both DHFR and ADK over approximately three orders of magnitude when expressed on both the chromosome and the plasmid.

Synonymous substitutions in *E. coli folA* cause fitness effects via changes in protein expression

Our previous work (Bhattacharyya et al., 2016) has shown that both under and over-expression of DHFR decrease the exponential-phase growth rate of *E. coli*. Consistent with this expectation, we find that DHFR under-expression in this study has a Michaelis–Menten-like effect on cellular fitness (Figure 6D, red points). However, unlike our prior study, under-expression in this case is caused solely by synonymous substitutions within the chromosomal copy of *folA*. Similarly, over-expression in the present study (Figure 6D, blue points) is caused by synonymous substitutions at a fixed inducer concentration. To determine whether changes in aggregation propensity might affect fitness, we also measured the total intra-cellular protein abundances for each synonymous construct. We observed a strong linear relationship between variations in the total and soluble protein abundances for both DHFR and ADK (Figure S4) both on the plasmid and on the chromosome, indicating that aggregation does not increase significantly upon over-expression and confirming that the observed fitness effects arise from changes in DHFR expression levels.

A genome-wide analysis reveals selection to avoid Shine–Dalgarno mis-interactions

Thus far, we have described a specific mechanism by which synonymous substitutions near the 5'-end dictate protein expression and, for one endogenous gene, affect cellular fitness. To evaluate the generality of these conclusions, we compared the folding free energies of WT and synonymous transcripts for all genes in the *E. coli* genome. Using the final 20 bases of the 5'-UTR and the entire coding sequence, we found the mean $\Delta G_{\mathrm{unfold}}$ (at $L =$

5 codons) of the WT sequences to be approximately $-12kT$ (Figure 7A). We then constructed a control data set by considering all possible synonymous variants of the first five codons of each *E. coli* gene. Assuming that the synonymous codons are selected independently and with equal frequencies, the most probable $\Delta G_{\mathrm{unfold}}$ distribution is shifted to the left of the WT distribution (Fig.6A). This observation is consistent with selection against strong interactions between the SD and coding sequences. To make a quantitative comparison between these histograms, we computed the probability, within the control distribution, of observing no more than the actual number of genes with $\Delta G_{\mathrm{unfold}}$ less than a specified threshold. The resulting *p*-values indicate that the control and WT distributions differ most significantly with respect to the number of genes that have $\Delta G_{\mathrm{unfold}} \lesssim -14kT$ (Figure 7B, red curve). This conclusion is unchanged if we instead bias the selection of synonymous codons in accordance with the empirical genome-wide codon-usage frequencies within the first five coding positions (Figure 7C, blue curve). This analysis demonstrates that the sequence context, and not simply the relative usage of each codon, contributes to the difference between the WT and the control $\Delta G_{\mathrm{unfold}}$ distributions.

Finally, we sought to determine whether *E. coli* genes are optimized to avoid base-pairing with the SD sequence. Based on our analysis of *folA* and *adk* mRNA levels, we reasoned that any *local* secondary structure at the 5'-end of a transcript would, in general, be close to equilibrium. We therefore calculated p_{unbound} for the final 20 bases of the 5'-UTR and the first 5 codons for all *E. coli* genes and their synonymous variants. By looking at the difference between p_{unbound} in each WT sequence and the mean of the uniformly weighted synonymous variants, we detected a statistically significant reduction in base-pairing within the 5'-UTR of the WT genes (Figure 7C). We then located the SD sequence for each gene by minimizing the hybridization free energy between a 7-base sliding window and the *E. coli* rRNA anti-SD sequence, verifying that the resulting distribution of SD-sequence locations agrees with previous experimental studies of SD-sequence placement (Chen et al., 1994). Remarkably, the greatest reduction in base-pairing in the WT sequences coincides precisely with the most likely location of the SD region. The overlap of these two features indicates that the reduction in mRNA secondary structure near the 5'-end is driven specifically by an avoidance of interactions leading to the occlusion of the SD sequence. By directly implicating the SD sequence, this statistical analysis lends additional support to our proposed ribosome-protection mechanism.

Discussion

Codon usage bias has been a fascinating topic of research for over 30 years (Brule and Grayhack, 2017; Chaney and Clark, 2015; Plotkin and Kudla, 2011). While uneven codon usage has been firmly established, the biological causes and consequences of non-random synonymous-codon selection remain largely unknown. Identifying specific mechanisms is challenging, since codon usage can affect all stages of protein biosynthesis, from transcription and translation to co-translational protein folding and potentially protein-complex assembly. Codon usage also varies significantly within individual genes (Clarke and Clark, 2010). Of particular interest is codon usage near the N-terminus, where the relative frequencies of synonymous codons tend to be most dissimilar from the remainder of the coding sequence. Because substitutions near the beginning of a gene are unlikely to affect co-translational protein

folding, it is widely believed that this N-terminal codon bias arises in some way from the behavior of mRNA transcripts.

Previous experimental approaches to assess the effects of synonymous substitutions have used randomized libraries of synonymous variants of several endogenous and non-endogenous genes expressed in *E. coli* (Boel et al., 2016; Goodman et al., 2013; Kudla et al., 2009; Welch et al., 2009). These studies provide a plethora of data for statistical analyses that can be used to assess the primary determinants of codon-level effects on protein expression. All of these studies showed large variations in protein expression, often reaching several orders of magnitude, as a result of substitutions near the N-terminus. Perhaps surprisingly, a similar or greater variability was observed for mRNA levels, immediately suggesting a link between synonymous codon usage and mRNA turnover. Linear regression analyses pointed to multiple relevant factors, including mRNA secondary structure stability, deleterious effects of specific codons, and several other signatures that were shown to predict protein expression in gene-design applications (Boel et al., 2016; Welch et al., 2009). However, such an approach cannot provide an understanding of the underlying physical mechanisms and evolutionary factors that determine the observed biases in codon usage.

To address these challenges, we chose an alternative, bottom-up approach where we explored possible mechanisms "one codon substitution at a time." By designing synonymous variants of two endogenous *E. coli* genes, we tested hypotheses regarding the relative importance of codon rarity, the context of rare codons within a sequence, and multiple predictors of mRNA folding and stability. In agreement with earlier findings, we observed dramatic variations in both mRNA and protein expression levels. We further determined that synonymous replacements do not significantly affect the efficiency of transcription *in vitro*. Consequently, we attribute the effects of synonymous replacements on mRNA levels to differential rates of mRNA degradation, which may occur in concert with premature transcription termination.

Our systematic analysis revealed highly position-specific effects of synonymous substitutions on mRNA and protein levels. We found that synonymous replacements of rare codons, particularly near the N-terminus, tend to have the greatest effects. In particular, we found that synonymous replacements of the first rare codon (encoding Ser3 in *folA* and Leu6 in *adk*) resulted in the most dramatic drops in mRNA levels. However, codon rarity *per se* is not a determining factor, since replacing WT rare codons with either optimal or alternative rare codons had similar detrimental effects. This observation argues against the importance of translational pauses near the 5'-end (Eriksen et al., 2017; Mitarai et al., 2008; Pedersen et al., 2011).

Our results are instead consistent with the hypothesis that the 5'-end of a transcript should be free of secondary structure to promote efficient translation initiation (Eriksen et al., 2017; Ringquist et al., 1992). Yet in contrast to the conventional mRNA stability definition (Boel et al., 2016; Goodman et al., 2013; Kudla et al., 2009), which requires a subjective truncation of the RNA molecule, we propose a physically justified approach that captures the effects of substitutions throughout a transcript. We have shown that the unusual specificity of substitutions within the first 6 codons results primarily from the formation of secondary structures involving the upstream SD sequence. Because the participating bases are close to one another in the primary sequence, this effect can typically be predicted by examining the equilibrium secondary structure of the transcript. This interpretation is supported by the behavior of a synonymous *adk* construct that incorporates an exact anti-SD sequence but retains all native rare codons, as well as

by our statistical analysis of the *E. coli* genome, which specifically implicates avoidance of SD-sequence base-pairing as a contributing factor to N-terminal codon bias.

In addition, our detailed investigation reveals a key role of non-equilibrium folding effects. Since the folding time for a typical mRNA transcript can be much greater than the time required for transcription, long-range contacts that are not part of the minimum-free-energy secondary structure may transiently sequester the SD sequence. To this end, we demonstrated that long-range contacts involving the 5'-UTR, which are affected by substitutions far-downstream from the N-terminus, correlate significantly with variations in the mRNA levels, even though the predicted *equilibrium* secondary structure near the 5'-end remains unchanged. Our approach thus accounts explicitly for the co-transcriptional nature of *in vivo* RNA folding. Nevertheless, our treatment of non-equilibrium secondary structures is only approximate, and further theoretical analysis is needed to understand these effects completely.

Taken together, our results point to a unifying ribosome-protection model that explains the effects of synonymous substitutions on both mRNA degradation rates and protein expression levels. Previous studies have shown that mutations in the coding region which lead to a higher propensity to bind to the SD sequence accelerate RNaseE-mediated mRNA degradation through competitive inhibition of the RBS (Cho and Yanofsky, 1988; Yarchuk et al., 1992). Similarly, mutations in the RBS that affect pairing with the anti-SD sequence have been shown to predictably increase (Arnold et al., 1998; Hambraeus et al., 2002; Jurgen et al., 1998; Sharp and Bechhofer, 2003; Wagner et al., 1994) or decrease (Jain and Kleckner, 1993; Matsunaga et al., 1997) the rate of mRNA decay. These studies imply that the process of translation protects mRNA transcripts from degradation. Consequently, our model predicts that synonymous substitutions have commensurate effects on mRNA and protein expression levels, despite the fact that the rate of mRNA transcription remains constant. Our systematic analysis thus allows us to draw a clear mechanistic link between ribosome protection and codon usage bias.

Finally, we can connect the effects of synonymous substitutions to changes in cellular fitness, both in the specific case of *folA* and, more generally, for the *E. coli* genome as a whole. Comparing the expression of synonymous variants on the chromosome versus the plasmid, we found that the WT *folA* sequence appears to be well optimized to maximize protein production under its endogenous promoter. In contrast, we identified numerous substitutions that increase the mRNA and protein levels of *adk* relative to the WT sequence under its endogenous promoter. This difference may result from the sensitivity of the *E. coli* growth rate to under-expressing DHFR but not ADK (Bhattacharyya et al., 2016), suggesting a possible mechanistic rationale for the evolutionary conservation of codon usage in specific genes. Consistent with this hypothesis, we observed a statistically significant decrease in the genome-wide probability of SD-sequence occlusion relative to random synonymous sequences, providing evidence of widespread positive selection for this feature. Notably, this difference cannot be explained simply by the relative frequencies of synonymous codons near the N-terminus, which again emphasizes the importance of sequence specificity in determining the interactions between the coding and SD sequences.

In summary, we have used a systematic analysis of synonymous substitutions in two endogenous *E. coli* genes to establish a mechanism of N-terminal codon bias. We have shown that secondary structure formation involving the Shine–Dalgarno sequence is the primary determinant of the effects of codon choice because of its crucial function in promoting translation initiation, which modulates both mRNA and protein expression levels. This analysis improves

upon earlier studies by validating a physically motivated definition of 5'-end stability and revealing the previously unrecognized, yet crucial, role of co-transcriptional mRNA folding. Our model therefore connects 'silent' mutations, through RNA folding dynamics, to their effects on cellular fitness, providing a general mechanistic insight into the origin of codon-usage bias.

Methods

Definition of rare codons. We consider a codon to be rare if its usage within the most abundant *E. coli* proteins (Ikemura, 1985) is less than 10% of the usage of the amino acid for which it codes.

Plasmids and strain construction. We cloned synonymous codon variants of *folA* and *adk* genes in the pBAD plasmid between *NdeI* and *XhoI* sites under control of an arabinose inducible promoter. We used a megaprimer based PCR method or an inverse PCR method using partially complementary primers for site-directed mutagenesis to mutate codons on the plasmid. For both methods, KOD Hot Start DNA polymerase was used. The fully optimized 'MutAll' sequence and the rare-optimized 'MutRare' sequences for both *folA* and *adk* were synthesized as g-block gene fragments from IDT, and cloned in pBAD vector. Expression of pBAD plasmid in all cases was done in *E. coli* BW27783 cells (CGSC #12119).
We made MG1655 strains having synonymous mutations in the chromosomal copies of the *folA* and *adk* genes using a genome editing technique described previously (Bershtein et al., 2012).

Media conditions. We carried out bacterial growth in M9 minimal salts supplemented with 0.2% glucose, 1mM $MgSO_4$, 0.1% casamino acids, and 0.5 µg/ml thiamine (supplemented M9 medium). For BW27783 cells containing the pBAD plasmid, ampicillin was added to a concentration of 100µg/ml during growth. mRNA and protein abundance measurements were carried out using a saturating arabinose concentration of 0.05%.

mRNA abundance measurement. To determine *in vivo* mRNA levels, BW27783 cells transformed with pBAD plasmid were grown overnight in supplemented M9 medium containing 100µg/ml ampicillin at 37°C. The next day, the culture was diluted 1:100 into fresh M9 medium containing 0.05% arabinose and 100µg/ml ampicillin and grown for 4 hours at 37°C. Based on OD_{600} of the cultures, a culture volume corresponding to 5×10^8 cells was spun down, re-suspended in 1ml of M9 medium, and treated with 2ml of RNAprotect Bacteria reagent (Qiagen), after which the total RNA was extracted using an RNeasy mini kit according to the manufacturer's instructions. The concentration of the purified RNA samples was estimated using nanodrop. 1µg of the purified total RNA was subjected to reverse transcription to produce cDNA using a QuantiTect Reverse Transcription kit (Qiagen). The levels of *folA* and *adk* genes were estimated by quantitative PCR (qPCR) using 20bp gene specific primers and a QuantiTect SYBR Green PCR kit (Qiagen). For synonymous variants of *folA* gene on the plasmid, chromosomally expressed *adk* gene was used as a reference, while for synonymous variants of *adk* gene on the plasmid, chromosomally expressed *tufA* gene was used as a reference.

Protein abundance measurement. BW27783 cells transformed with pBAD plasmid were grown in supplemented M9 medium for 4 hours at 37°C, chilled on ice for 30 min and then spun down in a centrifuge. The cells were lysed with 1×BugBuster (Novagen) and 25units/ml of Benzonase such that the OD600 of the mix before lysis was 0.2. 10µl of the lysate (before centrifugation) was mixed with 2µl of 50% glycerol+10% SDS, heated at 95°C for 10

minutes, and used for estimation of the total protein abundance. The remaining lysate was centrifuged for 30 minutes at high speed and used for estimation of the soluble protein abundance. Intracellular DHFR and ADK protein amounts in both the total and soluble fractions were determined by SDS-PAGE followed by Western Blot using rabbit anti-DHFR/anti-ADK polyclonal antibodies (custom raised by Pacific Immunology).

Error analysis. The differences between replicates were plotted versus the replicate averages for both mRNA and soluble-protein measurements (Figure S5). Linear fits were performed to estimate the absolute errors, which are constant, and the relative errors, which are proportional to the mean value of each measurement, under the constraint that the absolute error is non-negative. Four *adk* constructs, for which the soluble-protein measurements fall below the absolute error estimate, were excluded from the soluble-protein correlation calculations. The error bars shown in Figures 1, 3, and 4 indicate the estimated standard deviation of the measurements. The error bars in Figure 2 indicate the standard error of the mean of three replicates.

***In vitro* transcription.** The coding regions of WT, MutAll and MutRare versions of *folA* gene were PCR amplified with primers to incorporate a T7 promoter upstream of the coding region, and this linearized DNA was used as a template for *in vitro* transcription using a Maxiscript T7 kit (Thermo Fisher Scientific) using the manufacturer's instructions. The DNA template was digested with the Turbo DNase included in the kit, after which the mRNA was purified using an RNeasy mini kit, reverse transcribed to generate cDNA, and quantified using qPCR as described above.

Fitness measurements. Overnight cultures of BW27783 cells containing pBAD plasmid were grown with ampicillin at 37°C while MG1655 strains containing chromosomally engineered mutants of *folA* and *adk* were grown at 30°C. The next day, cultures were diluted 1:100 and grown at 37°C for 12 hours in Bioscreen C system (Growth Curves USA). For BW27783 cells, 0.05% arabinose was added for induction. OD data were collected at 600nm at 15 min intervals. The resulting growth curves were fit to a bacterial growth model to obtain growth rate and lag time parameters (Zwietering et al., 1990).

RNA folding calculations. We used NUPACK (Zadeh et al., 2011) for all RNA folding calculations. This program computes the partition function, $Z \equiv \sum_{\mathrm{ss}} \exp\left(-\Delta g_{\mathrm{ss}}/kT\right)$, where the sum runs over all secondary structures, excluding pseudo-knots. The free energies of individual secondary structures, Δg_{ss}, are predicted using experimentally determined RNA base-pairing rules. The folding free energy of a transcript is then defined as $\Delta G_{1,N} = -kT \log Z$. Unless otherwise noted, the full transcript, from the predicted +1 site to the stop codon, was used. Calculations were performed assuming standard conditions ($[\mathrm{Na}^+] = 1\ \mathrm{M};\ T = 37°\mathrm{C}$), with the results reported relative to the thermal energy, kT. While these conditions are not physiological, we expect the relative free-energy differences among synonymous sequences to be robust.

The equilibrium probability that a base is unbound is given by $p_{\mathrm{unbound},i} = 1 - Z^{-1} \sum_{\mathrm{ss}} \mathbf{1}_{i,\mathrm{ss}} \exp\left(-\Delta g_{\mathrm{ss}}/kT\right)$, where $\mathbf{1}_{i,\mathrm{ss}}$ indicates that the base at position i is paired in a given secondary structure. To examine the role of

specific bases (Figure 5C), we used a bootstrap analysis to account for the effect of outliers on the correlation between the log-ratio mRNA levels and the log-ratio of $p_{\mathrm{unbound},i}$. To this end, we sampled our synonymous constructs with replacement to create approximately 1,000 alternative datasets with the same length but randomly repeated entries. The bootstrapped *p*-value is then the arithmetic mean of the one-tailed Pearson correlation *p*-values for all generated samples.

Figure captions

Figure 1: **Synonymous substitutions in the chromosomal copy of *E. coli folA* affect cellular fitness, soluble protein abundance, and mRNA levels.** (A) The locations of optimized codons within each *folA* construct are indicated by colored squares. (B) Replacing WT codons by their most frequently used synonymous variants has a deleterious effect on *E. coli* growth, reducing the exponential growth rate and increasing the lag time. (C) Synonymous substitutions affect both the soluble DHFR abundance (measured by quantitative Western blot) and steady-state mRNA transcript levels (measured by qPCR). All values are normalized to the WT.

Figure 2: **Codon usage has no significant effect on transcription.** The WT, MutAll and MutRare variants of the *folA* gene were fused to the bacteriophage T7 promoter, and transcription was carried out *in vitro* using T7 RNA polymerase. The amount of transcribed mRNA was quantified using qPCR, and mRNA levels were normalized with respect to the WT. Although the codon-optimized variants exhibit a trend toward slightly lower mRNA levels, the variations are too small to explain the *in vivo* data.

Figure 3: **The first rare codon has the largest effect on intracellular mRNA levels.** Using an arabinose-inducible pBAD promoter, intra-cellular mRNA levels at an inducer concentration of 0.05% were measured using quantitative PCR and compared to their WT level (see Table S1). Systematic optimization of both *folA* (A) and *adk* (B) revealed that optimizing the first rare codon in each gene has a dominant effect. (C) Far-downstream mutations, introduced on either the 'MutRare,' 'MutC-term,' or WT backgrounds, also significantly modulate mRNA levels, albeit to a lesser extent.

Figure 4: **The pivotal role of the first rare codon is not due to rarity.** (A) All Serine codons were incorporated at position 3 of *folA*. Other than the WT codon (AGT), only TCA was tolerated. (B) Similarly, all Leucine codons were incorporated at position 6 of *adk*. CTA and CTC were found to increase the mRNA levels relative to the WT. These constructs show that neither codon rarity nor GC-content is responsible for the observed effects. (C,D) Substitutions at neighboring positions may rescue the effects of optimizing position 3 of *folA* or position 6 of *adk*, indicating that the role of the codons at these positions is context-dependent. Re-introducing the WT codons at these key positions on the MutRare backgrounds likewise highlights the importance of the sequence context.

Figure 5: $\Delta G_{\mathrm{unfold}}$ **captures the effects of synonymous substitutions throughout the entire mRNA transcript.** (A) Schematic representation of local contacts (involving only the first L bases) and non-local interactions. (B) The Pearson correlation coefficient between the log-ratio mRNA levels and two different folding-stability calculations. The correlation with $\Delta\Delta G_{\mathrm{unfold}}$ has a maximum at $L \simeq 5$ codons. (C–D) The statistical significance of the correlation between the log-ratio mRNA levels and the log-ratio of p_{unbound} is maximized for the base at position -10, which is part of the SD sequence. Correlations are calculated using the subset of constructs with substitutions closest to the 5'-end (i.e., the red and orange points in panel D), and *p*-values are determined from a bootstrap analysis of these data. In contrast, constructs with substitutions elsewhere in the coding region have no dependence on p_{unbound}. (E) The log-ratio mRNA levels for both genes correlate best with $\Delta\Delta G_{\mathrm{unfold}}$, shown here at $L = 6$ codons. (F) Analysis of the residuals after subtracting the contribution from base-pairing near the 5'-end demonstrates that $\Delta\Delta G_{\mathrm{unfold}}$ captures important long-range interactions. (G) Moving specific constructs back onto the chromosome indicates that the effects of synonymous substitutions are dependent on the 5'-UTR sequence. All measurements and predictions are presented relative to the corresponding pBAD or chromosomal WT construct. In panels D, E, and F, the various pBAD constructs (see Table S1) are color-coded as follows: WT (black); synonymous replacements of the first rare codon (red); synonymous mutations adjacent to the first rare codon, keeping the latter optimized (orange); systematic mutations of other rare codons (blue); sequences with varied $\Delta G_{1,L}$ (gray); and the *adk* construct that incorporates an anti-SD sequence (green).

Figure 6: **A ribosome protection mechanism explains sequence-specific variations in mRNA levels, protein abundances, and cellular fitness.** (A) Schematic representation of translation initiation, mRNA folding and degradation. An accessible SD sequence hybridizes with the complementary 16s-rRNA, which leads to efficient translation initiation and provides protection from degradation by nucleases. Alternatively, formation of intra-molecular base pairs involving the SD sequence inhibits translation initiation. Soluble protein abundances correlate with both (B) $\Delta\Delta G_{\mathrm{unfold}}$ and (C) mRNA levels. (D) Variations in soluble protein abundances arising from synonymous substitutions in *folA* predict cellular fitness (i.e., the exponential-phase growth rate normalized by the WT growth rate, either without the plasmid, for chromosomal incorporations, or at zero arabinose concentration, for pBAD constructs). The under-expression arm exhibits a Michaelis–Menten-like behavior, while the over-expression arm has a linear dependence on the soluble protein abundance.

Figure 7: **A statistical analysis of *E. coli* N-terminal codon bias provides evidence for selection against SD mis-interactions.** (A) A comparison of the genome-wide distributions of ΔG_{unfold} for the wild-type (WT) and alternative synonymous sequences, assuming that synonymous codons are used with equal frequencies. The first 15 bases of the coding region are included in these calculations, and we have used $L = 5$ codons in the calculations. (B) P-values (see text) assess the statistical significance of the difference between the WT and control distributions, assuming that synonymous codons are selected either uniformly or according to the empirical genome-wide usage within the first five codons. (C) The average difference in the equilibrium base-pairing probability between the WT and the mean of the uniformly weighted synonymous sequences for each gene. The most significant decrease in secondary-structure formation within the 5'-UTR (unshaded region) coincides with the most probable location of the SD sequence (red curve).

Supplementary figure captions

Figure S1: (A) The Pearson correlation coefficient, relating the log-ratio mRNA levels and two ΔG calculations, as a function of L (cf. Figure 5B). For these calculations, we have considered constructs with synonymous substitutions only in the first 8 codons. (B) The mRNA levels show no statistically significant dependence on the 'downstream' mRNA folding free energy, $\Delta G_{L,N}$. Here we have used $L = 6$ codons.

Figure S2: The probability distribution of the distance (with respect to the primary sequence) between paired bases i and j in the predicted minimum-free-energy secondary structures of *E. coli* mRNA transcripts.

Figure S3: Soluble protein abundances of the synonymous codon variants appear to plateau relative to the intra-cellular mRNA abundances at the highest expression levels. Circles represent *folA*, while squares represent *adk*. Soluble protein abundances of all variants were measured using 0.05% arabinose on the pBAD plasmid and normalized with respect to the WT level at this inducer concentration. Dashed lines show the approximately linear relationship between the protein and mRNA abundances at low to moderate expression. In (A), the red circles show the soluble abundance of WT DHFR obtained from a range of arabinose concentrations (0.0000122% to 0.05%).

Figure S4: The normalized soluble and total intra-cellular protein abundances exhibit a clear linear dependence. Consequently, the aggregation propensity of both proteins does not appear to increase substantially upon over-expression. Note that $[\text{soluble}]_{\text{WT}}/[\text{total}]_{\text{WT}} \simeq 0.9$ for *adk* and $\simeq 0.95$ for *folA*.

Figure S5: Error analysis for mRNA and soluble-protein levels on the plasmid. Linear fits were used to estimate the typical difference between replicates (see Methods). We then calculated the average standard deviation of each measurement, s_x, as a function of the mean of each measurement, $\bar{x}$. All values are shown here relative to the WT chromosomal expression level.

References

Adkar, B.V., Manhart, M., Bhattacharyya, S., Tian, J., Musharbash, M., and Shakhnovich, E.I. (2017). Optimization of lag phase shapes the evolution of a bacterial enzyme. Nat Ecol Evol *1*, 149.

Arnold, T.E., Yu, J., and Belasco, J.G. (1998). mRNA stabilization by the ompA 5' untranslated region: two protective elements hinder distinct pathways for mRNA degradation. RNA *4*, 319-330.

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., and Bluthgen, N. (2013). Efficient translation initiation dictates codon usage at gene start. Molecular systems biology *9*, 675.

Bershtein, S., Mu, W., and Shakhnovich, E.I. (2012). Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. Proc Natl Acad Sci U S A *109*, 4857-4862.

Bhattacharyya, S., Bershtein, S., Yan, J., Argun, T., Gilson, A.I., Trauger, S.A., and Shakhnovich, E.I. (2016). Transient protein-protein interactions perturb E. coli metabolome and cause gene dosage toxicity. Elife *5*.

Boel, G., Letso, R., Neely, H., Price, W.N., Wong, K.H., Su, M., Luff, J.D., Valecha, M., Everett, J.K., Acton, T.B., *et al.* (2016). Codon influence on protein expression in E. coli correlates with mRNA levels. Nature *529*, 358-363.

Brule, C.E., and Grayhack, E.J. (2017). Synonymous Codons: Choose Wisely for Expression. Trends Genet *33*, 283-297.

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. Genetics *129*, 897-907.

Chaney, J.L., and Clark, P.L. (2015). Roles for Synonymous Codon Usage in Protein Biogenesis. Annu Rev Biophys *44*, 143-166.

Chaney, J.L., Steele, A., Carmichael, R., Rodriguez, A., Specht, A.T., Ngo, K., Li, J., Emrich, S., and Clark, P.L. (2017). Widespread position-specific conservation of synonymous rare codons within coding sequences. PLoS Comput Biol *13*, e1005531.

Chen, H., Bjerknes, M., Kumar, R., and Jay, E. (1994). Determination of the optimal aligned spacing between the Shine–Dalgarno sequence and the translation initiation codon of Escherichia coli m RNAs. Nucleic acids research *22*, 4953-4957.

Cho, K.O., and Yanofsky, C. (1988). Sequence changes preceding a Shine-Dalgarno region influence trpE mRNA translation and decay. J Mol Biol *204*, 51-60.

Clarke, T.F.t., and Clark, P.L. (2010). Increased incidence of rare codon clusters at 5' and 3' gene termini: implications for function. BMC genomics *11*, 118.

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

Deana, A., and Belasco, J.G. (2005). Lost in translation: the influence of ribosomes on bacterial mRNA decay. Genes Dev *19*, 2526-2533.

Eriksen, M., Sneppen, K., Pedersen, S., and Mitarai, N. (2017). Occlusion of the Ribosome Binding Site Connects the Translational Initiation Frequency, mRNA Stability and Premature Transcription Termination. Front Microbiol *8*, 362.

Eyre-Walker, A., and Bulmer, M. (1993). Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res *21*, 4599-4603.

Galzitskaya, O., and Finkelstein, A.V. (1998). Folding rate dependence on the chain length of RNA-like heteropolymers. Fold Des *3*, 69-78.

Goodman, D.B., Church, G.M., and Kosuri, S. (2013). Causes and effects of N-terminal codon bias in bacterial genes. Science *342*, 475-479.

Grantham, R., Gautier, C., and Gouy, M. (1980a). Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res *8*, 1893-1912.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pave, A. (1980b). Codon catalog usage and the genome hypothesis. Nucleic Acids Res *8*, r49-r62.

Hall, M.N., Gabay, J., Debarbouille, M., and Schwartz, M. (1982). A role for mRNA secondary structure in the control of translation initiation. Nature *295*, 616-618.

Hambraeus, G., Karhumaa, K., and Rutberg, B. (2002). A 5' stem-loop and ribosome binding but not translation are important for the stability of Bacillus subtilis aprE leader mRNA. Microbiology *148*, 1795-1803.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol *2*, 13-34.

Jacobs, W., and Shakhnovich, E.I. (2017). Evidence of evolutionary selection for co-translational folding. PNAS *in press*.

Jain, C., and Kleckner, N. (1993). IS10 mRNA stability and steady state levels in Escherichia coli: indirect effects of translation and role of rne function. Mol Microbiol *9*, 233-247.
Jurgen, B., Schweder, T., and Hecker, M. (1998). The stability of mRNA from the gsiB gene of Bacillus subtilis is dependent on the presence of a strong ribosome binding site. Mol Gen Genet *258*, 538-545.
Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. Science *324*, 255-258.
Li, G.W., Oh, E., and Weissman, J.S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature *484*, 538-541.
Lynn, D.J., Singer, G.A., and Hickey, D.A. (2002). Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res *30*, 4272-4277.
Mackie, G.A. (2013). RNase E: at the interface of bacterial RNA processing and decay. Nat Rev Microbiol *11*, 45-57.
Matsunaga, J., Simons, E.L., and Simons, R.W. (1997). Escherichia coli RNase III (rnc) autoregulation occurs independently of rnc gene translation. Mol Microbiol *26*, 1125-1135.
Mitarai, N., Sneppen, K., and Pedersen, S. (2008). Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization. J Mol Biol *382*, 236-245.
Muto, A., and Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci U S A *84*, 166-169.
Pedersen, M., Nissen, S., Mitarai, N., Lo Svenningsen, S., Sneppen, K., and Pedersen, S. (2011). The functional half-life of an mRNA depends on the ribosome spacing in an early coding region. J Mol Biol *407*, 35-44.
Plotkin, J.B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet *12*, 32-42.
Presnyak, V., Alhusaini, N., Chen, Y.H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R., *et al.* (2015). Codon optimality is a major determinant of mRNA stability. Cell *160*, 1111-1124.
Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J., Stormo, G.D., and Gold, L. (1992). Translation initiation in Escherichia coli: sequences within the ribosome-binding site. Mol Microbiol *6*, 1219-1229.
Sharp, J.S., and Bechhofer, D.H. (2003). Effect of translational signals on mRNA decay in Bacillus subtilis. J Bacteriol *185*, 5372-5379.
Sharp, P.M., and Li, W.H. (1987). The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol *4*, 222-230.
Steege, D.A. (2000). Emerging features of mRNA decay in bacteria. RNA *6*, 1079-1090.
Taniguchi, Y., Choi, P.J., Li, G.W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X.S. (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science *329*, 533-538.
Tokuoka, M., Tanaka, M., Ono, K., Takagi, S., Shintani, T., and Gomi, K. (2008). Codon optimization increases steady-state mRNA levels in Aspergillus oryzae heterologous gene expression. Appl Environ Microbiol *74*, 6538-6546.
Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell *141*, 344-354.
Voges, D., Watzele, M., Nemetz, C., Wizemann, S., and Buchberger, B. (2004). Analyzing and enhancing mRNA translational efficiency in an Escherichia coli in vitro expression system. Biochem Biophys Res Commun *318*, 601-614.
Wagner, L.A., Gesteland, R.F., Dayhuff, T.J., and Weiss, R.B. (1994). An efficient Shine-Dalgarno sequence but not translation is necessary for lacZ mRNA stability in Escherichia coli. J Bacteriol *176*, 1683-1688.
Welch, M., Govindarajan, S., Ness, J.E., Villalobos, A., Gurney, A., Minshull, J., and Gustafsson, C. (2009). Design parameters to control synthetic gene expression in Escherichia coli. PloS one *4*, e7002.
Yarchuk, O., Jacques, N., Guillerez, J., and Dreyfus, M. (1992). Interdependence of translation, transcription and mRNA degradation in the lacZ gene. J Mol Biol *226*, 581-596.
Zadeh, J.N., Steenberg, C.D., Bois, J.S., Wolfe, B.R., Pierce, M.B., Khan, A.R., Dirks, R.M., and Pierce, N.A. (2011). NUPACK: analysis and design of nucleic acid systems. Journal of computational chemistry *32*, 170-173.
Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C.H., Fu, J., Chen, S., and Liu, Y. (2016). Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proc Natl Acad Sci U S A *113*, E6117-E6125.
Zwietering, M.H., Jongenburger, I., Rombouts, F.M., and van 't Riet, K. (1990). Modeling of the bacterial growth curve. Appl Environ Microbiol *56*, 1875-1881.

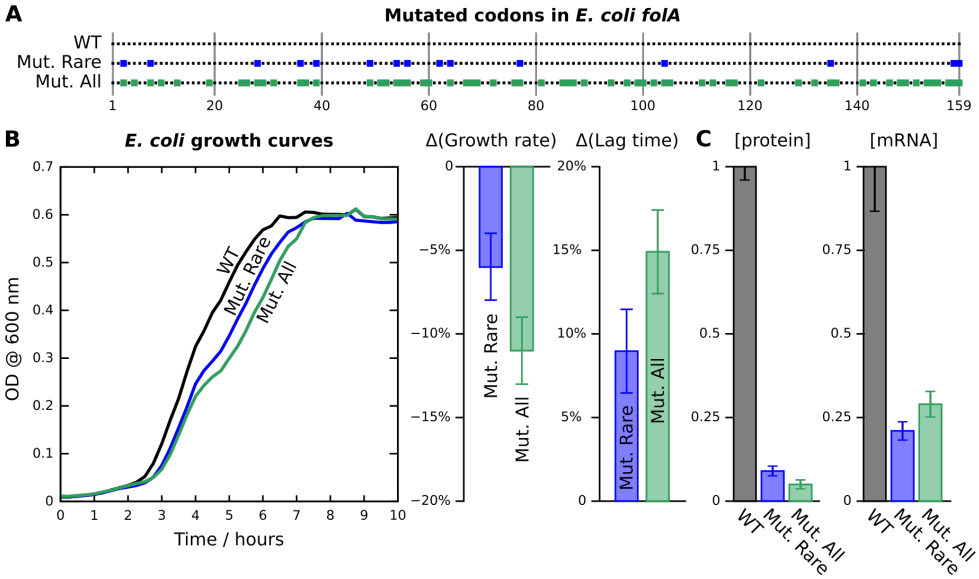
A
Mutated codons in E. coli folA
WT
Mut. Rare
Mut. All
1
20
40
60
80
100
120
140
159
B
E. coli growth curves
OD @ 600 nm
Time / hours
WT
Mut. Rare
Mut. All
Δ(Growth rate)
Δ(Lag time)
C
[protein]
[mRNA]

FIG. 1.

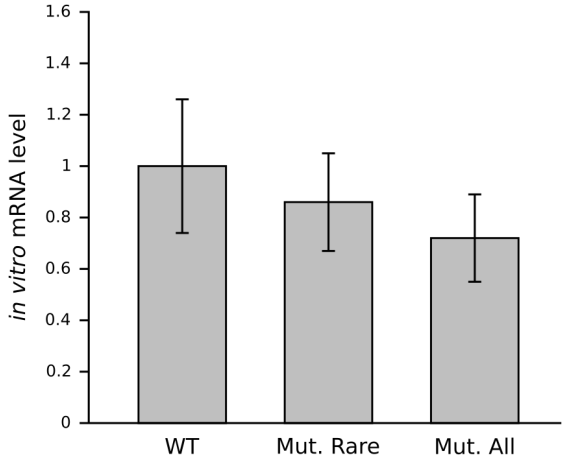

FIG. 2.

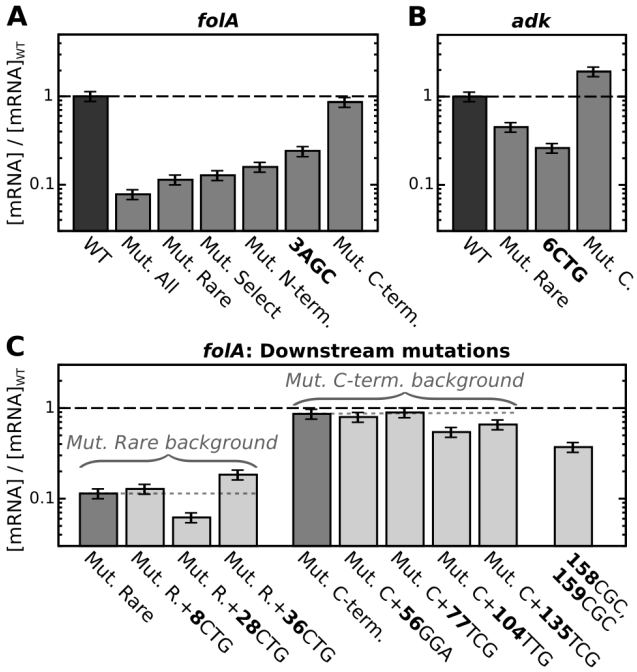
A
folA
[mRNA] / [mRNA]WT
1
0.1
WT
Mut. All
Mut. Rare
Mut. Select
Mut. N-term.
3AGC
Mut. C-term.
B
adk
1
0.1
WT
Mut. Rare
6CTG
Mut. C.
C
folA: Downstream mutations
[mRNA] / [mRNA]WT
Mut. C-term. background
Mut. Rare background
1
0.1
Mut. Rare
Mut. R.+8CTG
Mut. R.+28CTG
Mut. R.+36CTG
Mut. C-term.
Mut. C+56GGA
Mut. C+77TCG
Mut. C+104TTG
Mut. C+135TCG
158CGC,
159CGC

FIG. 3.

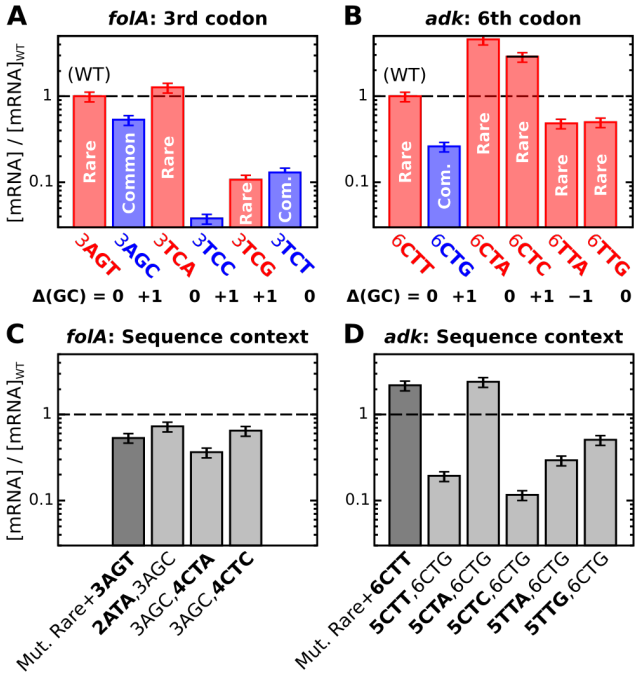
A
folA: 3rd codon
B
adk: 6th codon
[mRNA] / [mRNA]WT
(WT)
1
0.1
Rare
Common
Rare
Rare
Com.
3AGT
3AGC
3TCA
3TCC
3TCG
3TCT
Δ(GC) = 0 +1 0 +1 +1 0
Rare
Com.
Rare
Rare
Rare
Rare
6CTT
6CTG
6CTA
6CTC
6TTA
6TTG
Δ(GC) = 0 +1 0 +1 −1 0
C
folA: Sequence context
D
adk: Sequence context
Mut. Rare+3AGT
2ATA,3AGC
3AGC,4CTA
3AGC,4CTC
Mut. Rare+6CTT
5CTT,6CTG
5CTA,6CTG
5CTC,6CTG
5TTA,6CTG
5TTG,6CTG

FIG. 4.

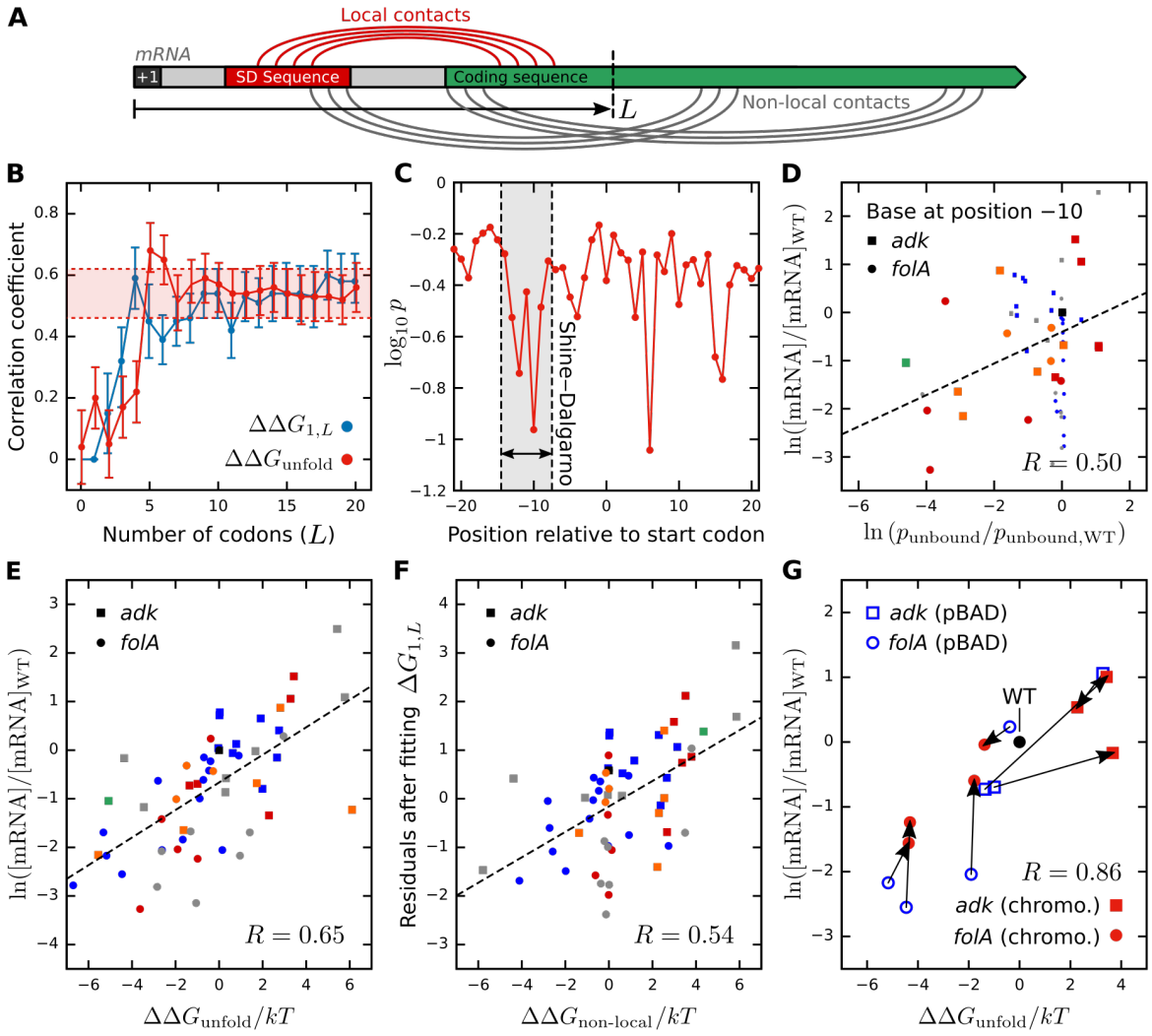
A
Local contacts
mRNA
+1
SD Sequence
Coding sequence
L
Non-local contacts
B
Correlation coefficient
Number of codons (L)
ΔΔG1,L
ΔΔGunfold
C
log10 p
Shine–Dalgarno
Position relative to start codon
D
Base at position −10
adk
folA
R = 0.50
ln([mRNA]/[mRNA]WT)
ln (punbound/punbound,WT)
E
adk
folA
R = 0.65
ln([mRNA]/[mRNA]WT)
ΔΔGunfold/kT
F
adk
folA
R = 0.54
Residuals after fitting ΔG1,L
ΔΔGnon-local/kT
G
adk (pBAD)
folA (pBAD)
WT
R = 0.86
adk (chromo.)
folA (chromo.)
ln([mRNA]/[mRNA]WT)
ΔΔGunfold/kT

FIG. 5.

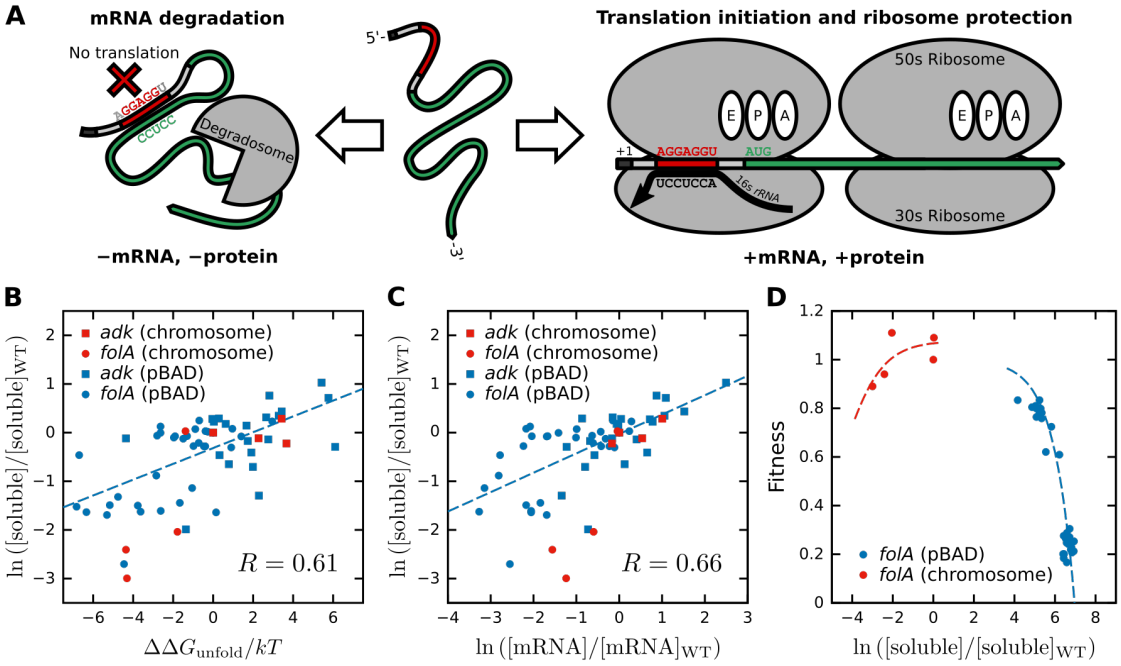
A
mRNA degradation
No translation
AGGAGGU
CCUCC
Degradosome
5'-
-3'
Translation initiation and ribosome protection
50s Ribosome
E
P
A
+1
AGGAGGU
AUG
UCCUCCA
16s rRNA
30s Ribosome
−mRNA, −protein
+mRNA, +protein
B
adk (chromosome)
folA (chromosome)
adk (pBAD)
folA (pBAD)
R = 0.61
ln ([soluble]/[soluble]WT)
ΔΔGunfold/kT
C
R = 0.66
ln ([mRNA]/[mRNA]WT)
D
Fitness
ln ([soluble]/[soluble]WT)

FIG. 6.

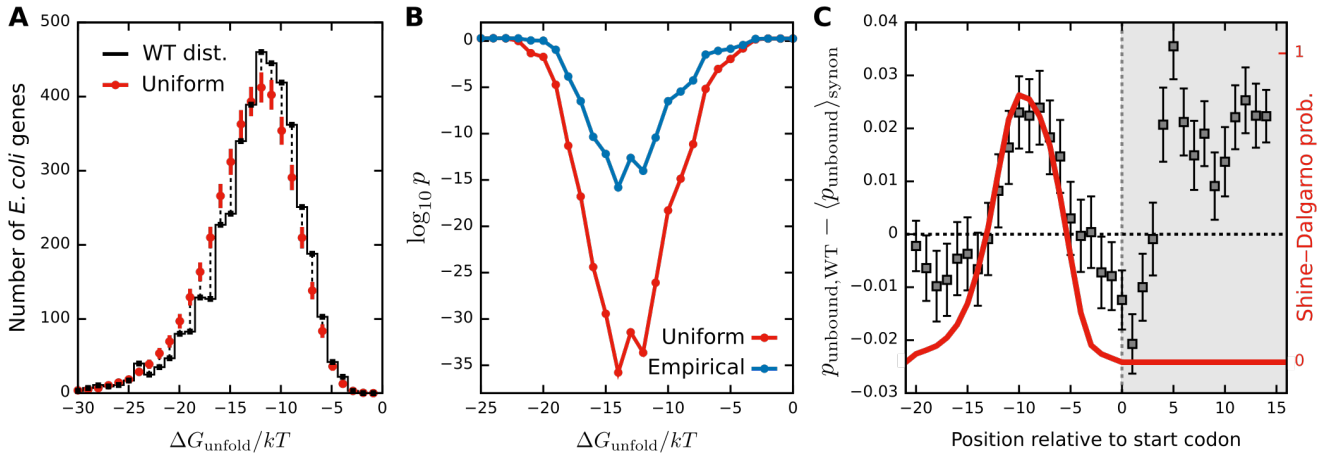
A
WT dist.
Uniform
Number of E. coli genes
$\Delta G_{\mathrm{unfold}}/kT$
B
Uniform
Empirical
$\log_{10} p$
$\Delta G_{\mathrm{unfold}}/kT$
C
$p_{\mathrm{unbound,WT}} - \langle p_{\mathrm{unbound}} \rangle_{\mathrm{synon}}$
Shine–Dalgarno prob.
Position relative to start codon

FIG. 7.

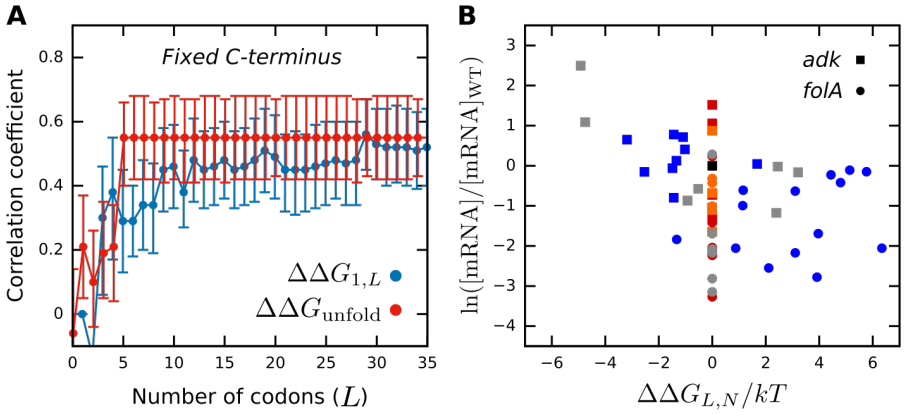
A
Fixed C-terminus
Correlation coefficient
Number of codons (L)
$\Delta\Delta G_{1,L}$
$\Delta\Delta G_{\mathrm{unfold}}$
B
adk
folA
$\ln([\mathrm{mRNA}]/[\mathrm{mRNA}]_{\mathrm{WT}})$
$\Delta\Delta G_{L,N}/kT$

FIG. S1.

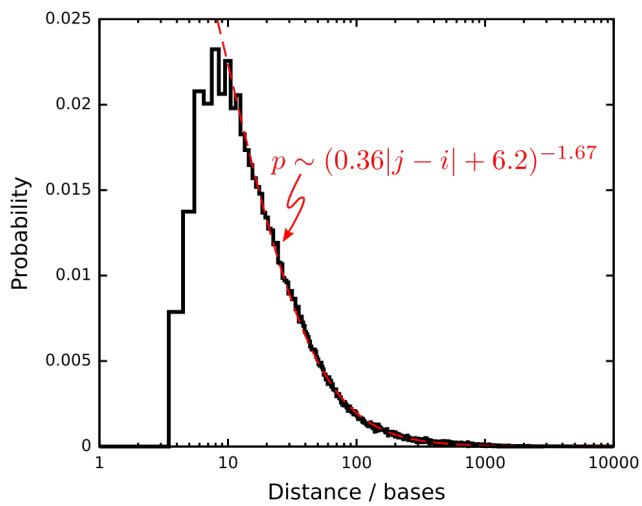
0.025
0.02
0.015
0.01
0.005
0
Probability
1
10
100
1000
10000
Distance / bases
$p \sim (0.36|j-i|+6.2)^{-1.67}$

FIG. S2.

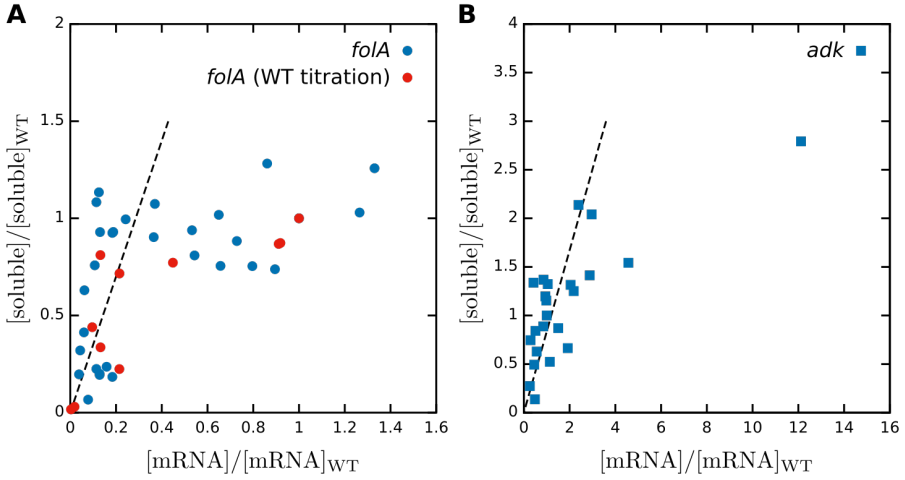
A
folA
folA (WT titration)
[soluble]/[soluble]WT
[mRNA]/[mRNA]WT
B
adk
[soluble]/[soluble]WT
[mRNA]/[mRNA]WT

FIG. S3.

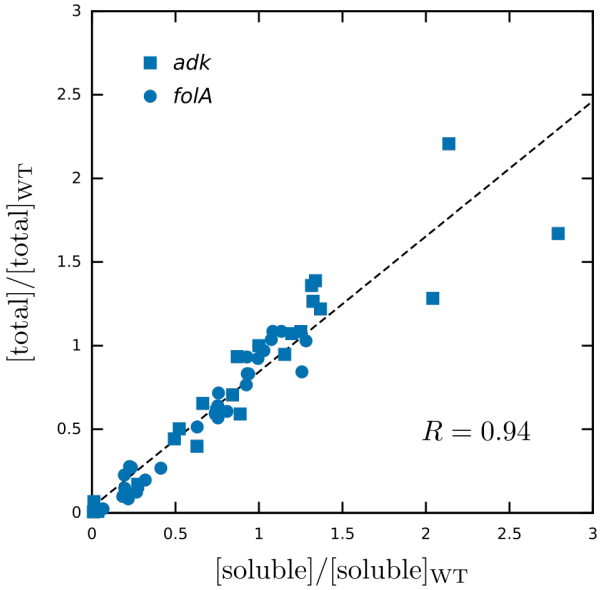
adk
folA
$R = 0.94$
[total]/[total]$_{\mathrm{WT}}$
[soluble]/[soluble]$_{\mathrm{WT}}$
0
0.5
1
1.5
2
2.5
3

FIG. S4.

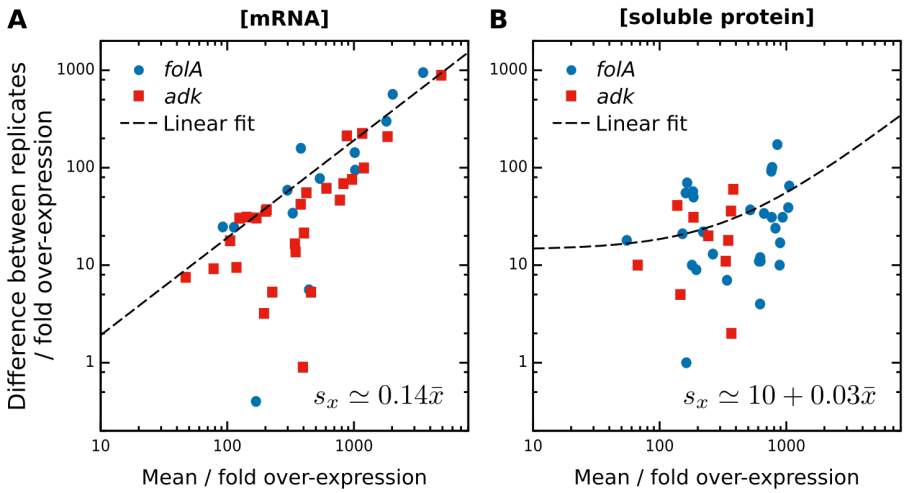
A
[mRNA]
B
[soluble protein]
folA
adk
Linear fit
Difference between replicates / fold over-expression
Mean / fold over-expression
$s_x \simeq 0.14\bar{x}$
$s_x \simeq 10 + 0.03\bar{x}$
1000
100
10
1

FIG. S5.