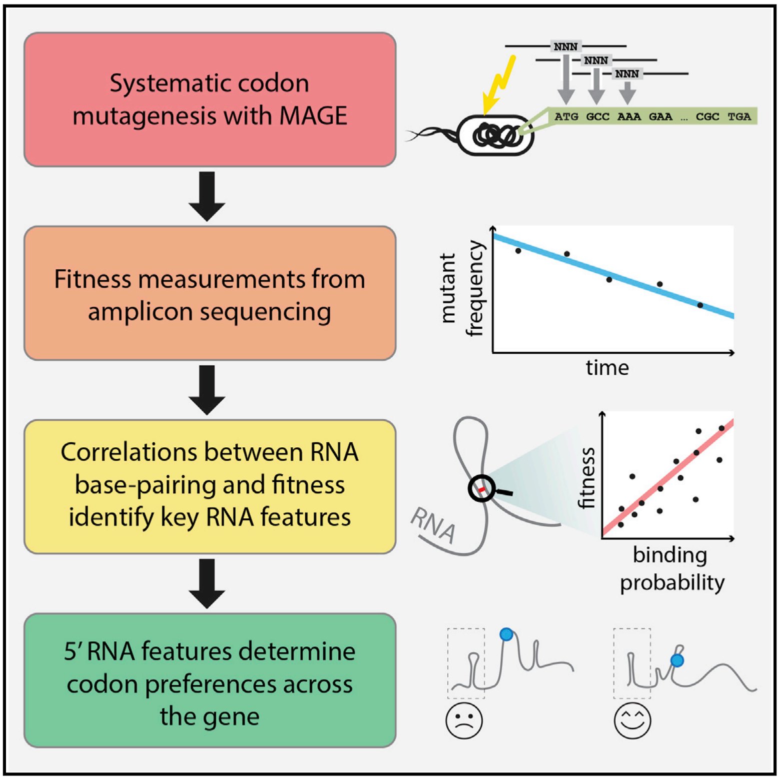
Article

Cell Systems

RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq

Graphical Abstract



Systematic codon mutagenesis with MAGE

Fitness measurements from amplicon sequencing

Correlations between RNA base-pairing and fitness identify key RNA features

5′ RNA features determine codon preferences across the gene

Authors

Eric D. Kelsic, Hattie Chung, Niv Cohen,
Jimin Park, Harris H. Wang,
Roy Kishony

Correspondence

hw2429@columbia.edu (H.H.W.),
rkishony@technion.ac.il (R.K.)

In Brief

Kelsic et al. develop and apply MAGE-seq to identify RNA structures that determine optimal codon preferences in an essential *E. coli* gene.

Highlights

- Systematic mutagenesis of an essential gene reveals context-dependent codon preferences
- 5′ RNA structure determines codon preferences far from the start codon
- Synonymous codons that disrupt the native 5′ RNA structure are more deleterious
- MAGE-seq enables rapid and quantitative phenotyping of mutant libraries in *E. coli*

Data Resources

E-MTAB-4020

Kelsic et al., 2016, Cell Systems *3*, 563–571
December 21, 2016 © 2016 Elsevier Inc.
http://dx.doi.org/10.1016/j.cels.2016.11.004

RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq

Eric D. Kelsic,[1,2,7] Hattie Chung,[1,7] Niv Cohen,[3] Jimin Park,[4] Harris H. Wang,[4,5,*] and Roy Kishony[1,6,8,*]
[1]Department of Systems Biology
[2]Wyss Institute for Biologically Inspired Engineering
Harvard Medical School, Boston, MA 02115, USA
[3]Faculty of Physics, Technion - Israel Institute of Technology, Haifa 3200003, Israel
[4]Department of Systems Biology
[5]Department of Pathology and Cell Biology
Columbia University Medical Center, New York, NY 10032, USA
[6]Faculty of Biology and Faculty of Computer Science, Technion - Israel Institute of Technology, Haifa 3200003, Israel
[7]Co-first author
[8]Lead Contact
*Correspondence: hw2429@columbia.edu (H.H.W.), rkishony@technion.ac.il (R.K.)
http://dx.doi.org/10.1016/j.cels.2016.11.004

SUMMARY

Synonymous codon choices at the beginning of genes optimize 5′ RNA structures for enhanced translation initiation, but less is known about mechanisms that drive codon optimization downstream within the gene. To understand what determines codon choices across a gene, we generated 12,726 in situ codon mutants in the *Escherichia coli* essential gene infA and measured their fitness by combining multiplex automated genome engineering mutagenesis with amplicon deep sequencing (MAGE-seq). Correlating predicted 5′ RNA structure with fitness revealed that codons even far from the start of the gene are deleterious if they disrupt the native 5′ RNA conformation. These long-range structural interactions generate context-dependent rules that constrain codon choices beyond intrinsic codon preferences. Genome-wide RNA folding predictions confirm that natural codon choices far from the start codon are optimized in part to prevent disruption of native structures near the 5′ UTR. Our results shed light on natural codon distributions and should improve engineering of gene expression for synthetic biology applications.

INTRODUCTION

Genome-wide biases in codon usage have been attributed to intrinsic preferences and to context-dependent effects. Intrinsic preferences refer to a general benefit of some codons over others, based on translation speed (Elf et al., 2003; dos Reis et al., 2004; Sharp and Li, 1987; Sørensen et al., 1989), translation accuracy (Drummond and Wilke, 2008; Stoletzki and Eyre-Walker, 2007), and other properties (Gingold and Pilpel, 2011; Novoa and Ribas de Pouplana, 2012; Plotkin and Kudla, 2011). In contrast, context-dependent codon preferences occur when codon effects depend on short-range neighboring sequences or on longer range distant sequences. Short-range context dependence appears, for example, when adjacent codons create inhibitory sequences (Gamble et al., 2016; Tats et al., 2008). Longer range dependences can be generated at the beginning of genes by selection for decreased strength of RNA secondary structures in the 5′ UTR (Bentele et al., 2013; Goodman et al., 2013; Gu et al., 2010; Kudla et al., 2009). However, less is known about long-range context-dependent effects in later gene regions.

Understanding context-dependent codon preferences within a gene requires systematically measuring the effects of codon substitutions in the gene's native chromosomal context. However, current methods for systematic mutagenesis and phenotyping of large mutant populations typically use plasmid-based expression rather than in situ chromosomal modification (Boucher et al., 2014; Fowler and Fields, 2014). Alternative chromosomal editing methods are usually unable to achieve the throughput necessary for comprehensive analysis. Computational genome-wide methods have been successful at identifying intrinsic preferences but may average out context-dependent effects that occur at specific positions. Such limitations have impeded the identification of factors that determine optimal codon usage within individual genes.

Here, we quantify intrinsic and context-dependent codon preferences throughout an essential gene by generating systematic libraries of single-codon and codon-pair mutants directly on the *E. coli* chromosome. Codon choice is particularly important within essential genes, where it can greatly affect fitness (Agashe et al., 2013; Lajoie et al., 2013; Lind et al., 2010). To understand such effects, we focused on understanding codon preferences within a single gene, *infA*, a highly expressed and essential gene of *E. coli* that encodes the universally conserved translation Initiation Factor 1 (IF1) (Croitoru et al., 2004; Gualerzi et al., 1989). This focus enabled us to comprehensively explore the fitness landscapes of single-codon and codon-pair mutants and to directly quantify and compare intrinsic versus context-dependent codon effects. We apply motif analysis and RNA structural analysis to identify key features that generate

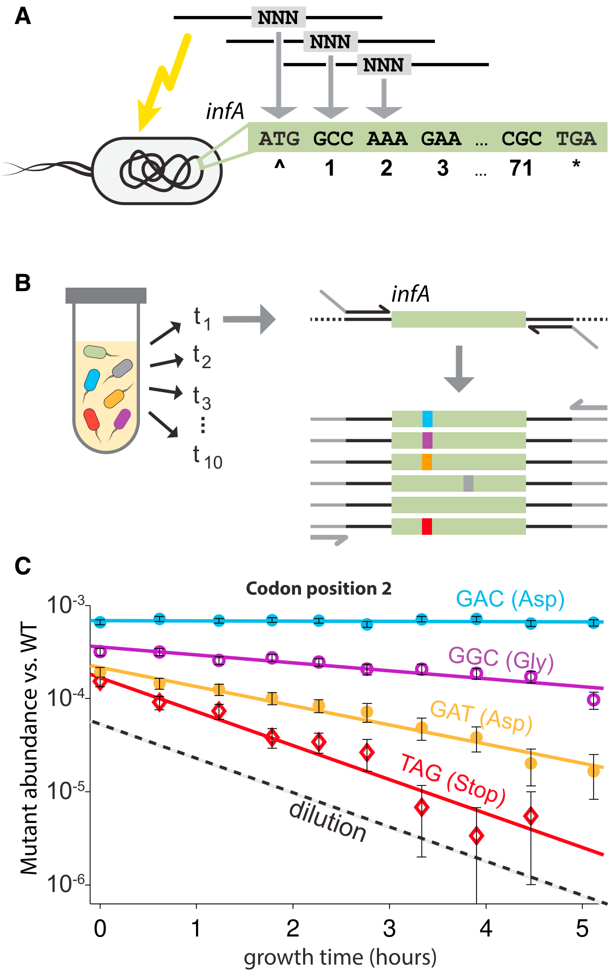


Figure 1. Systematically Generating and Measuring Fitness of All Single-Codon Substitutions across *infA* Using MAGE-Seq

(A) MAGE oligos for creating all single-codon mutants scanning along *infA* on the *E. coli* chromosome.

(B) Mutants were pooled and competed in continuous exponential growth. Samples were taken at every population doubling, and mutant frequencies were measured using deep sequencing of PCR amplicons.

(C) Mutant fitness is calculated from the slope of best-fit lines tracking mutant abundance relative to the wild-type allele over time. Dotted line shows the dilution rate, which is the expected slope for non-growing cells. Error bars are 2 SDs.

context-dependent codon preferences and generalize these findings with evolutionary conservation and genome-wide bioinformatics.

RESULTS

Comprehensive Single-Codon Mutagenesis of *infA* Using Multiplex Automated Genome Engineering Sequencing

To measure fitness of codon mutants in high throughput, we developed MAGE-seq, a method combining multiplex automated genome engineering (MAGE) (Wang and Church, 2011; Wang et al., 2009), an optimized iterative variant of single-stranded DNA-mediated recombineering (Sawitzke et al., 2013) that generates large-scale and systematically designed libraries of chromosomal variants (Figure 1A), with fitness measurements based on amplicon deep sequencing (Figures 1B and S1; Table S1; STAR Methods). Scanning across the genome using MAGE-seq enables the study of whole genes or regions of interest with arbitrary lengths by tiling: mutations are introduced using MAGE within regions of up to 30 nucleotides in length, then the pool of mutants is exposed to selective pressures that increases or decreases the abundance of individual mutants. Sequencing each MAGE pool with paired overlapping Illumina reads and quantifying mutant enrichment relative to the wild-type (WT) allele enables the assembly of fitness data for all mutants. We use separate barcodes during library prep to detect and reduce sources of experimental error. MAGE-seq thus enables rapid generation and quantitative phenotyping of mutant libraries and can be used to map fitness landscapes for functional elements anywhere on the *E. coli* genome.

We applied MAGE-seq to *infA*, a highly expressed and essential gene of *E. coli* that encodes translation Initiation Factor 1 (IF1). We created all possible single-codon mutants scanning along the entire length of the gene (63 codon mutants × 73 positions = 4,599 mutants; eight MAGE pools). We then inoculated the pooled mutant population into rich or minimal liquid media and grew them for ten generations while continuously diluting the cultures to maintain exponential phase. By sampling the population and deep-sequencing the mutated regions at multiple time points, we measured g, the exponential rate of change of mutant frequency relative to the wild-type allele (best-fit slope, Figure 1C). We then define the fitness of each mutant as $f = (g/d) + 1$, where d is the dilution rate of the culture ($d = 1.23$ and 0.43 doublings/hour for rich and minimal media, respectively). Thus, a mutant with $f = 1$ grows as fast as wild-type while a mutant with $f = 0$ does not grow and is depleted from the culture at the dilution rate. Performing the competition immediately after mutagenesis enabled us to detect all mutants in the library, including non-growing null mutants that are depleted at the rate of dilution (Figure 1C). Although MAGE is known to increase the background frequency of genomic mutations (Nyerges et al., 2014), the combination of a small library size and a large population ensured that each mutant was created many times ($>10^4$), thereby reducing the effects of background mutations. These experiments led to precise mutant fitness values that corresponded well with measurements from single-codon mutants generated via gene replacement (Figure S2) (Croitoru et al., 2004).

As expected, introducing early stop codons or mutating the start or stop codon of the gene is highly deleterious (Figures 2A and S3A). Nonsense mutations have null fitness values throughout the gene, except close to the C terminus, where the protein tolerates early truncation of 2 amino acids. Substitutions of the start codon have null fitness, except for alternative start codons such as GTG (Figure S3A). Mutations of the wild-type stop codon, which append 36 additional amino acids at the C terminus, are also deleterious.

Analysis of Single-Codon Mutant Fitness

Understanding synonymous mutations requires first accounting for the effects of different amino acids at each position. To separate the effects of codon and amino acid substitutions, we

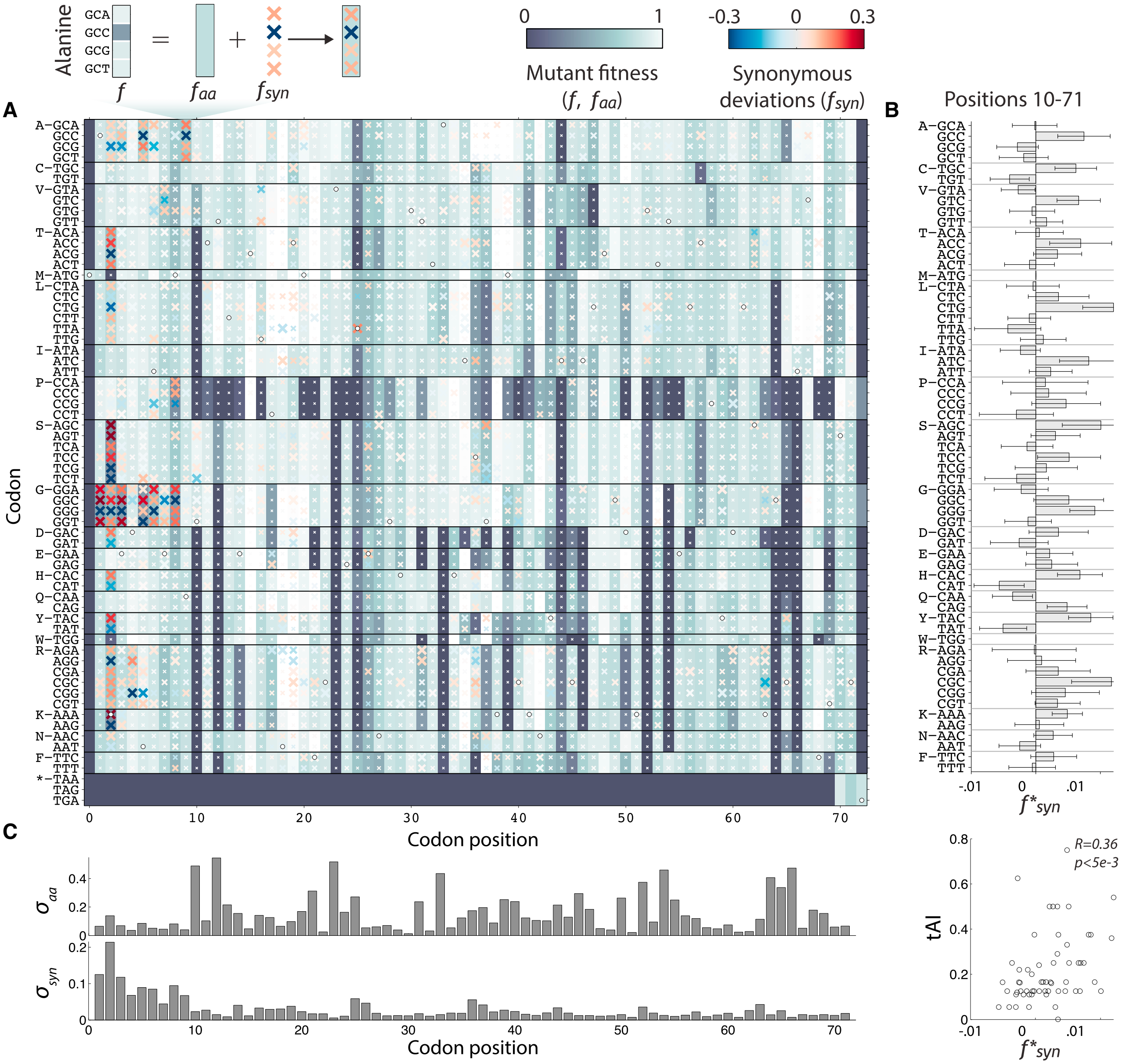


Figure 2. Context-Dependent Codon Preferences of *infA* Are Strongest at the Beginning of the Gene

(A) Fitness of all single-codon mutants of *infA* in minimal media (Figure S3 shows rich media). The optimal codon for each amino acid varies with position, indicating context dependence. Circles indicate WT codons; horizontal lines separate synonymous codons. Background color indicates average effect of each amino acid substitution (f_{aa}), while colored X's indicate synonymous fitness deviations (f_{syn}). For clarity, we remove X's for start and stop codons and set f_{syn} to zero in the later gene region for the most deleterious mutants (positions 10–71, f_{aa} < 0.75), which have higher measurement error (f_{std}). Larger X's indicate more significant Z scores (f_{syn}/f_{std}).

(B) Intrinsic codon preferences averaged over later gene regions (f^{*}_{syn}, positions 10–71). Error bars are 2 SEMs. Bottom: correlation between intrinsic codon preferences and the tRNA adaptation index (tAI).

(C) Comparison of SD of fitness for amino acid substitutions (σ_{aa}) and synonymous deviations (σ_{syn}, mutants with f_{aa} < 0.75 not included in calculation). Synonymous codon preferences are strongest at the beginning of the gene (positions 1–9).

averaged the fitness of all codons synonymous for each amino acid at each position to yield f_{aa}, then subtracted f_{aa} from each codon's fitness to create a synonymous fitness deviation matrix $f_{syn} = f - f_{aa}$ (Figures 2A and S3B). Focusing on the amino acid effects f_{aa}, we observed that specific sets of amino acid substitutions were deleterious at multiple positions across the gene. Principal-component analysis showed that the effects of most amino acid substitutions can be explained by only four principal components, reflecting key amino acid properties of hydrophobicity, flexibility, size, and charge (Figure S4; Table S2). The weights of these principal components vary across positions, indicating differential requirements for each property within the

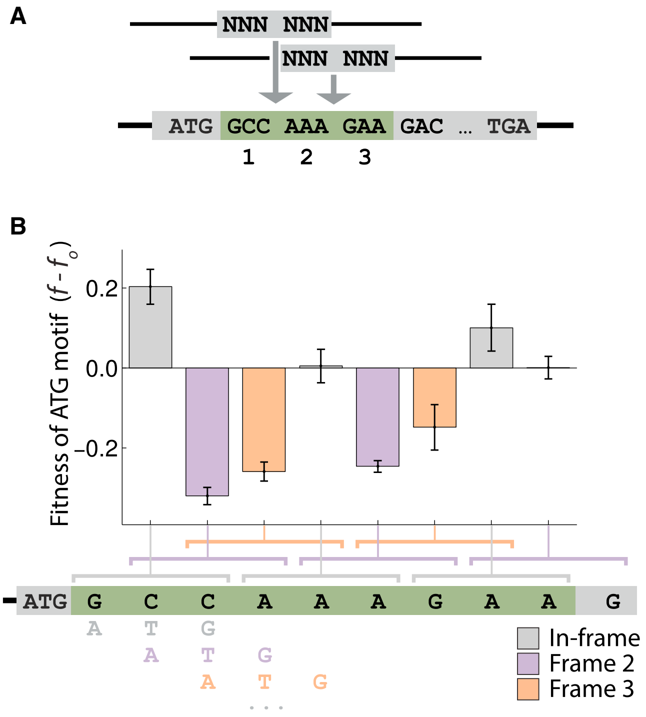


Figure 3. Analysis of Codon-Pair Interactions Near the Start of the Gene Reveals Deleterious Effects of Frameshifted Start Codons
(A) MAGE oligos for creating all codon-pair mutants at positions 1–2 and 2–3.
(B) Average fitness of codon pair mutants with in-frame ATG start codons (gray bars) versus frameshifted start codons (colored bars), relative to mean library fitness (f_o). Error bars are 2 SEMs.

core of the protein and at the interface with the 30S ribosome (Figure S4E). This analysis provides an unbiased way to reveal the key patterns of how fitness depends on amino acid properties throughout a gene.

We next focused on the effects of synonymous codons. Synonymous codon substitutions had the strongest effect in the beginning of the gene (positions 1–9), yet were also detectable in later gene regions, especially in rich media (positions 10–71) (Figure S5C). For the latter gene region, we measured intrinsic codon preferences, f^*_{syn}, by averaging synonymous deviations across all positions (positions 10–71; codon averages were calculated with individual measurements weighted by inverse fitness variances). Intrinsic preferences, while small (SD f^*_{syn} ~0.004), were correlated with genomic codon bias measures such as tRNA adaptation index (tAI) (dos Reis et al., 2004) (Figure 2B). At the beginning of the gene, synonymous deviations were strong, often dominating over the effects of amino-acid substitutions (Figure 2C). In this region, codon preferences depend strongly upon context (i.e., the best codon for a given amino acid differed across positions) (Figure S3).

Systematic Codon-Pair Mutagenesis Reveals Beneficial and Deleterious Sequence Motifs

To better understand the origins of codon preferences within *infA*, we measured the fitness of codon-pair mutants at the start of the gene, where synonymous fitness deviations were strong and highly context dependent. We used MAGE-seq to generate systematic codon-pair libraries at positions 1–2 and 2–3, yielding 8,127 mutants, and measure their fitness in rich media (Figure 3A). Within this library, codon choice at one position strongly affects the codon preferences of a neighboring position, as measured by codon-pair epistasis ($E_{ab} = f_{ab} - f_a f_b$, where f_{ab} are the fitness measurements of the codon-pair mutant and f_a, f_b are the fitness of the single-codon mutants in a best-fit null model; Figure S6). We investigated whether these strong epistasis effects could be explained by the local context of neighboring sequences and by long-range interactions with other positions on the RNA.

Investigating local context, we found beneficial and deleterious sequence motifs that create codon-pair interactions. We calculated the effect of all 2–5 nt sequences on codon-pair epistasis by correlating the presence of each motif with fitness, thereby identifying beneficial and deleterious sequence motifs (Table S3). The most significant motifs contain the frameshifted start codons ATG and GTG, which were strongly deleterious (Figures 3B and S7A). Consistent with bioinformatics predictions (Zur and Tuller, 2013), these measurements provide direct experimental confirmation of the deleterious consequences of frameshifted start codons near the beginning of a gene. In contrast, frameshifted stop codons were often beneficial (although with smaller effects than for frameshifted start codons), presumably because they help terminate frameshifted translation (Figure S7B). Although such beneficial and deleterious motifs were significantly correlated with fitness ($p < 10^{-18}$ and $p < 10^{-58}$, respectively), because of their rare occurrence, they explain only a small fraction of the total variance in codon-pair fitness (1.1% and 3.3%, respectively; Table S3).

Identification of Beneficial and Deleterious 5′ RNA Base Pairings

We next investigated longer range interactions and determined the key structural configurations of the *infA* RNA that are important for fitness. We computationally folded the first 100 nt of the 5′ end of the RNA transcript (including the entire 36 nt UTR) for each of the 8,127 codon-pair mutant alleles, yielding a base-pairing probability matrix P_{ij}^m for each mutant m, with i and j representing nucleotide indices. We then asked how the base-pairing probabilities P_{ij}^m correlate with the fitness f^m of these mutants; calculating the Pearson correlation coefficient R_{ij} = corr(P_{ij}^m, f^m) revealed base pairings that were highly correlated or anti-correlated with fitness (Figure 4A). The strongest indicators of beneficial fitness were base pairings that form a hairpin centered upstream of the start codon (near −18 nt; Figure 4B). To verify the importance of this hairpin structure, we systematically mutated pairs of positions on the two sides of the presumed hairpin and measured fitness in rich media. Almost every deleterious mutation on either side of the hairpin could be compensated to near WT fitness by mutating the opposite side of the hairpin to restore base pairing (Figure 4C; only one deleterious mutant could not be compensated, Figure S8). Together, these results support the existence of a beneficial hairpin RNA structure at the 5′ end of the gene.

Much of the fitness of the codon-pair mutant library is explained by the beneficial base-pairing content of each mutant 5′ RNA structure. For each mutant m, we calculated the extent to which its structure contained beneficial and not deleterious

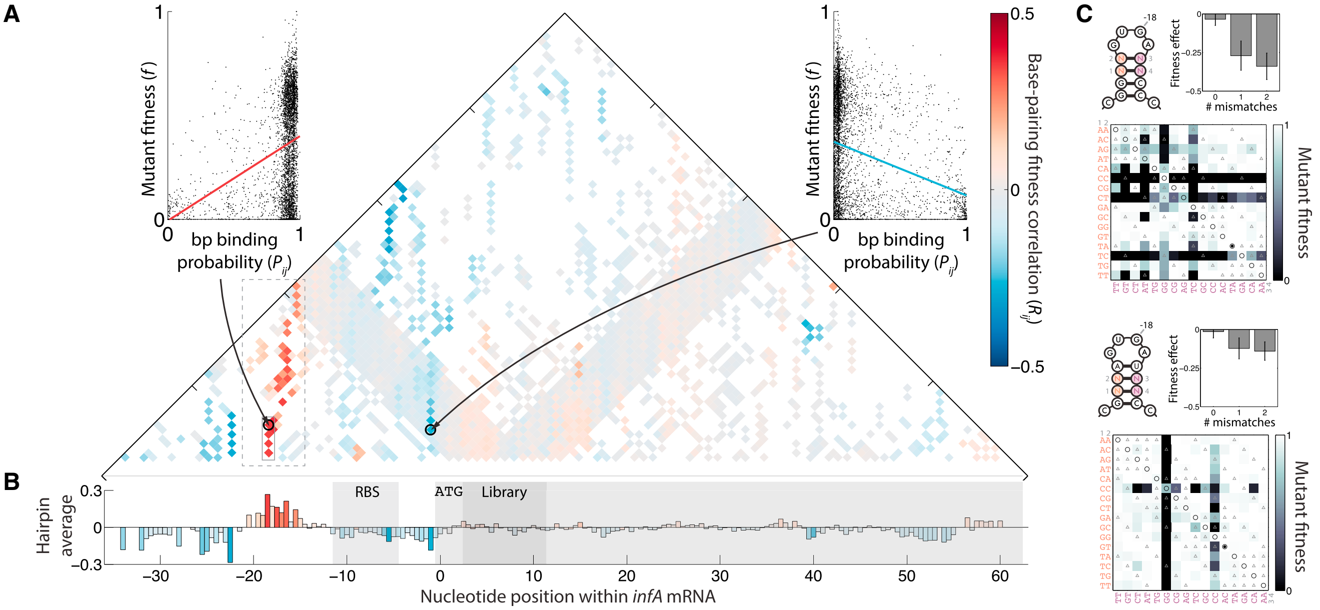


Figure 4. Correlations between RNA Base Pairing and Fitness Reveal Beneficial and Deleterious RNA Hairpin Structures

(A) Correlation matrix R_{ij}. Each base-pairing location is colored by the correlation between predicted base pair binding probabilities (P_{ij}) and codon-pair mutant fitness (f); white indicates base pairings that do not form for any mutants, red indicates positive correlation with fitness and blue indicates negative correlation with fitness. Insets show examples of base pairings with positive and negative correlations with fitness. The dotted gray box surrounds the approximate location of beneficial base pairings, while the solid gray box surrounds base pairings near the step loop as shown in (C).

(B) Average of the base-pairing correlations within RNA hairpins (vertical averages of A). The preferred RNA configurations are hairpins centered upstream of the start of the gene (near position −18 nt).

(C) Analysis of deleterious and compensatory mutations within the presumed beneficial RNA hairpin, showing mutated positions on top of the predicted minimum free energy RNA structure of the WT allele. Deleterious mutations near the step loop of the hairpin can be compensated by mutations on the opposite side of the hairpin that restore base pairing. Gray numbers 1–4 indicate positions of introduced mutations. Circles along the diagonal mark locations with perfect base pairing, triangles mark locations with one mismatch, and a black dot marks the WT 5′ UTR sequence. Insets show average fitness for zero to two mismatches. Error bars are 2 SEMs.

base pairings by summing R_{ij} scaled by P_{ij}^{m}, which we define as its RNA configuration score (*RCS*), $RCS^m = \sum_{i=1}^{100}\sum_{j=1}^{100} P_{ij}^{m} R_{ij}$. *RCS* was a strong predictor of codon-pair mutant fitness (explaining 21.7% of variance, $p < 10^{-10}$) and robustly predicted fitness even when using a small fraction of mutants (~1%) as training data (Figure S9A). *RCS* explained fitness better than the presence of frameshifted start and stop codons (in total 13.2% variance explained), Shine-Dalgarno-like sequences (12.2%) (Li et al., 2012) and RNA minimum free energy (12.6%) (Bentele et al., 2013; Goodman et al., 2013; Gu et al., 2010; Kudla et al., 2009) (Figure 5), as well as other metrics such as GC content (4.8%) and ribosome binding site accessibility (15.5%) (Salis et al., 2009) (STAR Methods). Furthermore, because RNA configuration was largely independent of these other metrics, combining *RCS* with multiple properties helped increase the fitness variance explained (up to 44.8%; Figure 5B). Some of the unexplained variance may come from potential fitness contributions at the protein level because of amino acid changes and also from measurement error. Cross-species comparative genomic analysis using *RCS* also showed signal for conservation of this key RNA structural configuration among closely related *infA* sequences (Figure S9B).

Constraints on 5′ RNA Structure Determine Downstream Codon Preferences

We next asked whether the requirement for specific RNA base pairings at the 5′ end could explain the fitness of codon changes not only at the beginning of the gene but also farther away from the start codon. Focusing on the library of single-codon changes across all positions of the gene, we quantified how much each codon change alters beneficial aspects of the 5′ RNA fold: we computationally folded the 5′ UTR along with the mutated gene sequence (including 20 nt of padding sequence after the mutant codon) and calculated the predicted effect of the first 100 nt of the 5′ structure on fitness using the base-pairing correlations inferred from the codon-pair mutants ($RCS^m = \sum_{i=1}^{100}\sum_{j=1}^{100} P_{ij}^{m} R_{ij}$, using P_{ij}^{m} of the folded single-codon mutant RNA and the previously determined R_{ij} from the codon-pair data). We found that the differences in *RCS* among synonymous codons significantly correlated with their synonymous fitness deviations (whole gene: $P^{whole} < 10^{-9}$; later gene positions 10–71: $P^{later} < 3 \times 10^{-4}$). Although the correlation coefficients were small ($R^{whole} = 0.24$, $R^{later} = 0.12$; Figure S5C), they were much larger than those for intrinsic codon preferences such as tAI ($R^{whole} = 0.02$ and $R^{later} = 0.03$) and f^*_{syn} ($R^{whole} = -0.01$ and $R^{later} = 0.03$; Figure 6A; more conservatively, fitting f^*_{syn} to the synonymous fitness deviations without weighting by the inverse of measurement uncertainty gives $R^{whole} = 0.17$ and $R^{later} = 0.13$, though this reflects overfitting of measurement error). We observed significant correlations between *RCS* and fitness deviations even at locations far from the start codon (Figures 6B and S9A). Indeed, we found that wild-type codon choices throughout most of the gene, even as far as 66 codons from the start, tended to avoid long-range disruption of 5′ structures with beneficial base pairing

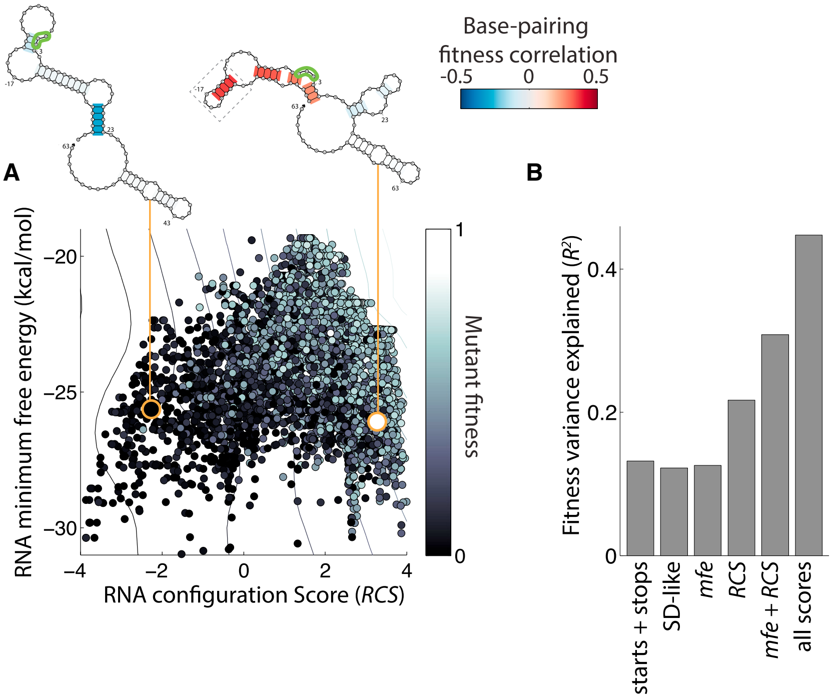


Figure 5. An RNA Configuration Score Explains Fitness Better Than RNA Folding Energy and Other Metrics

(A) Scatterplot of mutant minimum free energy (*mfe*) and RNA configuration score (*RCS*) for codon-pair mutants. Each point is colored by mutant fitness; contour lines are a best-fit regression; orange circles mark example mutants with comparable minimum free energy but bad versus good *RCS*, with example mutant RNA structures shown above: a green line surrounds the start codon; each RNA base pair is colored by the correlation of its base-pairing probability with fitness as in Figure 4A. The dashed box surrounds the beneficial hairpin region mutated in Figure 4C.

(B) Fitness variance explained by frameshifted start and stop codons, Shine-Dalgarno-like sequences (SD-like), minimum free energy, *RCS*, and linear combinations of these metrics.

(Figure 6C; $RCS^{WT} > RCS^{Synonymous\ Mutants}$, $p < 7 \times 10^{-3}$ for positions 10–66, $p < 0.02$ for positions 10–71). In contrast, the fitness dependence on RNA minimum free energy reversed between early and later gene regions: weaker RNA structure was beneficial near the start of the gene, while stronger RNA structure was beneficial in later regions (Figure S10B), consistent with genome-wide bioinformatics predictions (Gu et al., 2010) and with 5′ mutagenesis libraries (Goodman et al., 2013). These results show that 5′ RNA structural constraints can generate context-dependent codon preferences that extend far beyond the start of a gene.

Wild-Type Genes Tend to Use Codons that Do Not Disrupt Their 5′ Structures

As suggested by the wild-type codon choices within *infA*, we hypothesized that because the native RNA structure near each gene's UTR has certain optimal properties (such as accessibility for the ribosome or protection from nucleases via hairpin formation, etc.), downstream codon choices should be biased toward those that do not disrupt these 5′ structures. We tested this hypothesis across all genes of the *E. coli* genome. For each wild-type gene sequence, we calculated the structural similarity of its early RNA (40 nt of 5′ UTR + first 10 codons) when folded alone and when folded with an additional *k* subsequent codons, $S_{WT} = \Sigma_{ij} P_{ij}^{UTR+WT(10)} P_{ij}^{UTR+WT(10+k)}$ (P_{ij} is the RNA base-pairing probability, and *i* and *j* sum over the length of the early RNA). We then contrasted this WT similarity with the null similarity for alleles where the last 20 positions were randomized to synonymous codons, $S_{null} = \Sigma_{ij} P_{ij}^{UTR+WT(10)} P_{ij}^{UTR+WT(k-10)+Random(20)}$. Defining $\Delta S = S_{WT} - median(S_{null})$, we observed that *E. coli*'s wild-type codon choices are significantly better at maintaining 5′ RNA structures than codon randomized alleles ($\Delta S > 0$), with the bias toward codons that preserve these structures gradually decaying with distance from the beginning of the gene (up to 70 codons downstream; $p < 0.01$, Figure 6D). Enrichment for wild-type codons that do not disturb the native 5′ structure decreases slightly when controlling for the free energy of the wild-type allele (Figure S11), suggesting that this constraint is mediated at least in part by selection for stronger local RNA structures in later gene regions. These results show that preservation of native 5′ structures in certain genes can create selective pressure on codon usage even in distant regions of the RNA.

DISCUSSION

Systematic codon mutagenesis of the *E. coli infA* demonstrates that strong context-dependent codon preferences are created both by short-range sequence motifs and by long-range RNA structural constraints. In the short range, codons at the beginning of the gene are chosen to avoid deleterious sequence motifs such as frameshifted start codons and to enhance other motifs such as frameshifted stop codons. The beneficial effects of frameshifted stop codons are likely limited to sequences near the start codon, as recent work shows that frameshifted stop codons are generally depleted across coding regions (Tats et al., 2008). Long-range structural constraints also vary depending upon the gene region: consistent with previous reports (Bentele et al., 2013; Goodman et al., 2013; Gu et al., 2010; Kudla et al., 2009), weak RNA structure at the beginning of the gene and stronger structure in later regions is beneficial to fitness. However, in our *RCS* analysis the extent to which mutant RNA structures contained beneficial and not deleterious base pairings predicted the fitness of synonymous mutants both at the beginning of the gene and far beyond the start codon, augmenting more general predictors based on RNA folding energy and ribosome binding (Salis et al., 2009). The use of gene-specific empirical metrics such as *RCS* may therefore be useful in cases when general metrics fail to predict the effects of synonymous changes (Agashe et al., 2016; Knöppel et al., 2016). These results can also

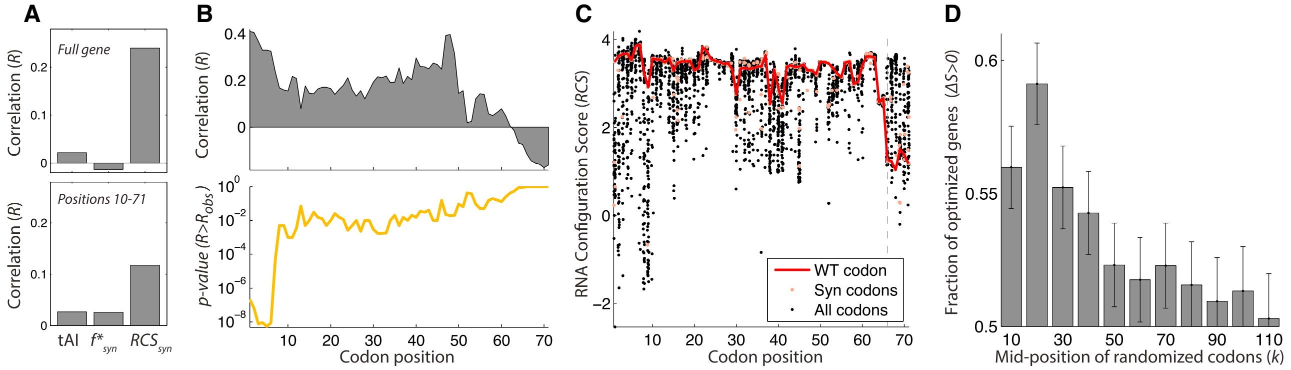


Figure 6. Preservation of RNA Structure at the Beginning of the Gene Determines Context-Dependent Codon Preferences throughout the Gene

(A) Correlation of synonymous fitness deviations with tAI, f^*_{syn}, and synonymous *RCS* deviations, for the full gene and for later regions (positions 10–71).

(B) *RCS* correlations as in (A), calculated for a sliding window of ten codons centered at each position. The p values show probability of measuring $R > R_{obs}$ based on a null model of shuffling synonymous codons within amino acids.

(C) *RCS* for single-codon mutants throughout the gene. Black dots show non-synonymous single-codon mutants, pink dots show synonymous single-codon mutants, and a red line connects the wild-type codons. Wild-type codons are near optimal with respect to *RCS* up until codon 66 (dashed line).

(D) Fraction of *E. coli* genes for which WT codons preserve the 5′ UTR configuration better than the median null allele, with 10 codons on either side of the indicated position being synonymously randomized (for position 10, we use only the WT 5′ UTR for the earlier region of the RNA; see STAR Methods). Fractions greater than 0.5 indicate genome-wide enrichment for WT codons that preserve these upstream RNA structures. Error bars are 2 SEMs.

help to explain biases toward stronger minimum free energy in later gene regions (Gu et al., 2010), because a stronger local fold in later regions of the gene is less likely to interact with upstream RNA structures. Although *infA* may be a special case in that it is essential, highly expressed and relatively small, our genome-wide analysis of RNA folding patterns highlights more generally the importance of upstream 5′ RNA structures in constraining codon choices even far away from the start of the gene. The existence of preferred RNA structures could be expected as differential accessibility for the ribosome and RNA nucleases can dramatically affect rates of translation initiation (Knöppel et al., 2016; Salis et al., 2009) and RNA degradation (Boël et al., 2016; Deana and Belasco, 2005; Presnyak et al., 2015). However, it will be important to profile large numbers of synonymous changes in other genes to determine how key RNA base pairings affect gene expression and fitness, and it will be interesting to compare such predicted base pairings to RNA structures measured in vivo (Del Campo et al., 2015). Our cross-species analysis of beneficial RNA structures among closely related *infA* alleles suggests that beneficial base pairings could also be identified using evolutionary conservation. Combining cross-species comparison with RNA structural analysis may therefore help predict the effect of synonymous mutations and refine evolutionary metrics for evaluating selection, such as *dN/dS*. Although our experimental results and analysis are specific to *E. coli*, it will be interesting to study how they extend to other bacteria and possibly to eukaryotic organisms, where it has also been shown that optimal codons can stabilize RNA folding (Gu et al., 2010; Katz and Burge, 2003) and avoid sequence motifs such as frameshifted start codons (Zur and Tuller, 2013). We expect that further application of these methods will help guide synthetic gene design, reveal evolutionarily important synonymous substitutions, and explain the forces that shape codon usage across species.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - ○ Generating mutant libraries with MAGE
  - ○ Preparing mutant populations for selection
  - ○ Sampling populations during selection
  - ○ Preparing population samples for measuring mutant frequencies via sequencing
  - ○ Analysis of mutant frequencies
  - ○ Measurement of mutant fitness
  - ○ Comparing mutant fitness measured using MAGE-seq to site-directed mutagenesis studies
  - ○ Analysis of amino acid changes on protein function
  - ○ Calculation of intrinsic codon preferences
  - ○ Codon-pair motif analysis
  - ○ RNA structure analysis
  - ○ Combining different metrics to explain the fitness of the codon-pair mutants
  - ○ Correlating codon preferences and RNA Configuration Score throughout the gene
  - ○ Testing the optimality of wild-type codons for forming beneficial RNA structures
  - ○ Evolutionary conservation of optimal RNA structures
  - ○ Genome-wide enrichment for codons that preserve 5′ RNA configuration
  - ○ Tips for high quality MAGE-seq experiments
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY
  - ○ Software
  - ○ Data Resources

SUPPLEMENTAL INFORMATION

Supplemental Information includes eleven figures, three tables, and four data files and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.11.004.

AUTHOR CONTRIBUTIONS

E.D.K., H.C., H.H.W., and R.K. designed the study. E.D.K., H.C., and H.H.W. developed the method. E.D.K., H.C., H.H.W., and J.P. did the experiments. E.D.K., H.C., and R.K. analyzed the data. E.D.K. and N.C. did the genome-wide folding analysis. E.D.K., H.C., H.H.W., and R.K. wrote the paper.

ACKNOWLEDGMENTS

We would like to thank the reviewers, R. Amit, Y. Pilpel, Y. Gerardin, D. Goodman, and M. Baym for providing feedback on early drafts of the manuscript, H. Salis for assistance with the RBS calculator, as well as other members of the Wang and Kishony labs for helpful discussions. E.D.K. acknowledges government support under and awarded by the Department of Defense, Office of Naval Research (ONR), National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a. H.H.W. acknowledges funding support from the National Institutes of Health (NIH) (1DP5OD009172-02, 1U01GM110714-01A1), the National Science Foundation (MCB-1453219), Sloan Foundation (FR-2015-65795), the Defense Advanced Research Projects Agency (W911NF-15-2-0065), and the ONR (N00014-15-1-2704). This work was supported in part by NIH grants R01 GM081617 (to R.K.) and European Research Council FP7 ERC Grant 281891 (to R.K.).

Received: November 1, 2015
Revised: May 25, 2016
Accepted: November 8, 2016
Published: December 21, 2016

REFERENCES

Agashe, D., Martinez-Gomez, N.C., Drummond, D.A., and Marx, C.J. (2013). Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. Mol. Biol. Evol. *30*, 549–560.

Agashe, D., Sane, M., Phalnikar, K., Diwan, G.D., Habibullah, A., Martinez-Gomez, N.C., Sahasrabuddhe, V., Polachek, W., Wang, J., Chubiz, L.M., and Marx, C.J. (2016). Large-Effect Beneficial Synonymous Mutations Mediate Rapid and Parallel Adaptation in a Bacterium. Mol. Biol. Evol. *33*, 1542–1553.

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

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., et al. (2016). Codon influence on protein expression in E. coli correlates with mRNA levels. Nature *529*, 358–363.

Boucher, J.I., Cote, P., Flynn, J., Jiang, L., Laban, A., Mishra, P., Roscoe, B.P., and Bolon, D.N.A. (2014). Viewing protein fitness landscapes through a next-gen lens. Genetics *198*, 461–471.

Carter, A.P., Clemons, W.M., Jr., Brodersen, D.E., Morgan-Warren, R.J., Hartsch, T., Wimberly, B.T., and Ramakrishnan, V. (2001). Crystal structure of an initiation factor bound to the 30S ribosomal subunit. Science *291*, 498–501.

Croitoru, V., Bucheli-Witschel, M., Hägg, P., Abdulkarim, F., and Isaksson, L.A. (2004). Generation and characterization of functional mutants in the translation initiation factor IF1 of Escherichia coli. Eur. J. Biochem. *271*, 534–544.

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

Del Campo, C., Bartholomäus, A., Fedyunin, I., and Ignatova, Z. (2015). Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function. PLoS Genet. *11*, e1005613.

dos Reis, M., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. *32*, 5036–5044.

Drummond, D.A., and Wilke, C.O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell *134*, 341–352.

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

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat. Methods *11*, 801–807.

Gamble, C.E., Brule, C.E., Dean, K.M., Fields, S., and Grayhack, E.J. (2016). Adjacent codons act in concert to modulate translation efficiency in yeast. Cell *166*, 679–690.

Giangrossi, M., Brandi, A., Giuliodori, A.M., Gualerzi, C.O., and Pon, C.L. (2007). Cold-shock-induced de novo transcription and translation of infA and role of IF1 during cold adaptation. Mol. Microbiol. *64*, 807–821.

Gingold, H., and Pilpel, Y. (2011). Determinants of translation efficiency and accuracy. Mol. Syst. Biol. *7*, 481.

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.

Gu, W., Zhou, T., and Wilke, C.O. (2010). A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput. Biol. *6*, e1000664.

Gualerzi, C.O., Spurio, R., La Teana, A., Calogero, R., Celano, B., and Pon, C.L. (1989). Site-directed mutagenesis of Escherichia coli translation initiation factor IF1. Identification of the amino acid involved in its ribosomal binding and recycling. Protein Eng. *3*, 133–138.

Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F.U., Kerner, M.J., and Frishman, D. (2008). Protein abundance profiling of the Escherichia coli cytosol. BMC Genomics *9*, 102.

Katz, L., and Burge, C.B. (2003). Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res. *13*, 2042–2051.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. *36*, D202–D205.

Knöppel, A., Näsvall, J., and Andersson, D.I. (2016). Compensating the fitness costs of synonymous mutations. Mol. Biol. Evol. *33*, 1461–1477.

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.

Lajoie, M.J., Kosuri, S., Mosberg, J.A., Gregg, C.J., Zhang, D., and Church, G.M. (2013). Probing the limits of genetic recoding in essential genes. Science *342*, 361–363.

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.

Lind, P.A., Berg, O.G., and Andersson, D.I. (2010). Mutational robustness of ribosomal protein genes. Science *330*, 825–827.

Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.

Novoa, E.M., and Ribas de Pouplana, L. (2012). Speeding with control: codon usage, tRNAs, and ribosomes. Trends Genet. *28*, 574–581.

Nyerges, Á., Csörgő, B., Nagy, I., Latinovics, D., Szamecz, B., Pósfai, G., and Pál, C. (2014). Conditional DNA repair mutants enable highly precise genome engineering. Nucleic Acids Res. *42*, e62–e62.

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., and Coller, J. (2015). Codon optimality is a major determinant of mRNA stability. Cell *160*, 1111–1124.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods *9*, 173–175.

Riley, M., Abe, T., Arnaud, M.B., Berlyn, M.K.B., Blattner, F.R., Chaudhuri, R.R., Glasner, J.D., Horiuchi, T., Keseler, I.M., Kosuge, T., et al. (2006). Escherichia coli K-12: a cooperatively developed annotation snapshot—2005. Nucleic Acids Res. *34*, 1–9.

Salis, H.M., Mirsky, E.A., and Voigt, C.A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. Nat. Biotechnol. *27*, 946–950.

Sawitzke, J.A., Thomason, L.C., Bubunenko, M., Li, X., Costantino, N., and Court, D.L. (2013). Recombineering: Highly Efficient In Vivo Genetic Engineering Using Single-Strand Oligos (Elsevier).

Senn, H., Lendenmann, U., Snozzi, M., Hamer, G., and Egli, T. (1994). The growth of Escherichia coli in glucose-limited chemostat cultures: a re-examination of the kinetics. Biochim. Biophys. Acta *1201*, 424–436.

Sharp, P.M., and Li, W.-H. (1987). The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. *15*, 1281–1295.

Sørensen, M.A., Kurland, C.G., and Pedersen, S. (1989). Codon usage determines translation rate in Escherichia coli. J. Mol. Biol. *207*, 365–377.

Starmer, J., Stomp, A., Vouk, M., and Bitzer, D. (2006). Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. PLoS Comput. Biol. *2*, e57.

Stoletzki, N., and Eyre-Walker, A. (2007). Synonymous codon usage in Escherichia coli: selection for translational accuracy. Mol. Biol. Evol. *24*, 374–381.

Tats, A., Tenson, T., and Remm, M. (2008). Preferred and avoided codon pairs in three domains of life. BMC Genomics *9*, 463.

Toprak, E., Veres, A., Yildiz, S., Pedraza, J.M., Chait, R., Paulsson, J., and Kishony, R. (2013). Building a morbidostat: an automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition. Nat. Protoc. *8*, 555–567.

Wang, H.H., and Church, G.M. (2011). Multiplexed genome engineering and genotyping methods applications for synthetic biology and metabolic engineering. Methods Enzymol. *498*, 409–426.

Wang, H.H., Isaacs, F.J., Carr, P.A., Sun, Z.Z., Xu, G., Forest, C.R., and Church, G.M. (2009). Programming cells by multiplex genome engineering and accelerated evolution. Nature *460*, 894–898.

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. J. Comput. Chem. *32*, 170–173.

Zur, H., and Tuller, T. (2013). New universal rules of eukaryotic translation initiation fidelity. PLoS Comput. Biol. *9*, e1003136.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
LB Lennox	RPI	L24065
MOPS Minimal media	Teknova	M2106
MOPS EZ Rich Defined media	Teknova	M2105
Deposited Data		
Sequencing FASTQ reads for Single Codon, Codon-pair, 5′ UTR hp mutant	ArrayExpress	https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-4020/
Fitness data for the *infA* single-codon	This Study	Data S1
Fitness data for the *infA* Codon-pair	This Study	Data S2
Fitness data for the *infA* 5′ UTR hp mutant	This Study	Data S3
Data on genome-wide RNA similarity scores in *E. coli*	This Study	Data S4
Experimental Models: Organisms/Strains		
EcNR2, *E. coli*: strain MG1655, mutS$^-$, λ-Red$^+$	Wang et al., 2009	https://www.addgene.org/26931/
Sequence-Based Reagents		
illustra bacteria genomicPrep Mini Spin Kit	GE Healthcare Life Sciences	28-9042-58
Q5 Hot-Start High-Fidelity 2X Master Mix	NEB	M0494S
Quant-iT DNA assay kit	Life Tech	Q-33120
MinElute PCR Purification Kit	QIAGEN	28004
KAPA Library Quantification Kit for NGS, Illumina platform	KAPA Biosystems	KK4824
Software and Algorithms		
SeqPrep	John St. John	https://github.com/jstjohn/SeqPrep
RBS Calculator	Salis et al., 2009	https://github.com/hsalis/Ribosome-Binding-Site-Calculator-v1.0
NUPACK	Zadeh et al., 2011	http://www.nupack.org
free2bind	Starmer et al., 2006	https://sourceforge.net/projects/free2bind/
Other		
Turbidostat for growth competition assay	Toprak et al., 2013	N/A

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents may be directed to Lead Contact Roy Kishony at the Technion - Israel Institute of Technology (rkishony@technion.ac.il).

METHOD DETAILS

Generating mutant libraries with MAGE

For the single-codon mutant libraries, we synthesized single-stranded 90nt oligonucleotides (IDT) with homology to the *E. coli* chromosome and with 3 consecutive degenerate N's in the center of each oligo (Table S1). We hand mixed with equal ratios of A, C, T and G to ensure equal representation of all codons. We combined oligonucleotides targeting 9 consecutive codons into 8 MAGE pools for multiplex transformation, with 10 consecutive codons for the last pool. For each MAGE pool, we performed 4 consecutive cycles of MAGE with oligo concentration of 10 μM, with recovery cultures in 3mL LB Lennox (RPI L24065) with Chloramphenicol at 30°C. These 4 cycles led to a final transformation efficiency of the mutant library of about 50%, with the remainder being the original wild-type sequence. We did all 4 cycles serially with the shortest possible recovery times (3-5 hr) to minimize loss of deleterious mutants from the population.

For the codon-pair libraries, we pooled oligos containing 6 degenerate N nucleotides at the center of the oligo, covering positions 1-2 and 3-4. We combined these with a pool of the three single-codon oligos for positions 1-3 (80% codon-pair oligos, 20% single-codon oligos).

For the 5′ UTR hairpin libraries experiment, we designed oligos with 2 degenerate N's on opposite sides of the hairpin stem, oligo #1 covering positions (−22, −21) and (−16, −15), oligo #2 covering positions (−23, −22) and (−15, −14), and oligo #3 covering positions (−24, −23) and (−14, −13). Positions are relative to the ATG start codon (A at position 0).

Preparing mutant populations for selection

Selection was conducted in either MOPS Minimal or MOPS EZ Rich Defined media (Teknova M2106, M2105). Starter populations for selection were prepared immediately after the last MAGE cycle to minimize loss of deleterious mutants. We used 2.5mL of the 3mL culture to inoculate the selection. After regrowth to an optical density of about 0.5 (λ = 600nm), the selection inoculum was spun at 10,000*g* for 1 min, washed twice with 1mL minimal media, and resuspended in 1.5mL of minimal media. The 1.5mL resuspension was split in two vials of 0.75mL, one for selection in minimal media and the other for selection in rich media. Two consecutive MAGE pools were combined to form one selection pool with 18-19 consecutive codon positions competing in the same culture; this minimized the number of selection cultures while fitting within the sequencing read length on the Illumina platform. Each selection pool was placed in a flat-bottom glass vial (Chemglass, CG-4902-08) with open-top screw caps (Chemglass, CV-3750-0024) and mini magnetic stir-bars (Big Science Inc., SBM2003MIC), brought to 12mL with the selection media, covered with AeraSeal (EXCEL) and grown at 30°C with continuous stirring.

For the codon-pair library, we measured 5 samples during 10 doublings in rich media. For the 5′ UTR hairpin library, we did 2 independent biological replicas of the MAGE transformations followed by growth in rich media taking 10 samples, sampling once per doubling.

Sampling populations during selection

Optical density (OD) was continuously monitored using an IR LED and photodiode (Toprak et al., 2013), calibrated against cultures measured on a spectrophotometer (λ = 600nm). 6mL of each population was sampled in 15mL Falcon tubes when the average OD of vials of the same media went above 0.3. Vials were replenished with 6mL of pre-warmed media, with slight adjustments in volume to keep the OD uniform across vials. Sample cultures were placed immediately on ice, then pelleted at 10,000*g* for 10min at 4°C. After discarding the supernatant, pellets were resuspended in 1mL phosphate buffered saline (1x PBS) and transferred to 1.5mL Eppendorf tubes, pelleted at 10,000*g* for 1min (discarding the supernatant) and then frozen at −80°C.

Preparing population samples for measuring mutant frequencies via sequencing

Genomic DNA was extracted from pelleted samples using the Illustra Bacteria Genomic Prep Mini Spin Kit (GE) eluting in 30 μL. Illumina adapters and sequencing barcodes were attached via two rounds of PCR. We used a pool of forward and reverse primers with a varying number of internal N sequences to improve clustering and read diversity (Table S1). We did two independent PCR replicas for each library DNA sample and for independent PCR replicas of DNA from a single colony of the WT. For PCR1, 2 μL DNA template was added to 10 μL Q5 Hot-Start High-Fidelity 2X Master Mix (NEB), 1 μL of 10 μM forward primer mix, 1 μL of 10 μM reverse primer mix and 6 μL PCR grade water, with program: 30sec at 98°C, 20 cycles of (10sec at 98°C, 15sec at 67°C, 15sec at 72°C), and 2min at 72°C. PCR1 products were then diluted 1:10 in PCR grade water. Each PCR2 reverse primer had a unique 8nt barcode sequence that allowed us to demultiplex samples from the same sequencing reaction (Table S1). For PCR2, 2 μL of diluted PCR1 product was added to 10 μL Q5 Master Mix, 1 μL forward primer, 1uL reverse primer with appropriate barcode (for sample time and pool) and 6 μL PCR grade water, with program: 30sec at 98°C, 10 cycles of (10sec at 98°C, 30sec at 72°C), and 2min at 72°C. We checked each PCR product for a clean band on an agarose gel, measured DNA concentrations using the Quant-iT DNA assay kit (Life Tech, Q-33120) kit, added a small amount of water to each sample to normalize their concentrations, then pooled all samples and cleaned-up using QIAGEN MinElute PCR Purification Kit. Final library DNA concentrations were quantified via qPCR with KAPA Library Quantification Kit (KAPA Biosystems) and sequenced with overlapping 100bp paired end reads on the Illumina HiSeq platform.

Analysis of mutant frequencies

Every position within the mutant library was covered by overlapping paired-end reads, enabling highly accurate measurements of mutant frequencies. SeqPrep (github.com/jstjohn/SeqPrep) was used to merge each pair of reads into a consensus sequence. We used a combination of Python and Perl scripts to trim reads to the library region, remove any reads with degenerate sequences or low quality scores (< Q20), and to count the number of times we saw each allele. We removed reads that contained multiple mutations outside of the target single-codon or codon-pair libraries. We obtained an accurate measurement of errors originating from the WT allele by sequencing a library prepared using WT genomic DNA at high coverage. Miscalled bases occurred on average for less than 1 in 10^4 reads per position per base (Figure S1A). Miscalled mutant counts were generally very small for sequences that were more than one SNP different from the WT allele. Mutant counts were highly reproducible across independent PCR replicas (Figure S1B), enabling us to subtract the expected number of counts that were coming from sequencing miscalls of the WT allele (Figure S1C) to obtain corrected counts for each allele. We divided corrected counts for each allele by the number of WT counts to obtain relative mutant frequencies (Figures S1E–S1G).

We used two independent PCR replicas for each library to assess errors introduced by sequencing and to remove outliers coming from PCR amplification. Comparing the frequencies of each mutant across replicas, we determined that measurement errors were just slightly above what would be expected from counting noise (Figure S1D).

Measurement of mutant fitness

We calculated relative mutant growth rates *g,* by finding the best linear fit to the logarithm of rate of change in mutant frequencies relative to the WT allele over time. We normalized this by *d,* the expected dilution rate in the culture due to growth of the WT. The codon-pair libraries and UTR hairpin experiments were done on different days and therefore under slightly different experimental conditions, and we observed slower growth than for the single mutant libraries during the initial periods of selection. Therefore for these datasets we set *d* so that the stop codon null mutants had an average fitness of 0.

For the UTR hairpin libraries we measured fitness for each replica separately (Figure S2E). Fitness values were normalized by the rate of dilution for null mutants (stop codons at position 1) and averaged across replicas.

Since we measure all mutant frequencies relative to the WT allele, our fitness values represent fitness relative to WT. As the WT allele will have 2 offspring after a single generation, the Malthusian fitness of a mutant with fitness *f* is therefore $2^f/2$.

Comparing allele frequencies between media identified positions where amplification of residual MAGE oligos sometimes prevented accurate measurements of allele frequencies for the single-codon libraries (as determined by differences in barcode frequencies or by comparing total number of reads for each mutant at each position). Such oligo amplification only occurred for positions near the center of sequencing pools, where MAGE oligos bridged the sequencing PCR primers. Altogether these events occurred at only a few time points and positions and usually affected only a single barcode (less than 2% of mutants in rich media and less than 0.3% of mutants in minimal media had issues with both barcodes, Figures S1E and S1F). We removed these time points from the growth fits at any positions where they occurred, which enabled fitness measurements of all mutants in both media except at positions 43-44 in rich media. We prevented further oligo amplification for the codon-pair libraries by increasing the number of washes following sampling from the selection culture (Figure S1G).

Comparing mutant fitness measured using MAGE-seq to site-directed mutagenesis studies

Previous works have measured the growth rates of a few *infA* mutants using plasmid-based expressions systems or site-directed mutagenesis and allele replacement (Croitoru et al., 2004). While these studies did not create synonymous variants, comparison of our measured growth rates with single amino acid *infA* mutants shows good agreement: comparing in rich media at 30°C we observed Pearson correlation coefficient of $R = 0.92$ ($p < 10^{-4.7}$), while in minimal media at 30°C we observed $R = 0.76$ ($p < 10^{-2.4}$) (Figure S2D). Note that we used MOPS rich and minimal medias while Croituro et. al used LB for rich media and an alternative glucose limited minimal media (Senn et al., 1994).

Analysis of amino acid changes on protein function

Fitness values were similar between rich and minimal media, though mutations were slightly less deleterious in minimal media (Figure S2). For f_{aa}, we used the average mutant fitness between minimal and rich media (using only minimal media fitness for positions 43-44 due to greater uncertainty for rich media in these positions). We applied Principal Component Analysis (PCA) to f_{aa} after removing start, stop and nonsense codons and centering the columns by their mean. The resulting principal components (PCs) are vectors of length 20 describing the key patterns of how fitness depends on amino acid choice. The first four PCs explain 88% of the covariance in the matrix. We find that these PCs correlate strongly with distinct biochemical properties from a database of amino acid indices (Kawashima et al., 2008): PC1 with buried residues and hydrophobicity, PC2 with flexibility (propensity for chain reversal), PC3 with size (steric hindrance) and PC4 with net charge (Table S2). For structural analysis and for identifying positions where IF1 interacts with the ribosomal RNA, the *infA* sequence was aligned to the IF1 sequence of *Thermo thermophilus* (Carter et al., 2001) (single insertion between residues 3-4 of *infA* and a single deletion at residue 70).

Calculation of intrinsic codon preferences

We averaged synonymous fitness deviations (f_{syn}) across positions 10-71, weighting the value at each position by the inverse of standard deviation squared, with standard deviations based on uncertainty of the slope in the linear fits (Figure S5). This use of the weighted average reduces the contribution of positions with low fitness, such as deleterious amino acid substitutions, where measurement error is higher.

Codon-pair motif analysis

The multiplicative pairwise null model was created by assigning fitness values f_a and f_b between 0 and 1 to each codon in the pair and calculating the fitness of codon-pairs as $f_{ab} = f_a f_b$. We optimized the fitness values for each of the 256 individual codon values (64 codons × 2 codons in each pair × 2 codon-pairs) by minimizing the sum of the residuals squared when the null model was subtracted from the actual fitness values. Mutants with $f<0$ were set to $f = 0$. We created all possible 6nt motifs matching *2-5* nucleotides across the codon-pair. Codon-pair interactions were analyzed by finding the deviations in mutant fitness from the expected fitness of the codon-pair based on the multiplicative null model. We tested for correlations between this fitness deviation and the presence or absence of each motif. Many of the most deleterious motifs are created by the frameshifted start codons ATG and GTG. Table S2 shows all motifs with significant *p* values after adjusting for multiple hypothesis testing using the Bonferroni correction (M_N different hypotheses, where M_N is the number of possible motifs with *N* nucleotides).

RNA structure analysis

RNA Configuration Score (*RCS*) was more predictive of fitness than RNA minimum folding energy and other RNA metrics. When compared to RNA folding energy as measured by minimum free energy, *RCS* explained a greater amount of variance (21.7% for *RCS* versus 12.6% for minimum free energy, both with $p < 10^{-10}$). *RCS* also exceeds predictions based on RBS accessibility (Salis et al., 2009) (15.5%, $p < 10^{-10}$; Methods), although correlation between these metrics ($R = 0.295$, $p < 10^{-10}$) suggests that the beneficial RNA structures impact fitness by at least in part by modulating rates of translation initiation. Other RNA features contributed to fitness but to a smaller extent: for example, greater affinity of Shine-Dalgarno-like mutant sequences for the ribosome's anti-Shine-Dalgarno site (which leads to ribosomal stalling (Li et al., 2012)) explained 12.2% of variance ($p < 10^{-10}$), while the intrinsic property of GC content explained only 4.8% of variance ($p < 10^{-10}$). Combining multiple metrics, we found that *RCS* was largely orthogonal to effects of minimum free energy: combining *RCS* and minimum free energy in a linear model explained 30.9% of fitness variance ($p < 10^{-10}$), while including these two metrics with the effects of Shine-Dalgarno-like sites and all frameshifted start and stop codons explained 44.8% ($p < 10^{-10}$), highlighting how context-dependent codon preferences are simultaneously affected by multiple RNA properties.

RNA structures and RNA minimum folding energies were calculated using the NUPACK (Zadeh et al., 2011) 'pairs' and 'mfe' functions, with settings: Temperature = 30, material = 'rna', dangles = 'all', and with a 0.001 cutoff for outputting probabilities of base-pairing. RBS scores were calculated using the RBS calculator (Salis et al., 2009), modified to run at temperature = 30. All codon-pair calculations use the first 100nt of the *infA* RNA (from the primary promoter P2(Giangrossi et al., 2007), which yields an mRNA with 36nt in the 5′ UTR). While calculations with RNAs that extended further downstream yielded similar results, RNA secondary structural configuration explained more fitness variance in codon-pair mutants with shorter mRNA's (100nt) than when using the entire mRNA including the 3′ UTR (319nt). Minimum-free-energy structures in Figure 6A were created using the NUPACK online GUI (http://www.nupack.org), and modified to color basepairs by their fitness correlation scores. Affinity of codon-pair mutants for the anti-Shine-Dalgarno site were calculated using *free2bind* (Starmer et al., 2006) with the first 15 nt of each mutant gene (including the start codon) and the anti-Shine-Dalgarno probe sequence 3′CCUCCU5′, with Temperature = 30.

Combining different metrics to explain the fitness of the codon-pair mutants

Fitness correlations were calculated for all codon-pair mutants excluding mutants containing stop codons. We combined different metrics using a linear model. For metrics $\{X_1, X_2 \dots X_n\}$, we solve for the coefficients $\{a_0, a_1, a_2 \dots a_n\}$ that minimize the root mean square distance between $M = a_0 + a_1X_1 + a_2X_2 + \dots + a_nX_n$ and mutant fitness values f, then calculate the Pearson correlation between M and f. For the frameshifted start codons we allow coefficients for each start codon to be chosen independently: $M = a_0 + a_1X_{ATG} + a_2X_{TTG} + a_3X_{GTG}$, and similarly for frameshifted stop codons.

Correlating codon preferences and RNA Configuration Score throughout the gene

We calculated the *RCS* of single-codon mutants by folding the RNA from the transcription start site until 20nt after the mutated codon using the same settings as for the codon-pair library. The R_{ij} matrix was calculated only using the codon-pair data. We calculated the *RCS* score based on the RNA configuration of single codon mutants within the first 100nt of the folded RNA (subset of $P_{ij}{}^{m}$ matrix with $1 \leq i,j \leq 100$). We calculated synonymous fitness deviations (f_{syn}) and synonymous *RCS* deviations (RCS_{syn}) by subtracting out the average affect of each amino acid at each position. We tested for correlations between the synonymous deviation matrices since this avoids the potential for false positive or negative correlations due to amino acid effects (as the first two nucleotides of each codon are also correlated with amino acid properties) From all calculations we removed amino acids with only one codon (methionine and tryptophan) since synonymous deviations are zero by default, and we excluded stop codons and all codons for which $f_{aa} < 0.25$ (< 7% of mutants), since uncertainty in fitness measurements is highest for highly deleterious amino acids. For Figure 6B and Figure S10 we used single codon fitness measurements from rich media since the codon-pair fitness values used for calculating R_{ij} were also measured in rich media.

We calculated Pearson correlation coefficients and associated p values when correlating RCS_{syn} values and f_{syn} values, using a sliding window of 10 codons centered at each codon position. Codons within the first half of the gene make the strongest contribution to these correlations (positions 1-9: $R = 0.33$, $p < 5{*}10^{-8}$, positions 26-35: $R = 0.16$, $p < 6{*}10^{-5}$), although correlations are still significant at later positions (positions 51-60: $R = 0.12$, $p < 0.015$) (Figure S10). We calculated p values using a null model with random shuffling of synonymous codons within amino acids (10^3 trials), with p values based on the number of random shufflings with either higher or lower correlation coefficients than those of the actual *RCS*, or using the cumulative probability distributions with mean and variance given by the null model when there were 0 trials with higher/lower correlations than for *RCS*. We calculated the local effects of RNA minimum free energy by computing the folding energies of RNAs containing the mutant codon and 20nt of upstream and downstream sequence (43nt in total). Local RNA folding energy was positively correlated with synonymous codon preferences only at the beginning of the gene (positions 1-9: $R = 0.33$, $p < 5{*}10^{-7}$) and was negatively correlated with fitness in later regions (positions 10-71, $R = -0.051$, $p < 0.006$; Figure S10B, Methods). These correlations with local minimum free energy match genome-wide trends toward reduced RNA folding at the beginning of genes and toward stronger folding in later regions (Gu et al., 2010).

Testing the optimality of wild-type codons for forming beneficial RNA structures

To identify whether WT codon choices within *infA* were optimized for forming specific RNA structures, we tested whether wild-type codons in later regions of the gene had significantly better *RCS* than other single-codon mutants: we calculated the average difference between wild-type codon *RCS* and average mutant *RCS* for a given number of positions, and then calculated a Z-score and p value for this difference (by comparing the actual value to a distribution of scores obtained by randomly shuffling synonymous codons within amino acids at each position, 10^5 random trials). We found that WT codons had near optimal *RCS* values compared to all possible single-codon mutants, up until codon 66. We tested for significance when all single-codon mutants were allowed (*all*), and also when only synonymous codons for the wild-type amino acid were included in the average (*syn*). Including codon positions 10-66: $P_{all} = 4.5*10^{-5}$, $P_{syn} = 0.0063$; for the entire gene (positions 1-71): $P_{all} = 1.6*10^{-4}$, $P_{syn} = 0.018$.

Evolutionary conservation of optimal RNA structures

9,824 IF1 ortholog alignments from the *infA* protein sequence of *E. coli* were generated via HHblits (Remmert et al., 2011) using e-value sensitivity of $E = 10^{-2}$. We retrieved the gene sequence for each protein ortholog and took 36nt from the 5′ UTR and the first 64nt of the gene, yielding 3821 unique sequences. We calculated *RCS* for each sequence and for *infA* alleles with random nucleotide substitutions on the WT *E. coli* sequence (maintaining average nucleotide content). *RCS* of closely related *infA* orthologs was higher than for randomly mutated sequences, indicating conservation of this RNA configuration (Figure S9).

Genome-wide enrichment for codons that preserve 5′ RNA configuration

For each gene in the *E. coli* genome K-12 MG1655(Riley et al., 2006), we folded RNA from the beginning of the gene alone and stored the probabilities of any two positions within the RNA being base-paired (base-pairing matrix P_{ij}^{early}). For the early region of the gene, we use 40nt from the 5′ UTR, the start codon and the next 10 codons from the WT sequence ($P_{ij}^{early} = P_{ij}^{UTR+WT(10)}$), except when testing for optimality of the first 20 codons (Figure 6D, position 10) in which case we use only 40nt from the 5′ UTR and the start codon ($P_{ij}^{early} = P_{ij}^{UTR+WT(0)}$). We then calculated the binding probability matrix when the same RNA was folded together with the addition of *k* downstream codons (either all WT codons or randomly choosing synonymous codons in the final 20 positions: matrices $P_{ij}^{UTR+WT(k+10)}$ and $P_{ij}^{UTR+WT(k-10)+Random(20)}$). For the null allele the final 20 codons were randomly chosen according to their genomic frequencies. Positions in Figure 6D and Figure S11 indicate the middle of the synonymously randomized region. We calculated the extent to which bindings near the beginning of the gene were disrupted when adding additional sequences using the similarity metrics $S_{WT} = \Sigma_{ij} P_{ij}^{UTR+WT(10)} P_{ij}^{UTR+WT(10+k)}$, and $S_{null} = \Sigma_{ij} P_{ij}^{UTR+WT(10)} P_{ij}^{UTR+WT(k-10)+Random(20)}$ where *i* and *j* sum over all of the base-pairing probabilities of the early RNA. To test whether WT codon choices are optimized to prevent disruption of 5′ structures, we compared S_{WT} to $median(S_{null})$. Thus, genes for which $\Delta S = S_{WT} - median(S_{null}) > 0$ are enriched for codons that preserve the RNA configuration of 5′ RNA structures near the beginning of the gene. For each gene, we calculated the $median(S_{null})$ from 300 random trials at each position. For RNA folding we used ViennaRNA (Lorenz et al., 2011). Error bars in Figure 6D and Figure S11 are calculated based on the standard error of the binomial distribution. We did not detect significant correlation ($p < 0.05$) between ΔS and gene expression data (Ishihama et al., 2008).

Tips for high quality MAGE-seq experiments

MAGE-seq enables systematic mutagenesis of functional elements anywhere on the *E. coli* genome. The method comprises 3 steps:

1) Generation of mutant library via MAGE (Wang and Church, 2011; Wang et al., 2009).
2) Selection on the mutant population with a functional assay: diverse functional assays are applicable (growth, FACS, etc.) so long as the assay differentially alters the frequencies of functional and non-functional mutants within the population.
3) Deep-sequencing mutated regions before and after selection: converting mutant counts into mutant frequencies relative to the WT allele, and finally calculation of mutant fitness by analyzing the change in mutant frequencies due to the assay.

As in standard MAGE transformations, MAGE oligos should be approximately 90nt with mutations surround by flanking homologous sequences targeting the lagging strand of the genome. Oligonucleotides libraries can be combined into a single pooled reaction as long as the mutated positions are within about 30nt of each other, which is necessary to avoid introducing multiple mutations. In this way, transformation with a second MAGE oligo will restore a pre-existing mutation back to the WT sequence. For measuring fitness landscapes across genes or regions longer than 30nt, split the target region into multiple MAGE pool regions of approximately 30nt or less in length. MAGE transformations typically achieve efficiency of around 20% without selection. In order to decrease the WT fraction of the population (and enable greater sequencing depth of the mutant library), we do 4 MAGE transformations serially (with the minimum required recovery time between transformations), after which WT abundances drops to about 50% of the population. This process typically takes around 24 hr. When mutations affect growth rates, we recommend beginning the competition experiment immediately after library generation, so that depletion of null mutants in the population can be directly observed by sequencing.

Thorough washing of the pellet is especially critical after the final MAGE transformation (2 washes in fresh media recommended) as it removes residual oligonucleotides from the culture that could be amplified during amplicon sequence of the genome, creating high background that prevents accurate measurement of mutant frequencies via sequencing. Although the oligos for our single codon mutant library scanned along across all positions with constant length homology regions on both sides of the mutant codon, through later experiments we determined that best practice is to hold constant the 30nt flanking sequences on either side (L and R) of the MAGE region, and scan the degenerate sequences (N) along within the middle of the 30nt MAGE region (for example LN-----R, L-N----R, …, L----N-R, L-----NR, etc.). In this way all MAGE oligos will have at least 30nt of homology outside of any other mutation, while no MAGE oligo will bridge the PCR1 primers. In any case we recommend ordering MAGE oligos and PCR1 primers on separate plates or in separate tubes, to minimize the potential for accidental cross contamination.

Multiple MAGE pools can be combined during selection or during sequencing, provided that the entire library region of every pool is covered by paired end overlapping reads. For example, when sequencing with 100bp paired-end Illumina reads, two adjacent 30nt MAGE pools may be combined into one 60nt sequencing pool prior to competition experiments, reducing the number of samples. Overlapping paired-end reads are beneficial as this reduces the frequencies of sequencing errors. For any MAGE pool being sequenced, we also recommend sequencing an amplicon of the WT allele, as this allele is most abundant in the population following transformation and is therefore the greatest source of sequencing errors. Sequencing errors coming from this WT control are highly reproducible on the Illumina platform, which enables subtraction of the expected counts due to WT sequencing error from any mutant population. This is especially helpful for obtaining accurate fitness measurements for deleterious alleles that differ from the WT allele by only 1 nt. Independent sample prep starting with PCR1 and separate barcoding in PCR2 is helpful for identifying sources of error during PCR and sequencing. Biological replicas starting from before the initial MAGE transformation are also helpful as fitness accuracy can be increased by averaging independent fitness measurements across replicas.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis as reported in the results and figure legends were calculated with custom scripts written in MATLAB. R-values indicate Pearson correlation coefficients.

DATA AND SOFTWARE AVAILABILITY

Software

MATLAB code for *E. coli* genome analysis is available upon request.

Data Resources

Fitness data for the *infA* single-codon, codon-pair and UTR mutants available as supplemental data files Data S1, Data S2, and Data S3.

Data on genome-wide RNA similarity scores in *E. coli* available as supplemental data file Data S4.

The accession number for the raw FASTQ files reported in this paper is ArrayExpress: E-MTAB-4020.