SUPPORTING INFORMATION

Evidence for Context-Dependent Complementarity of Non-Shine-Dalgarno Ribosome Binding Sites to *Escherichia coli* rRNA

Pamela A. Barendt[1], Najaf A. Shah[2], Gregory A. Barendt[3], Parth A. Kothari[1], and Casim A. Sarkar[1,2,4]

[1]Department of Bioengineering, [2]Genomics and Computational Biology Graduate Group, [3]Penn Medicine Academic Computing Services, [4]Department of Chemical & Biomolecular Engineering
University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

SUPPLEMENTARY TABLE LEGENDS

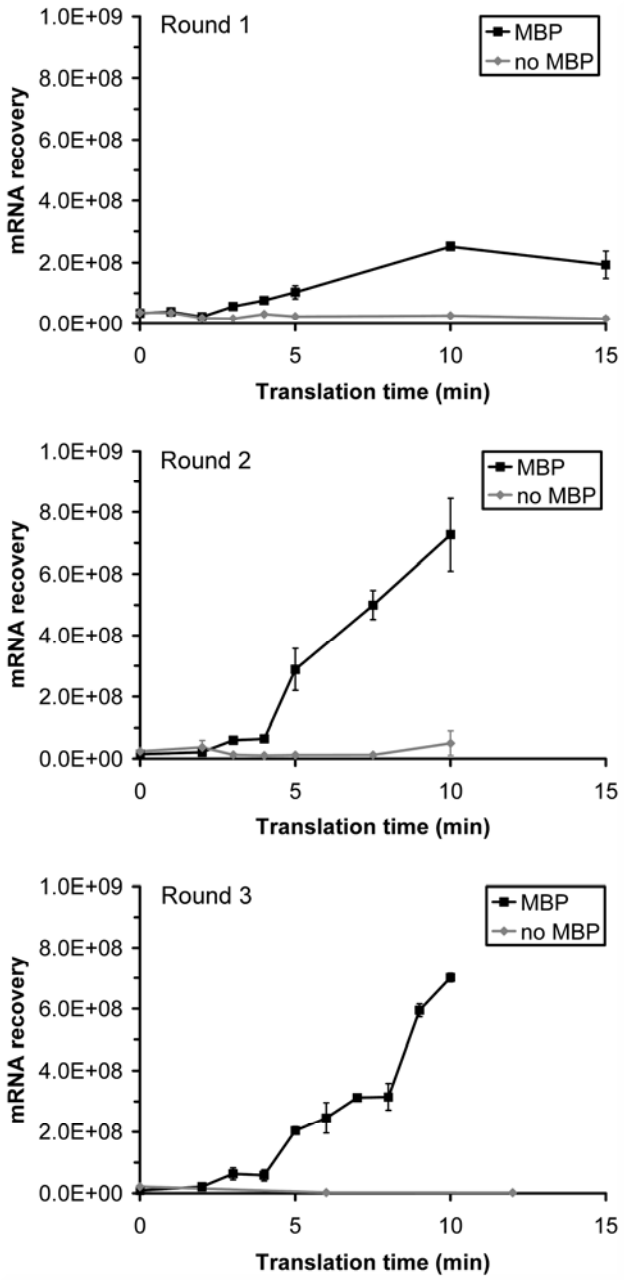
Supplementary Table 1. mRNA-rRNA complementarity. The first column in each group provides the index of the first 16S rRNA base in the "motif" column. The incidence of complementarity in the data (SD, non-SD, and *E. coli*) and the corresponding *p*-values (P.rand, based on random sequences as the null distribution; P.perm, based on permuted sequences as the null distribution) are presented. SD, Shine-Dalgarno.

Supplementary Table 2. Motif search results. The incidence of each motif in the data and the corresponding *q*-values (Q.rand, based on random sequences as the null distribution; Q.perm, based on permuted sequences as the null distribution) are presented.

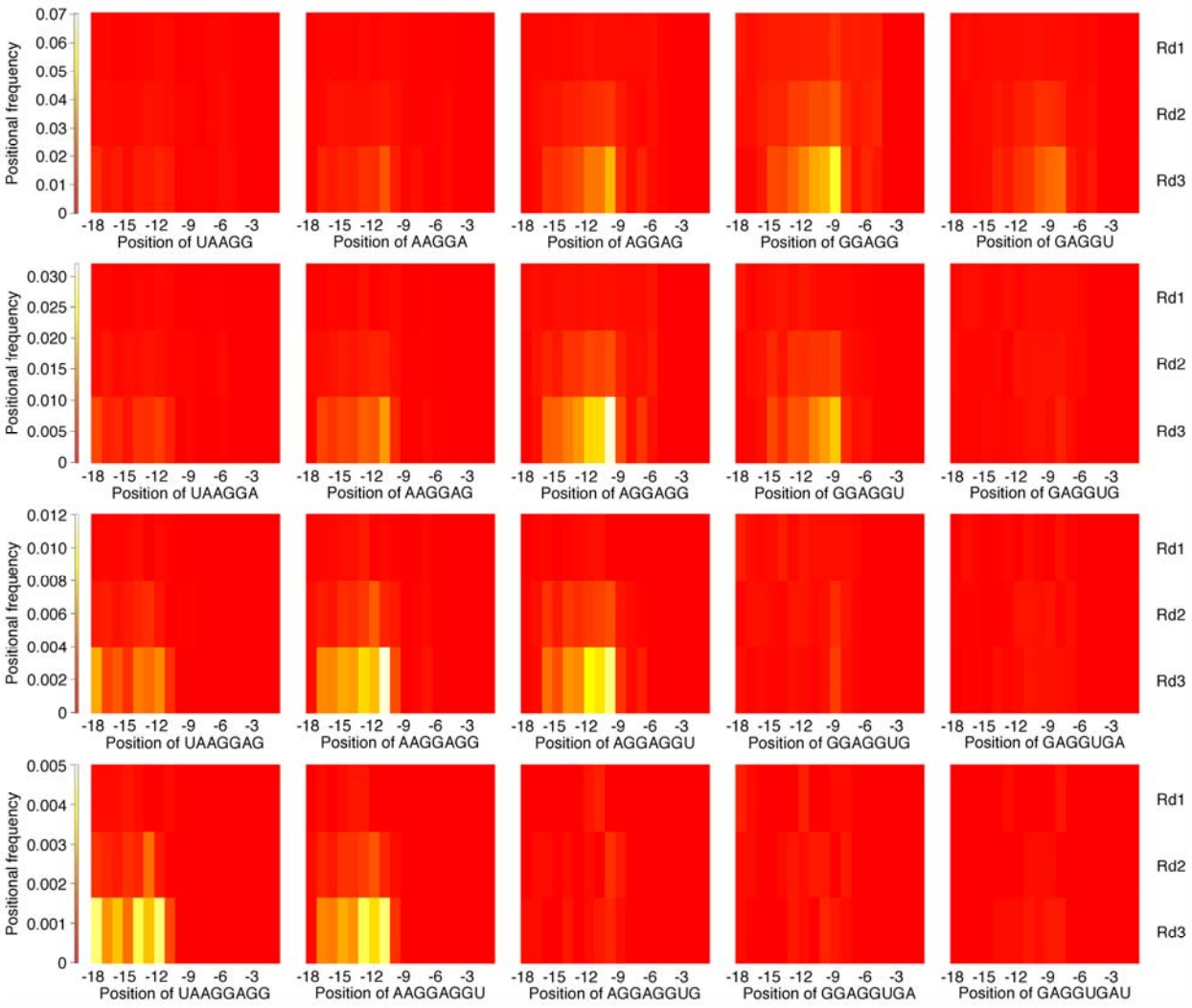
Supplementary Table 3. Co-occurrence of significant motifs. The number of sequences that contain both motif 1 and motif 2 is equal to "coincidence." Co-occurrence metric = coincidence/(motif 2 incidence).

Supplementary Table 4. Oligonucleotide sequences.

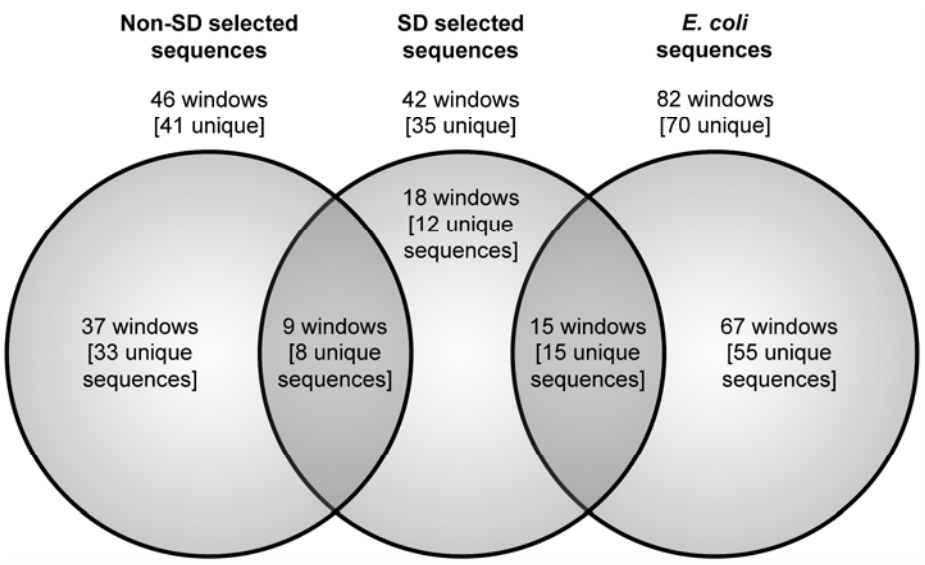
SUPPLEMENTARY FIGURES



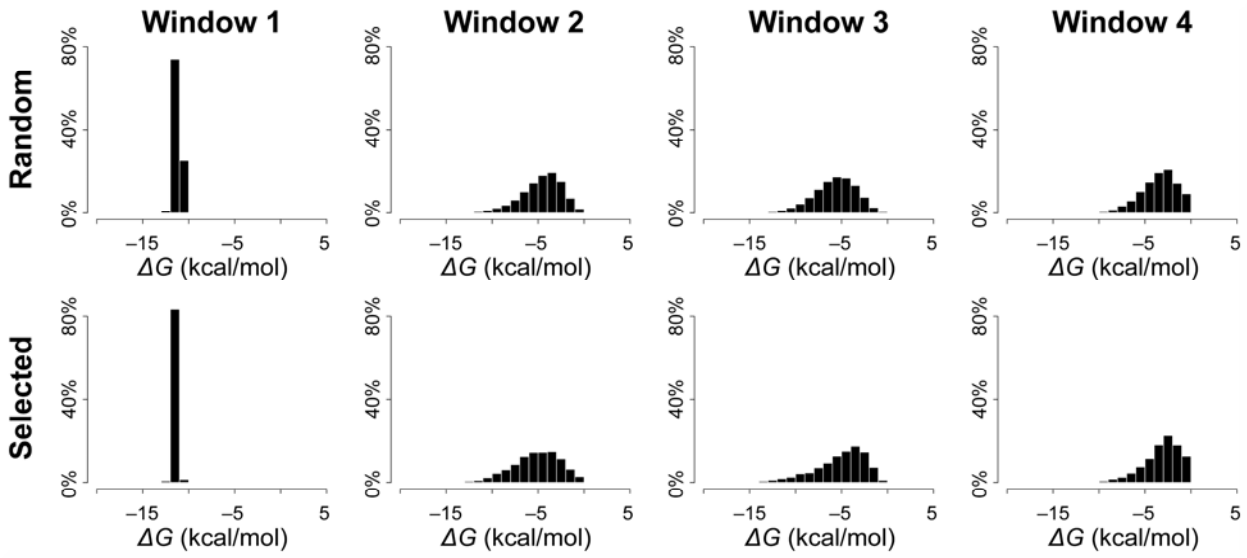
Supplementary Figure 1 mRNA recovery. mRNA recovery was quantified by qRT-PCR after each round. In each round, several different translation times were used. In round 1, 10 min translation produced the optimal signal. This mRNA was taken to round 2. From round 2, the 7.5 min translation product was taken forth to round 3. From round 3, the 7 min translation product was sequenced. Error bars indicate the half range of duplicate wells. MBP, maltose-binding protein; qRT-PCR, quantitative reverse transcription–polymerase chain reaction.



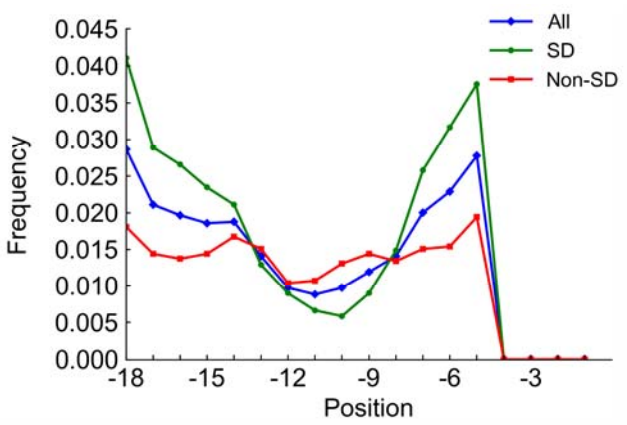
Supplementary Figure 2 Position-dependent enrichment of five-, six-, seven-, and eight-base sequences complementary to the 3' tail of the 16S rRNA over three rounds (Rd1, Rd2, Rd3) of selection. Position is specified by the first base of the motif relative to the start codon.



Supplementary Figure 3 Number of significant windows on the 16S rRNA complementary to selected sequences (SD/non-SD) and natural *E. coli* 5' UTRs. Notably, there was no intersection between significant windows complementary to selected non-SD 5' UTRs and significant windows complementary to *E. coli* 5' UTRs. 5' UTR, 5' untranslated region; SD, Shine-Dalgarno.



Supplementary Figure 4 Histograms of *ΔG* values. Histograms of *ΔG* values (in kcal/mol) in four 30-base sliding windows (offset by 10 bases) in a 60-base region centered on the 18-base randomized region in the theoretical naïve (top) and selected (bottom) library are shown. The similarity of the distributions suggests a lack of strong pressure for less or more secondary structure than a random library.



Supplementary Figure 5 Total frequency of top 10 five-base motifs most frequently co-occurring with five-base SD motifs vs. position. The top 10 five-base motifs most frequently co-occurring with five-base SD motifs (AAGGA, AGGAG, GGAGG, or GAGGU) are listed as “Motif 2” in Table 2 (UAUUA, AUUAA, UUAUA, UAUAU, GUUAA, GUAAU, GUAUA, UUGAA, UUAAG, and AUAUU). These motifs are over-represented in general (Supplementary Table 2) and are most likely found in positions upstream or downstream of the optimal five-base SD positions (which are near the middle of the 18-base randomized region) in selected library members containing a five-base SD motif (green line). In selected library members not containing a five-base SD motif, there is little dependence on position (red line). Position is specified by the first base of the motif relative to the start codon. SD, Shine-Dalgarno.

SUPPLEMENTARY METHODS

Materials. Oligonucleotides (Supplementary Table 4) were purchased from Integrated DNA Technologies. DNA purification was performed by agarose gel electrophoresis with SYBR Safe (Invitrogen) and QIAquick gel extraction kit (Qiagen) or by using QIAquick PCR purification kit (Qiagen). Restriction enzymes, T4 DNA ligase, and Phusion DNA polymerase were purchased from New England Biolabs.

Construction of RBS library. The process of library construction was designed to minimize bias. First, a primer-extension product was made using two Ultramer oligonucleotides: Short_leader_lib_fwd (5'-ATA CGA AAT TAA TAC GAC TCA CTA TAG GGA CAC CAC AAC GGT TTC CCN NNN NNN NNN NNN NNN NNA TGG CGG ACT ACA AAG ATG ACG ATG-3') and BsaI_FLAG_rev2_CGCCAT (5'-ACT GAT TAG GTC TCT CAT CGT CAT CTT TGT AGT CCG CCA T-3'). These were annealed and extended using the following program: 98°C for 1 min; 35 cycles of 69°C for 20 sec and 72°C for 4 sec; 72°C for 1 min; and 4°C thereafter. The product was purified using agarose gel electrophoresis. Oligonucleotides BsaI_FLAG_fwd3 (5'-ACT GAT TAG GTC TCC GAT GAC AAA GGA TCC GA-3') and tolAk (5'-CCG CAC ACC AGT AAG GTG TGC GGT TTC AGT TGC CGC TTT CTT TCT-3') (*1*) were used to amplify FLAG-off7-tolA from pRDVstops-off7 (*2*). The Short_leader_lib_fwd/BsaI_FLAG_rev2_CGCCAT primer-extension product (105 bp) and FLAG-off7-tolA (786 bp) were separately digested with BsaI to create compatible sticky ends (5'-CCAT-3' and 5'-ATGG-3,' respectively). The digests were purified and ligated at 16°C for 45 min. The full-length product (857 bp) was gel-purified according to Qiagen, except that RNase-free water was used for elution instead of EB buffer. This product was used directly for transcription, as previously described (*3*).

Reverse transcription. Reverse transcription was performed with AffinityScript reverse transcriptase (Agilent Technologies) and tolA_stops_HindIII_rev (5'-GGC CAC CAG ATC CAA GCT T-3'), which anneals just downstream of off7. The previously reported *in situ* reverse transcription protocol (*2*) was adapted as follows: 20.25 µL Solution 1 (19.4 µL water, 0.6 µL RNasin Plus [Promega], and 0.25 µL tolA_stops_HindIII_rev) was pipetted into the well, incubated at 65°C for 9 min, and removed from heat for 5 min. Then, 9.75 µL Solution 2 (3 µL dNTPs [5 mM each], 3 µL 10x AffinityScript buffer, 3 µL 0.1 M DTT, and 0.75 µL AffinityScript reverse transcriptase) was added, and the reaction was incubated at 45°C for 1 h, then heat-inactivated at 70°C for 15 min. The mRNA/complementary DNA (cDNA) was quantified by quantitative reverse transcription–polymerase chain reaction (qRT-PCR) on the Applied Biosystems 7300 Real-Time PCR System using TaqMan Universal PCR Master Mix (Applied Biosystems), off7_fwd (5'-TCC ATC GAC AAC GGT AAC GA-3'), tolA_stops_HindIII_rev, and off7_probe (6-FAM-5'-TGG CTG AAA TCC TG-3'). Based on the results from the first round, 10 µL of each reverse transcription reaction corresponding to 10 min translation (with maltose-binding protein on plate) was taken as template for a

100 µL PCR with primers T7_no_BsaI_TTTCC (5'-ATA CGA AAT TAA TAC GAC TCA CTA TAG GGA CAC CAC AAC GGT TTC C-3') and BsaI_FLAG_rev2_CGCCAT (5'-ACT GAT TAG GTC TCT CAT CGT CAT CTT TGT AGT CCG CCA T-3'). T7_no_BsaI_TTTCC anneals just upstream of the 18-base randomized region to maximize recovery. The 105-bp product was gel-purified. The off7-tolA segment was generated by amplifying pRDVstops-off7 with BsaI_FLAG_fwd3 (5'-ACT GAT TAG GTC TCC GAT GAC AAA GGA TCC GA-3') and tolAk (5'-CCG CAC ACC AGT AAG GTG TGC GGT TTC AGT TGC CGC TTT CTT TCT-3') (*1*) to form a 786-bp product. These two products were digested with BsaI and ligated, and the ligation reaction was gel-purified to generate the full-length construct (857 bp) for transcription.

Subsequent selection rounds. After the first selection round, two additional rounds were performed with multiple translation times each (Supplementary Figure 1). The 7.5 min sample from round 2 was taken to round 3. In rounds 2 and 3, the RNA:ribosome ratio was increased to ~4:1, and RT-PCR was performed with tolA_stops_HindIII_rev and T7_no_BsaI_TTTCC. Also, in round 2, an additional PCR was performed with T7_no_BsaI (5'-ATA CGA AAT TAA TAC GAC TCA CTA TAG GGA CAC CAC AAC GG-3') and BsaI_FLAG_rev2 (5'-ACT GAT TAG GTC TCT CAT CGT CAT CTT TGT AGT C-3') using the T7_no_BsaI_TTTCC/tolA_stops_HindIII_rev product as template, while FLAG-off7-tolA was amplified with BsaI_FLAG_fwd3 and tolAk. These products were digested with BsaI and ligated, and the ligation reaction was gel-purified to generate the template for transcription for round 3. Sequencing of the 10 min, 7.5 min, and 7 min samples from rounds 1, 2, and 3, respectively, was performed using a Roche/454 GS FLX sequencer at the University of Pennsylvania DNA Sequencing Facility.

Data analysis. In the present study, each window of length k (k = 4–8) on the *E. coli* 16S rRNA was compared to each k-base window in each sequence read of the randomized region, and the number of exact reverse-complements was recorded. To assess the statistical significance of rRNA-5' UTR association in the experimental library, this method was applied to 10^5 null libraries of random and permuted sequences followed by Bonferroni correction, yielding two sets of p-values, P.rand and P.perm, respectively (Supplementary Table 1). The random libraries consisted of uniformly random sequences, with each base equally likely at a given position. The permuted libraries were constructed by individually permuting each sequence in the experimental library; hence, P.perm assesses the significance of the rRNA-5'UTR association while controlling for the base composition of the sequences in the selected libraries, and thereby allows us to ascertain whether selection acts on base composition alone.

Significant six-base windows ($p < 0.01$) sharing five bases with at least one other significant window were considered part of a group of significant windows. PyMOL (*4*) was used to map these groups on the crystal structure (PDB accession code 3DF1; (*5*)). As in our previous study, no correlation between the position of these groups on the crystal structure and the position of the complementary motif within the randomized region was evident.

For the naïve motif search and co-occurrence analyses, virtual libraries (with random or permuted 18-base sequences, as described above) were generated and the incidence of each *k*-base motif was recorded. False discovery rate (FDR) was applied to correct for multiple tests, and the resulting *q*-values (Q.rand and Q.perm) are presented (Supplementary Table 2). Each significant *k*-base motif (FDR < 0.01; *k* = 4–6) was assessed to determine if it was more likely to co-occur in an 18-base sequence containing another significant motif than would be expected if the motifs occurred independently. A co-occurrence metric, [number of 18-base sequences containing non-overlapping motifs 1 and 2]/[number of 18-base sequences containing motif 2], was used to quantify this association. Non-zero co-occurrence metrics are reported (Table 2 and Supplementary Table 3)**.**

The analysis of mRNA secondary structure was adapted from previous work (*2*, *6*). Sequencing reads of the round 3 library were computationally trimmed to yield mRNA molecules consisting of the entire 21-base region immediately upstream of the randomized region, the 18-base randomized region immediately upstream of the start codon, and another 21-base region starting at the start codon. Each 60-base mRNA molecule was processed to yield four overlapping 30-base windows (offset = 10 bases). The secondary structure of each 30-base window was assessed using the UNAFold suite (*7*), and the corresponding *ΔG* values were noted. For comparison, a virtual library of 350,000 mRNA molecules with random 18-base regions (probability of each base = 0.25) was assessed for secondary structure using the same procedure.

Construction of clones for expression. pET-3a-newRBS-FLAG-off7-emerald GFP (emGFP) constructs were made by two consecutive PCRs. A clone-specific forward primer (i.e., 5'-CGA CTC ACT ATA GGG ACA CCA CAA CGG TTT CCC XXX XXX XXX XXX XXX XXX ATG GCG GAC TAC AAA GAT GAC-3' where the Xs correspond to a specific clone; Like_E._coli_1 through Poly-U_short_fwd in Supplementary Table 4) and BlpI_emGFP_short_rev (5'-TAG TTA TTG CTC AGC TTA CTT GTA CAG CT-3') were used to amplify newRBS-FLAG-off7-emGFP from one of the pET-3a-RBS-FLAG-off7-emGFP constructs from our previous work (*2*). The pET-3a-RBS-FLAG-off7-emGFP variant that we used as a template for this PCR contained the WT pRDV RBS in the longer leader context (*2*). The PCR product was directly used as template in a second PCR with primers BglII_5'_UTR_fwd (5'- ACT GAT TAA GAT CTC GAT CCC GCG AAA TTA ATA CGA CTC ACT ATA GGG ACA CCA CAA CGG-3') and BlpI_emGFP_short_rev. Poly-G was constructed by assembly PCR of two pieces: one made by PCR using mutually annealing primers BglII_5'_UTR_fwd and For_poly-G_short_rev (5'-GTC ATC TTT GTA GTC CGC CAT CCC CCC CCC CCC CCC CCC GGG AAA CCG TTG TGG TGT CCC TAT AGT GAG TCG-3') and a second piece made by PCR using For_poly-G_short_fwd (5'-ATG GCG GAC TAC AAA GAT GAC-3') and BlpI_emGFP_short_rev with the pET-3a-RBS-FLAG-off7-emGFP variant with the WT pRDV RBS in the longer leader context (*2*) as template. Final PCR products were purified using agarose gel electrophoresis and cloned into pET-3a (Novagen) between BglII and BlpI.

***In vivo* experiments.** Sequence-verified minipreps were used to transform *E. coli* BL21(DE3)pLysS (Agilent Technologies) to determine the expression level of emGFP. Individual colonies were inoculated into LB with 100 μg/mL ampicillin (to maintain pET-3a) and 50 μg/mL chloramphenicol (to maintain pLysS) and grown at 37°C for ~16 h. Ampicillin was omitted from the LB of the negative control (background strain). Cultures were then diluted 1:50 in 1 mL LB without antibiotic and grown at 37°C for 3 h, at which point half of each culture was induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). Cultures were grown at 37°C for an additional 4 h and subsequently analyzed on a Guava flow cytometer (Millipore). The average median fluorescence based on three separate experiments was used to determine whether expression was appreciable (i.e., greater than two-fold over background fluorescence of the strain).

REFERENCES

1. Binz, H. K., Amstutz, P., Kohl, A., Stumpp, M. T., Briand, C., Forrer, P., Grütter, M. G., and Plückthun, A. (2004) High-affinity binders selected from designed ankyrin repeat protein libraries, *Nat. Biotechnol. 22*, 575–582.
2. Barendt, P. A., Shah, N. A., Barendt, G. A., and Sarkar, C. A. (2012) Broad-specificity mRNA-rRNA complementarity in efficient protein translation, *PLoS Genet. 8*, e1002598.
3. Dreier, B., and Plückthun, A. (2011) Ribosome display: a technology for selecting and evolving proteins from large libraries, *Methods Mol. Biol. 687*, 283–306.
4. Delano, W. L. (2002) The PyMOL Molecular Graphics System.
5. Borovinskaya, M. A., Shoji, S., Fredrick, K., and Cate, J. H. D. (2008) Structural basis for hygromycin B inhibition of protein biosynthesis, *RNA 14*, 1590–1599.
6. 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.
7. Markham, N. R., and Zuker, M. (2008) UNAFold: software for nucleic acid folding and hybridization, *Methods Mol. Biol. 453*, 3–31.