ACS SyntheticBiology

Research Article

pubs.acs.org/synthbio

Avoidance of Truncated Proteins from Unintended Ribosome Binding Sites within Heterologous Protein Coding Sequences

Weston R. Whitaker,[†,§] Hanson Lee,[†,‡] Adam P. Arkin,[†,‡,∥] and John E. Dueber*[,†,‡,∥]

[†]Departments of Bioengineering, University of California, Berkeley, California 94720, United States

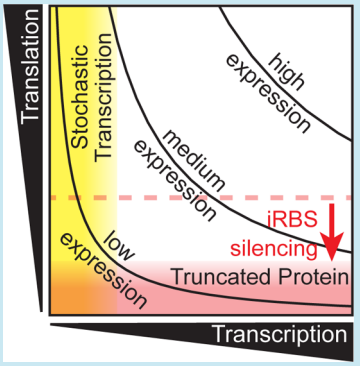
[‡]Energy Biosciences Institute, University of California, Berkeley, 2151 Berkeley Way, Berkeley California 94704, United States

[§]The University of California, Berkeley and University of California, San Francisco Graduate Program in Bioengineering, Berkeley, California 94720, United States

[∥]Physical Biosciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, United States

Supporting Information

ABSTRACT: Genetic sequences ported into non-native hosts for synthetic biology applications can gain unexpected properties. In this study, we explored sequences functioning as ribosome binding sites (RBSs) within protein coding DNA sequences (CDSs) that cause internal translation, resulting in truncated proteins. Genome-wide prediction of bacterial RBSs, based on biophysical calculations employed by the RBS calculator,[1] suggests a selection against internal RBSs within CDSs in *Escherichia coli*, but not those in *Saccharomyces cerevisiae*. Based on these calculations, silent mutations aimed at removing internal RBSs can effectively reduce truncation products from internal translation. However, a solution for complete elimination of internal translation initiation is not always feasible due to constraints of available coding sequences. Fluorescence assays and Western blot analysis showed that in genes with internal RBSs, increasing the strength of the intended upstream RBS had little influence on the internal translation strength. Another strategy to minimize truncated products from an internal RBS is to increase the relative strength of the upstream RBS with a concomitant reduction in promoter strength to achieve the same protein expression level. Unfortunately, lower transcription levels result in increased noise at the single cell level due to stochasticity in gene expression. At the low expression regimes desired for many synthetic biology applications, this problem becomes particularly pronounced. We found that balancing promoter strengths and upstream RBS strengths to intermediate levels can achieve the target protein concentration while avoiding both excessive noise and truncated protein.



KEYWORDS: *internal ribosome binding sites, truncated protein, gene optimization, protein expression, RBS calculator*

Coding DNA sequences (CDSs), besides encoding proteins, can have a number of embedded regulatory elements that may initiate transcription[2,3] or translation.[4−9] These initiation sequences are underrepresented in the CDSs of the organisms where they are recognized, presumably to avoid misregulation or wasted cellular resources.[10,11] CDSs taken from heterologous organisms or generated synthetically may manifest behavior that is difficult to predict or interpret when incorporated into a new organism, since they have not experienced selective pressure against problematic sequences. In this work, we focused on prokaryotic translation initiation sites encoded within CDSs, here termed internal ribosome binding sites (iRBSs, Figure 1a).

Alternative translation initiation sites are uncommon naturally but are occasionally utilized by both prokaryotes and eukaryotes despite their very different translation initiation mechanisms. Internal ribosome entry sites (IRESs), which drive cap-independent translation in eukaryotes, appear to be used primarily as regulatory control points[4] but are also used in some cases as a means of generating alternative isoforms of genes.[5] In prokaryotes, alternative translation initiation sites are less studied but likewise appear to generate alternative isoforms in natural systems for several documented cases.[6−9] Despite their natural roles, unintended iRBSs (also known as cryptic RBSs)[12,13] could be problematic for heterologous gene expression, resulting in truncated proteins. Expression of truncated protein products would represent, at a minimum, a waste of cellular energy and, at worse, a problematic, unexpected activity. For example, many eukaryotic signaling proteins have C-terminal catalytic output domains that are regulated by N-terminal input domains. A truncated protein may lack this regulation and show constitutive activity.[14] Similarly, truncated products of synthetic fusion proteins may produce erroneous output.[13] Recent work has suggested these sites may also lead to translational stalling, resulting in lower translation rates.[15] Thus, several potential complications may arise when expressing CDSs with iRBS sequences in prokaryotes.

Received: January 7, 2014

   dx.doi.org/10.1021/sb500003x | *ACS Synth. Biol.* XXXX, XXX, XXX−XXX

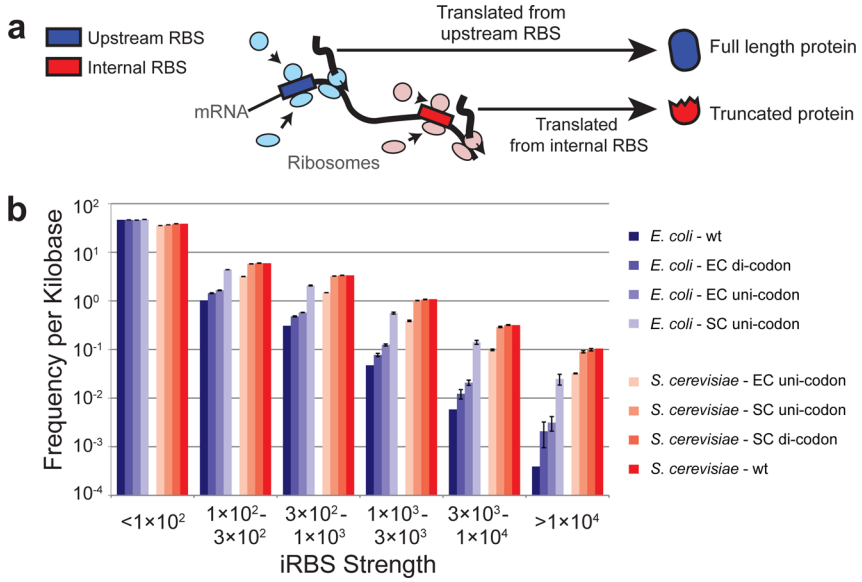


Figure 1. Frequency of internal ribosome binding sites (iRBS) within natural and recoded sequences. (a) Representation of how internal translation initiation from an in-frame iRBS produces truncated protein. The red rectangle denotes the iRBS and the blue rectangle denotes the intended 5′ upstream RBS. (b) Comparison of iRBS frequency for *E. coli* (blue) and *S. cerevisiae* (red) CDSs for natural and recoded genomes as predicted by the RBS calculator.[1] Every protein coding sequence in the *E. coli* and *S. cerevisiae* genomes was analyzed with the RBS calculator, excluding hypothetical or predicted genes. The codons for these protein coding sequences were recoded three independent times and analyzed with the RBS calculator. High strength iRBSs were rarely found in native *E. coli* CDSs (*E. coli*-wt, dark blue) compared to the recoded versions or all *S. cerevisiae* variants, suggesting a selective pressure against these sequences in *E. coli*. wt: native sequences. EC or SC unicodon: recoded CDSs with *E. coli* or *S. cerevisiae* codon usage frequency while preserving the amino acid sequences. EC or SC dicodon: recoded CDSs with *E. coli* or *S. cerevisiae* dicodon usage frequency according to Itzkovitz et al., while preserving amino acid sequences.

The Shine–Dalgarno sequence (SD) is thought to be the primary indicator of a strong translation initiating sequence, and sequence similarity to SD is commonly used as a way to gauge translational strengths. However, quantitative studies have found translational strengths to be highly dependent on the sequences adjacent to RBSs.[1,16] Salis et al. demonstrated improved ability to predict RBS strength by considering the free energy of both mRNA–rRNA hybridization and displacement of mRNA secondary structure.[1] The RBS calulator employs a statistical thermodynamic model of RNA hybridization and folding to predict the translation initiation rate. An optimization algorithm varies potential RBS sequence and shown to forward engineering RBS strengths spanning a range of 5 orders of magnitude with a 47% probability of achieving a strength estimate within 2-fold of the targeted level. Thus, the RBS calculator, based on biophysical calculations, represents a considerable, albeit not perfect, improvement in our ability to probe natural sequences for translation sites compared to using SD sequence similarity alone. Additionally, this tool provides a means of altering codon usage to increase mRNA secondary structure as a means of reducing iRBS strengths even if the SD sequence is found in highly constrained residues (e.g., lysine, AAG/AAA and glutamate, GAA/GAG). We estimated that 4.7% and 18% of the predicted iRBSs over 10^3 arbitrary units (au) and 10^4 au, respectively, in *S. cerevisiae* CDSs are constrained to the $(R)_6(n)_6ATG$ motif (i.e., the amino acid sequence (E/K)(E/K)xxM). If expressed in *E. coli*, the lengths of the resulting truncated proteins would average 51.9% ± 26.2% and 52.4% ± 26.9% (mean ± SD) of the full length proteins for predicted iRBSs over 10^3 au and 10^4 au, respectively. The constraints on coding for these amino acids would be expected to dictate the presence of a strong initiation site regardless of codon usage, unless silenced by the introduction of mRNA secondary structure to obscure this site from recruiting ribosomes. The authors of the RBS calculator have incorporated this function into the Operon Calculator[17] for gene optimization and removing potential iRBSs from CDSs. As synthetic biology applications increasingly call for low levels of protein expression,[18−22] iRBS removal will become more critical. This is particularly the case for expressing nonprokaryotic or synthetic sequences in prokaryotes, where the resultant iRBS strengths, by random chance, may be sufficient to produce a significant amount of truncated protein that is comparable, or even at a higher concentration, than the desired full-length product.

Here, we used the RBS calculator to compare the frequency of RBSs within prokaryotic and eukaryotic CDSs to estimate the prevalence of strong iRBSs. Our analysis showed that there is a considerably higher probability of finding in-frame iRBSs in eukaryotic CDSs than prokaryotic ones. Second, we demonstrated that iRBSs can result in truncated protein expression, which can be reduced or even eliminated with alternative codon usage designed to increase mRNA secondary structure as predicted by the RBS calculator. Third, we investigated whether interdependency exists between translational strengths of the iRBS and the upstream RBS. For example, a strong upstream RBS may be expected to load a sufficient number of ribosomes on the mRNA to occlude the iRBS site, obstructing *de novo* translation initiation. On the other hand, the presence of a high density of ribosomes may unfold mRNA secondary structure and expose iRBSs,[16] leading to more internal translation initiation. Our results indicated that the upstream RBS has little effect on internal translation from iRBSs. Finally, although the majority of iRBSs can be substantially weakened by silent mutations, some sequences, due to coding restraints or inaccurate mRNA secondary structure predictions, cannot be

dx.doi.org/10.1021/sb500003x | *ACS Synth. Biol.* XXXX, XXX, XXX−XXX

simply recoded to completely eliminate translation initiation. Since we found the strengths of iRBSs are independent of upstream RBS strengths, strengthening the upstream RBS can increase the percentage of full-length protein. For achieving a desired protein concentration, however, this necessitates lowering transcription levels, which results in increased noise,[23,24] particularly when targeting low protein concentrations. Our test case, expressing a model protein at a targeted low expression level, showed the best combination of minimal truncated protein and stochastic noise at a balance of moderate promoter and upstream RBS strengths.

RESULTS AND DISCUSSIONS

Internal Ribosome Binding Sites Are Selected against in Prokaryotic CDSs but not in Eukaryotic CDSs. How common are iRBSs when expressing nonprokaryotic CDSs in *E. coli*? To estimate probability, we compared the frequency and strength of iRBSs as predicted by the RBS calculator[1] within the genomes of *E. coli* (a representative prokaryote that utilizes RBS sequences) and *S. cerevisiae* (a representative eukaryote that does not utilize RBS sequences) (see Methods). It can be predicted that internal translation initiation sites are more frequent when expressing eukaryotic CDSs in a prokaryotic host because SD sequences (e.g., AGGAGG) appear less frequently in prokaryotic CDSs than in eukaryotic ones (even when codon usage and dicodon counts are taken into consideration[11]). However, biophysical calculations including the impact of mRNA secondary structure have not previously been used to estimate iRBS strengths. All CDSs, excluding hypothetical genes, were analyzed with the RBS calculator to predict the translational strength associated with each potential in-frame start codon. Internal start codons within 35 base pairs (bps) of the annotated translation start and stop codons were discarded because they may over-represent biologically relevant initiation sites and present difficulty for accurately representing the 35 bps of transcript found to be necessary for mRNA secondary structure calculations.[1] The *S. cerevisiae* CDSs were found to contain a substantially greater frequency of higher-strength iRBS predictions than *E. coli* CDSs, with more than a 200-fold increase in the likelihood of containing very high iRBS sites over 10^4 au (Figure 1b). The median of the strongest iRBS in a *S. cerevisiae* CDS is 802 au (i.e., 50% of the *S. cerevisiae* CDSs have at least one iRBS over 802 au) while that of an *E. coli* CDS is only 59 au. While iRBSs can occur anywhere in *S. cerevisiae* CDSs, there is a slight bias for iRBSs to appear more frequently toward the N-terminus of *S. cerevisiae* CDSs (linear regression gives $r = 0.3$, $p = 0.017$). Thus, there is a high probability that internal translation initiation would occur when expressing native eukaryotic gene sequences in prokaryotes.

We hypothesized the difference in the iRBS frequency between *E. coli* and *S. cerevisiae* is a result of a negative selection against iRBSs present only in prokaryotes, and not in eukaryotes. The availability of an internal ribosome binding site depends on local mRNA secondary structure and can be altered with different codon usage. Therefore, if iRBSs are selected against, we expect that changing the codons while preserving the amino acid sequences would increase the frequency of iRBSs in *E. coli* CDSs. On the contrary, the frequency would remain unchanged for *S. cerevisiae* CDSs. To test this, we recoded each CDS by randomizing each codon while maintaining the same amino acid sequences and the same codon usage frequency, or further maintaining the same dicodon usage frequency according to Itzkovitz et al.[11] These recoded CDSs are then scanned for iRBSs (Figure 1b). We found that the frequency of encountering iRBSs in recoded *E. coli* CDSs was the highest when preserving single codon usage frequency (EC unicodon), lower when preserving dicodon usage frequency (EC dicodon), and the lowest with the native sequence (wt). The more spatial relationship between nucleotides is retained, the less likely the recoding is to introduce strong iRBSs. In contrast, no changes in frequency were observed between the recoded and the native *S. cerevisiae* CDSs, supporting the hypothesis that iRBSs are selected against in prokaryotic CDSs but not subject to this selective pressure in eukaryotic CDSs.

Certain codons resembling the SD sequence, AGGAGG, occur rarely in *E. coli* CDSs but commonly in *S. cerevisiae* CDSs. For example, AGG and AGA appear in 2.4% and 2.7% of all arginine codons in *E. coli*, but they appear in 48.2% and 20.8% of all arginine codons in *S. cerevisiae*. If these amino acids happen to be upstream of a methionine residue, an iRBS is more likely to form with the *S. cerevisiae* codon usage frequency than with the *E. coli* codon usage, suggesting that recoding *S. cerevisiae* CDSs with *E. coli* codon usage frequency (*S. cerevisiae* − EC unicodon) should decrease the chance of encountering an iRBS in *S. cerevisiae* CDSs. Conversely, if the *E. coli* CDSs are recoded with *S. cerevisiae* codon usage frequency (*E. coli* − SC unicodon), the chance of encountering iRBS should increase. Both are found to be true in our analysis (Figure 1b). However, although recoding *S. cerevisiae* CDSs with *E. coli* codon usage frequency decreases the probability of iRBSs, the iRBS remains more than an order of magnitude higher than that in the native *E. coli* CDSs. Thus, applying host codon usage alone is predicted by the RBS calculator to reduce the frequency of iRBSs, but it is not sufficient to remove all these undesired activities completely or to the levels natively observed in *E. coli*.

We compared iRBSs in *S. cerevisiae* CDSs predicted by the RBS calculator to predictions based solely on SD sequences.[25] As expected, SD sequences are good at predicting the strongest iRBSs but poor at predicting moderate strength iRBSs. For very strong iRBSs (>10^4 au), SD sequences are present in 229 out of 236 iRBSs (97%); for iRBS > 10^3 au, the value drops to 72% (2746/3808); for iRBS > 10^2 au, only 42% of them contains SD sequences (12592/29980). It should also be noted that many nucleotide sequences contain SD sequences but are not predicted by the RBS calculator to be iRBSs: 62% of SD sequences (21243/33853) have iRBS strengths <100 au. This indicates that using SD sequence alone may not be sufficient to identify potential internal translation initiation sites in heterologous genes.

Taken together, our *in silico* analysis strongly suggests iRBSs are selected against in *E. coli* CDSs but not in *S. cerevisiae* CDSs. Due to the lack of a requirement for RBS function in translating eukaryotic mRNA and the resultant lack of a negative selection against their presence, naively transferring eukaryotic CDSs into bacteria could lead to production of truncated protein. Transferring CDSs between related prokaryotes would be expected to be less problematic. For applications expressing eukaryotic proteins in prokaryotes, recoding proteins with host codon usage reduces the likelihood of problematic iRBSs occurring, and this risk can be further reduced by employing biophysical calculation to specifically reduce iRBSs.

Internal Ribosome Binding Site Produces Truncated Protein. To empirically test if sequences predicted by the RBS calculator function as iRBSs, we investigated an often-used *E. coli* codon-optimized eukaryotic gene monomeric red fluo-

dx.doi.org/10.1021/sb500003x | ACS Synth. Biol. XXXX, XXX, XXX−XXX

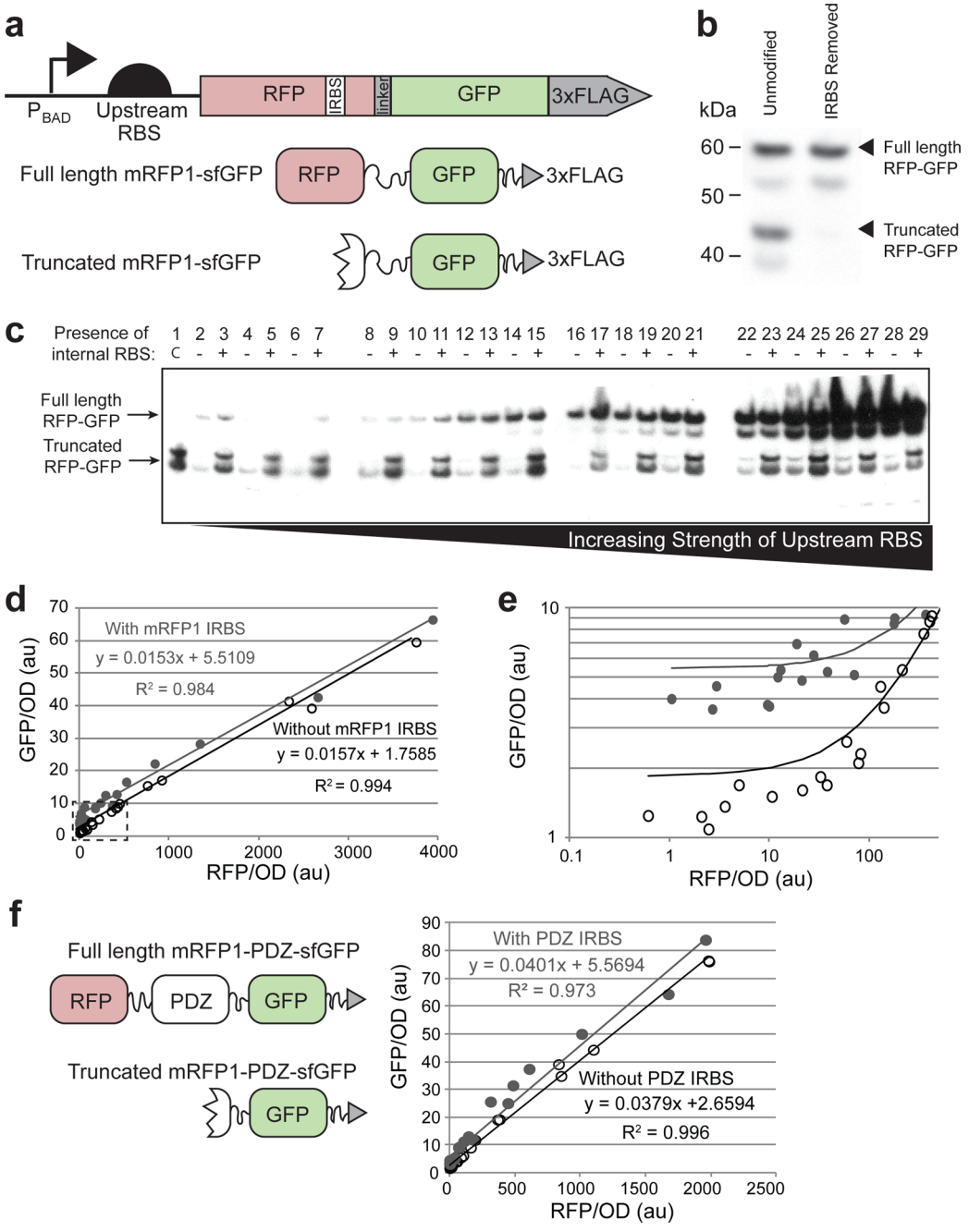


Figure 2. iRBS sequences can result in truncated proteins and are not inhibited by upstream translation. (a) Schematic of the mRFP1-sfGFP fusion construct with C-terminal 3xFLAG tag, the upstream RBS, and the P_{BAD} promoter. The mRFP1 sequence contains an iRBS with a calculated strength of 362 au initiated at MET163. The full length protein (57 kDa) produces both RFP and GFP fluorescence while the truncated product (39 kDa) only produces GFP fluorescence. (b) Western blot (using anti-FLAG antibodies) of the unmodified and iRBS-removed mRFP1-sfGFP construct expressed under the same 5′ upstream RBS. The iRBS was removed by introducing 11 silent mutations into the mRFP1-sfGFP construct with the aid of the RBS calculator. While full-length protein was present in both constructs, truncated product was only observed in the unmodified version at expected size. (c) A library was generated with varying strength upstream RBSs for constructs both with (+) or without (−) the iRBS. Western blot analysis showed bands corresponding to proteins with sizes expected from the full-length fusion protein and the truncated version if translation were to begin at the internal RBS site. Lane 1 is a control construct (C) containing a sequence beginning 50 base pairs upstream of the iRBS which is expected to produce only a truncated product. Lanes 2 through 29 correspond to increasing strengths of upstream RBS for full-length protein. Odd lanes correspond to constructs with unmodified mRFP1 sequences while even lanes correspond to constructs in which the iRBS strength has been reduced through silent mutations. (d) The GFP fluorescence from the members of the 5′ upstream RBS library is plotted as a function of their RFP fluorescence. The analysis of covariance (ANCOVA) showed that the linear fit (gray line) of the library members containing unmodified iRBS (gray filled circles) has the same slope but higher intercept compared to the linear fit (black line) of those with reduced iRBS strength (black circles). The increase in GFP due to the iRBS remained relatively constant as upstream RBS strength increased. (e) An expanded plot of the low upstream RBS region (dashed box in d) shows consistently higher GFP for constructs with iRBS. (f) A similar experiment was conducted for syntrophin PDZ domain, which is predicted to have two in-frame iRBSs. Syntrophin-PDZ domain was inserted between an iRBS-removed mRFP1 and a sfGFP (left). Again, unmodified PDZ domain constructs (right, gray filled circles) showed a constant higher GFP signal compared to those with iRBS removed (right, black circles) across the whole range of upstream RBSs tested.

dx.doi.org/10.1021/sb500003x | ACS Synth. Biol. XXXX, XXX, XXX−XXX

rescent protein (mRFP1)[26] as well as a natural mouse gene, syntrophin PDZ domain,[27] both of which are predicted by the RBS calculator to have in-frame iRBSs in their sequences. The RBS calculator predicts a single in-frame iRBS of 362 au that causes internal translation initiation from methionine 163 in mRFP1 (BBa_E1010, Registry of Standard Biological Parts), originally from mushroom coral (*Discosoma*), which should create a truncated protein of 63 amino acid long in addition to the full length version of 225 amino acids. No other in-frame site was found to be above 100 au. The sequence encoding the iRBS, GGTGAA, bears only some resemblance to the canonical SD sequence, AGGAGG, and is missed by algorithms that identify RBS sites solely by sequence identity. To test whether truncated protein resulted from the iRBS in mRFP1 as predicted by the RBS calculator, we fused a superfolding green fluorescent protein (sfGFP)[28] to the C-terminus of mRFP1 via a 12-residue glycine-serine linker (Figure. 2a). To enable quantitation of full and truncated protein product by Western blot and fluorescence analysis, a 3xFLAG epitope tag was added to the C-terminus of the fusion protein (Figure 2a). As a result, the full-length fusion protein is expected to be 57 kDa and emits both RFP and GFP fluorescence, while internal translation initiated at the iRBS within mRFP1 is expected to produce a truncated protein of 39 kDa that emits only GFP fluorescence. Indeed, we were able to detect the truncated protein with the predicted molecular weight by Western blot (Figure 2b).

If the truncated protein was the result of the iRBS, silent mutations predicted to lower the strength of the iRBS should reduce the amount of truncated protein. On the other hand, if the truncation was caused by proteolysis or another post-translational, amino acid sequence-dependent mechanism, silent mutations would not affect the amount of truncated protein. With the aid of the RBS calculator, 11 silent mutations were introduced into mRFP1 to reduce the predicted iRBS strength (from 362 au to 43 au). Only two of these mutations are within the canonical SD sequence, yielding GGCGAG and actually slightly increase the percent identity to the canonical AGGAGG. These sequence changes are designed to stabilize a mRNA secondary structure in which the iRBS is sequestered. As expected, the amount of full-length protein remained the same while the amount of truncated protein was reduced, as assayed by Western blot analysis (Figure 2b).

Next, we investigated the interdependency of iRBS strength on the strength of the 5′ upstream RBS by varying the nucleotide sequence upstream of the start codon. It is conceivable that stronger translation may lead to higher ribosome density on the mRNA strand, which in turn could mask the iRBS and thereby inhibit internal translation initiation. On the other hand, a high density of ribosomes may unfold mRNA secondary structure and expose iRBSs, potentially increasing translation.[16] A wide range of upstream RBS strengths were selected and verified to span more than a thousand-fold range of expression as measured with RFP fluorescence. Twenty-four constructs, varying only in upstream RBS sequences, were chosen containing either the unmodified or the iRBS-removed fusion protein-encoding gene, and analyzed by Western blot and fluorescent assays. The Western blot showed that the amount of truncated protein remained relatively constant while the amount of the full length protein increases with stronger 5′ upstream RBS strength (a subset of 14 different RBS strengths is shown in Figure 2c). We further quantified the results by measuring GFP and RFP fluorescence from each construct (Figure 2d and e). Since GFP fluorescence is produced by both truncated and full-length protein while, RFP fluorescence is only produced from the full length protein, we obtain the following equations:

$$\mathrm{GFP}_{\mathrm{total}} = \mathrm{GFP}_{\mathrm{truncated}} + \mathrm{GFP}_{\mathrm{full\ length}} \quad (1)$$

$$\mathrm{RFP}_{\mathrm{total}} = \mathrm{RFP}_{\mathrm{full\ length}} = \frac{1}{k}\mathrm{GFP}_{\mathrm{full\ length}} \quad (2)$$

where $\mathrm{GFP}_{\mathrm{total}}$ and $\mathrm{RFP}_{\mathrm{total}}$ is the total GFP or RFP fluorescence, respectively. $\mathrm{GFP}_{\mathrm{full\ length}}$ and $\mathrm{RFP}_{\mathrm{full\ length}}$ is the GFP and RFP fluorescence from the full-length protein, respectively. $\mathrm{GFP}_{\mathrm{truncated}}$ is the GFP fluorescence from the truncated protein, and k is a constant for converting RFP fluorescence per molecule to GFP fluorescence per molecule.

Rearranging the two equations, we get

$$\mathrm{GFP}_{\mathrm{total}} = k\cdot\mathrm{RFP}_{\mathrm{total}} + \mathrm{GFP}_{\mathrm{truncated}} \quad (3)$$

If the amount of truncated protein does not vary with the strength of the upstream RBS, the curve of GFP vs RFP fluorescence will be linear with the Y-intercept corresponding to the GFP fluorescence from the truncated protein. Moreover, the GFP vs RFP fluorescence curve of the unmodified mRFP1-sfGFP construct is expected to have the same slope but a larger Y-intercept than that with the iRBS-removed construct, corresponding to the additional GFP expressed from the truncated products. Indeed, we found the RFP and GFP regressed linearly with nearly identical slopes of 0.0153 and 0.0157 for the unmodified and iRBS-removed constructs, respectively (Figure 2d and e). The analysis of covariance (ANCOVA) showed that the slopes of the unmodified and iRBS-removed curves were not significantly different (p = 0.49). In contrast, the difference in the y-intercepts of 5.43 ± 0.51 and 1.85 ± 0.31 fluorescent units (mean ± SEM) for the unmodified and iRBS-removed constructs, respectively, were statistically significant ($p = 10^{-9}$). This suggested the fluorescence data were best explained by parallel lines of identical slopes offset by a constant expression level of truncated protein related by the y-intercept. Thus, the strength of the upstream RBS has no significant impact on translation initiation from the iRBS; that is, the iRBS contributes a fixed amount of truncated protein.

We repeated the fluorescence analysis with a eukaryotic protein domain, mouse syntrophin PDZ domain, that has been used for engineering synthetic assemblies in both *E. coli*[21,22] and *S. cerevisiae*.[29] The RBS calculator predicts two in-frame iRBSs in close proximity with the strengths of 2431 and 1345 au that can be altered with four silent mutations to reduce the predicted values to less than 10 au. This PDZ domain was inserted between the iRBS-removed mRFP1 and sfGFP for fluorescence measurement as described above. Again the removal of the iRBS decreased the GFP (y-intercept) by a relatively constant value (unmodified PDZ vs iRBS-removed PDZ = 5.95 ± 1.05 vs 2.31 ± 0.64 fluorescence units, respectively, $p = 6 \times 10^{-5}$, Figure 2f) while the difference in their slopes, 0.0401 vs 0.0379, was not statistically significant (p = 0.13). Together these results suggest sequences predicted to have iRBSs, even when not closely matching the canonical SD, can contribute truncated protein products. Further, these iRBSs act independently of the upstream RBS, consistent with recent ribosome profiling data where ribosome density is not high enough to cover the entire mRNA, even at a high translation level.[15]

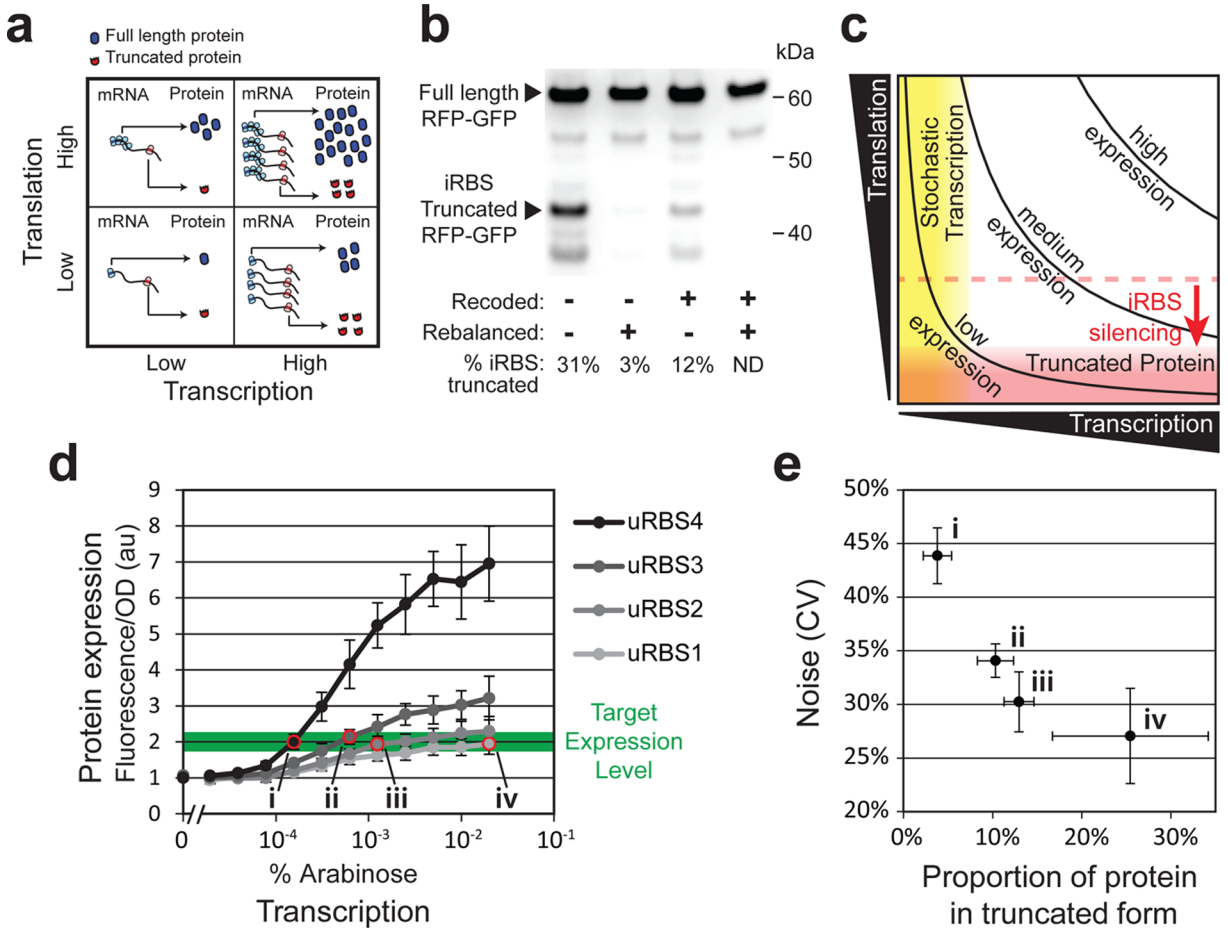


Figure 3. Balancing transcription and translation strength to adjust iRBS impact and protein expression noise. (a) Presence of an iRBS will result in internal translation initiation regardless of the upstream RBS strength, while the percentage of total protein that is truncated will depend on the strength of the 5′ upstream RBS. (b) Recoding proteins to silence iRBS sites and rebalancing expression with higher translation and lower transcriptions provide complementary approaches to reducing the impact of iRBS sites. A relatively weak uRBS with high induction of the RFP-GFP-3XFLAG construct (lane 1: pWW1923 induced with 0.01% arabinose) produces more truncated protein than expression with a stronger RBS and weaker transcription induction (lane 2: pWW1927 induced with 6×10^{-4}% arabinose). Silencing the iRBS with recoding, as done in Figure 2, reduces truncated protein for both the original and the rebalanced expression conditions as shown in lanes 3 and 4, respectively. (c) A schematic considering the regimes where expression noise and truncated protein production are problematic as translation and transcription levels are varied. When the translation rate from the upstream RBS is weak relative to an iRBS, the proportion of truncated protein produced dominates (light red area). Silencing an iRBS with alternative protein coding can be used to reduce the region in which truncated protein expression is problematic (red arrow). Expression from a weak promoter can result in stochastic transcription that increases cell-to-cell variability (light yellow area). Thus, transcription and translation rates must be balanced to avoid excessive noise from stochastic transcription or truncated protein from iRBSs, particularly at low expression levels where these problems are more difficult to avoid. (d) Expression of the mRFP1-sfGFP constructs was varied over a range of induction levels for several different RBS variants (ranging from weak, uRBS1, to strong, uRBS4, translation rates, see Table 1). A target expression level (highlighted in green) was chosen to compare noise and truncated protein proportion for different translation and transcription rates with similar protein expression (points i, ii, iii, and iv) as estimated by GFP fluorescence and confirmed by Western blot analysis (see Supporting Information Figure 2). (e) A trade-off between noise and truncated protein expression is observed at a target expression level (Figure 3d points i to iv). The proportion of truncated protein, determined by Western blot analysis (see Methods), increases as transcription rates increase (going from point i to iv). The coefficient of variation (CV), representing the cell-to-cell variability for each of the four samples, as measured by fluorescence microscopy (see Methods), decreases as translation rates decrease (going from point i to iv). Error bars represent the 95% confidence interval from three independent experiments (see Supporting Information Figure 2 for raw data).

Balancing Transcription and Translation Rates to Minimize Expression Noise and Protein Truncation when Low Protein Expression is Desired. Although using silent mutations to remove an iRBS is the fundamental way to reduce production of truncated protein, in some cases it may not be possible to completely remove iRBS activity due to amino acid sequence constraints. For example, all lysine (AAG/AAA) and glutamate (GAA/GAG) codons closely match the canonical SD sequence. We found that 18% (43/236 iRBSs) and 4.7% (179/3808 iRBSs) of iRBS over 10^4 and 10^3 au, respectively, in *S. cerevisiae* CDSs include the amino acid sequence motif (E/K)(E/K)xxM, which will exactly match the $(R)_6(n)_6$ATG motif regardless of codon usage choice. Additionally, miscalculations of difficult-to-predict mRNA structures (e.g., pseudoknots) near the iRBS may limit our ability to use predictions of the RBS calculator to silence the iRBS. For difficult-to-eliminate iRBS sequences, it may be best to use a strong 5′ upstream RBS to decrease the ratio of truncated to full-length protein since the iRBS contributes a fixed amount of truncated protein product independent of the upstream RBS

dx.doi.org/10.1021/sb500003x | *ACS Synth. Biol.* XXXX, XXX, XXX−XXX

Table 1. Sequences of the 5′ Upstream RBS Used in Figure 3[a]

	sequence	measured relative strength
uRBS1	^{45}GGTACCATTTAATAGGAGAATTTCTCGGCAGAGGGGAAT ^{1}ATG	16%
uRBS2	^{45}GGTACCTTTACAATGCCTAAGTTTAATTAGTAAAGAAGC ^{1}ATG	22%
uRBS3	^{45}GGTACCATATGCGCCCCTAACATCGGTCTTTAAAAAGGT ^{1}ATG	37%
uRBS4	^{45}GGTACCGGTATGAACAAACGATATTTATAATAAAGGAAT ^{1}ATG	100%

[a]The nucleotides are varied between a NheI site and a BglII site immediately upstream of the start codon. The relative strength of each RBS, measured by their RFP fluorescence, is listed on the right.

strength (Figure 2c and 3a), followed by a corresponding lowering of transcription (i.e., weaker promoter). This rebalancing of expression from a high transcription and low translation rate to a low transcription and high translation rate, while maintaining the same expression level, offers a complementary approach to reducing iRBS strength through recoding. This is shown in Figure 3b for the mRFP1-sfGFP construct from Figure 2c, where a moderate strength upstream RBS under high arabinose induction produces 31% truncated products. However, using a stronger upstream RBS and reducing arabinose induction reduces the truncated protein to 3% as estimated by Western blot analysis (see Methods). In combination, both recoding and rebalancing reduces truncated protein expression from this construct down to undetectable levels (Figure 3b).

Although strong upstream translation and weak transcription minimizes protein translation from iRBSs, stochastic gene expression (also referred to as noise) is known to result from weak transcription levels.[23,24,30] Therefore, as summarized in Figure 3c, a balance between transcription and translation rates may be required to minimize both the proportion of truncation products and cell-to-cell variability. Such a balance will be especially important when achieving low enzyme expression levels, a regime of increasing importance for many synthetic biology applications.[18−22] We developed a mathematic model to describe the trade-off between the cell-to-cell variability and the proportion of truncated proteins (see Supporting Information for details). In brief, for a given protein expression level the gene expression noise (η, expressed as the coefficient of variation CV) can be related to the proportion of truncated protein resulting from the iRBS (ϕ, which ranges between 0 and 1 as the relative amount of truncated proteins from the iRBS varies from zero to infinitely higher than the full length protein) with the following equation:

$$\eta \propto \sqrt{\left(\frac{k_{L,\mathrm{iRBS}}}{P}\right)\left(\frac{1}{\phi}-1\right)} \tag{4}$$

$k_{L,\mathrm{iRBS}}$ is the translation rate starting at the iRBS and P is the protein expression level.

When the translation rate of the upstream RBS is raised to reduce the proportion of truncated proteins (ϕ closer to 0), the gene expression noise increases due to the concomitant decrease in transcription rates necessary to maintain the desired protein concentration. Conversely, when ϕ increases due to lower upstream RBS strength, the transcription levels must be increased and, consequently, cell-to-cell variability goes down. Since the gene expression noise is inversely proportional to the root of protein expression level, low expression level (small P) would result in higher cell-to-cell variability and thus make the trade-off more prominent (Supporting Information Figure 1).

To investigate the trade-offs between the proportion of truncated protein and cell-to-cell variability, we vary the transcription and translation levels of the previously described mRFP1-sfGFP construct, while maintaining a set protein expression level. Several upstream RBS variants from above were chosen to span a wide expression range (uRBS1- uRBS4, with uRBS1 the weakest and uRBS4 the strongest, see Table 1 for sequences). These constructs were driven by the P_{BAD} promoter in the BW27783 strain with the transporter AraE integrated into the chromosome for constitutive expression, such that transcription rates can be varied homogeneously across the cell population with arabinose titration.[31] Each construct was integrated into the genome to eliminate the noise due to variability in plasmid copy number. Induction levels producing a similar amount of protein for the four different RBS strengths were chosen for analysis of cell-to-cell variability in gene expression and proportion of truncated protein (Figure 3d). The gene expression noise was determined by microscopy, as was performed in a previous study,[23] since the low signal was hard to measure confidently by flow cytometry. GFP was used instead of RFP fluorescence as GFP's higher quantum yield makes it easier to discern low intensity signal from the cell background. The single cell GFP concentration was automatically determined via a custom Matlab script (see Methods) and then fitted with a Gaussian curve to obtain the average intensity, the standard deviation, and CV. The normalized fluorescence distributions and their Gaussian fits are shown in Supporting Information Figure 2b. The proportion of truncated vs full-length proteins was determined by Western blot (see Supporting Information Figure 2a for gel images).

We found a trade-off between proportion of truncated protein and gene expression noise as predicted by the mathematical model (Figure 3e). Raising the transcriptional strength with a concomitant lowering of the upstream RBS strength, going from point i to point iv in Figure 3d and e, reduced gene expression noise, CV, from 44% to 27%, while increasing the proportion of truncated protein from 4% to 25%. At the moderate expression level illustrated here, either extreme of high noise or truncated protein expression can be avoided by balancing transcription and translation rates to intermediate levels (Figure 3d and e points ii and iii). This balance would be expected to become increasingly important at lower expression levels where further lowering of transcription or translation would exacerbate noise or truncation issues, respectively.

Summary. In conclusion, in-frame iRBSs can produce truncated protein when expressed in prokaryotes. Truncated products would be especially problematic for fusion proteins because it may cause partial functionality. As shown in our *in silico* analysis of predicted iRBS in *E. coli* and *S. cerevisiae* CDSs, the probability of encountering an iRBS is much higher when the CDSs are taken from eukaryotes—most likely because eukaryotic CDSs have not been subjected to negative selection against iRBSs. Similarly, gene synthesis based solely on host

dx.doi.org/10.1021/sb500003x | ACS Synth. Biol. XXXX, XXX, XXX−XXX

codon usage frequency will also run the risk of incorporating iRBSs. A signature of protein truncation due to iRBS instead of other post-translational events (e.g., proteolysis) is that the amount of truncated protein does not appear to be affected by the strength of 5′ upstream RBS, as shown by the two synthetic constructs we investigated. While RBS prediction software[1,17] can be helpful in approximating and reducing iRBS strength during gene optimization, the predictive capability is not currently universally precise. In addition, the coding constraints of some amino acids preclude complete elimination of iRBS function. Employment of a strong upstream RBS can reduce the proportion of the truncated product, albeit not a total elimination. However, when low protein expression is required, the upstream RBS strength needs to be balanced with an appropriate level of transcription such that both the percentage of internally translated product (due to low-strength upstream RBS) and the stochastic gene expression (due to weak promoters) can be minimized.

METHODS

Plasmids and Integration. Plasmids were constructed using a hybrid BglBrick-derived strategy[32] where the 5′ upstream RBS is placed between NheI and *Bgl*II sites, while the CDS is placed between *Bgl*II and XhoI sites. Strains with integrated constructs were based on the BW27783[31] strain and integrated into the galK locus as previously described,[23] using the Datsenko−Wanner method.[33] See Supporting Information Table 1 for a list of plasmids and strains used in this study.

Culture and Induction Conditions. Plasmids were transformed into chemically competent BW27783[31] cell strains using standard methods. LB/agar/antibiotic plates containing transformed strains were stored at 4 degrees for up to 5 days and inoculated into 300 μL of MOPS rich defined media (MRDM) (Teknova, Hollister, CA) with 0.4% glycerol as the carbon source and appropriate antibiotic in a 96 well plate for experimentation. This culture was grown for 12 h at 37 degrees in an ATR Multitron plate shaker at 1000 rpm, and then diluted 1:30 into 300 μL of prewarmed MRDM containing inducers for the experimental condition inducer concentrations. Strains were then similarly grown for 4 h until late log phase and immediately measured.

iRBS Strength Calculations. All RBS strength calculations were performed using the algorithm described by Salis and co-workers.[1] Code adapted for use with the Vienna RNA package[34] for improved portability was provided by Ying-Ja Chen and Christopher Voigt.

Genomic Shuffling, Recoding and iRBS Frequency Calculation. The CDSs of the entire *E. coli* str. K-12 substr. MG1655 and *S. cerevisiae* S288c genomes were downloaded from NCBI. Any CDSs annotated as "hypothetical" or "predicted" were excluded. All CDSs were shuffled three independent times as described by Itzkovitz and co-workers[11] to preserve bicodon pairs. Additionally all CDSs were recoded three independent times according to either *E. coli* or *S. cerevisiae* codon usage, preserving amino acid sequence but not bicodon frequencies. The RBS calculator was run on the entire CDS, and RBSs in all reading frames were considered except those within the first or last 35 base pairs, which were ignored. The RBS frequency is calculated by dividing the total amount of RBS by the combined length of all CDSs.

Fluorescent Assay. Measurement was done in a TECAN Safire2 machine with OD absorbance at 600 nM, GFP, and RFP excitation/emission of 481/507 nm and 584/607 nm, respectively, with a 5 nm bandwidth. Fluorescence readings were normalized by OD, though were generally within 2-fold between sample OD readings.

Western Blot. Western blots were prepared with protein gels and run under standard conditions for 10% Bis-Tris NuPAGE denaturing gels (Life Technologies), followed by transfer to nitrocellulose membrane, and then labeled with monocolonal ANTI-FLAG M2- peroxidase (HRP) antibody (Sigma) using standard procedure. Western blots were exposed on an ImageQuant LAS 4000 (GE Healthcare). Densitometry was performed using ImageJ analysis software (National Institutes of Health), and comparing the two prominent bands below 50 kDa to all bands to determine the percentage of truncated protein corresponding to the iRBS translation.

Microscopy and Image Analysis. Bacterial cells were fixed by 4% formaldehyde in PBS overnight at 4 °C, washed once with PBS, and resuspended in PBS before mounted on a slide. The images were taken under a Zeiss Axio Observer Microscope with 100× phase contrast objective. GFP images were taken with a green filter (emission, 470/40 nm; excitation, 525/50 nm). The cell boundaries were detected by their dark appearance using a custom Matlab script. Neighboring cells were further segmented using a watershed algorithm. The background intensity was measured by averaging fluorescence intensity across the region without cells. The concentration of GFP for each individual cell was calculated by summing the intensities of all its pixels, dividing by its area, and subtracting the background intensity. The distribution of single cell GFP concentration was then fitted with a single Gaussian curve to obtain the mean and the standard deviation.

ASSOCIATED CONTENT

Supporting Information

Derivation of the mathematical model, supplemental figures, and additional plasmid and strain information. This material is available free of charge *via* the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: 510-643-4616. Fax: 510-642-9725. Email: jdueber@berkeley.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Okoia Uket and Rami El-Kweifi from Prairie View A&M University and Aobo Wang from Zhejiang University, China, for assistance in experiments. We thank Ying-Ja Chen and Christopher Voigt for helping with the codes of RBS calculator, and members of the Dueber lab for discussions and comments during the preparation of this manuscript. This work is supported by Energy Biosciences Institute to Hanson Lee, by National Science Foundation (NSF) Synthetic Biology Engineering Research Center Grant EEC-0540879 and NSF Grant CBET-0756801 to Weston Whitaker and John Dueber.

REFERENCES

(1) 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*, 946−50.

(2) Kaplan, C. D., Laprade, L., and Winston, F. (2003) Transcription elongation factors repress transcription initiation from cryptic sites. *Science 301*, 1096−9.

(3) Cheung, V., Chua, G., Batada, N. N., Landry, C. R., Michnick, S. W., Hughes, T. R., and Winston, F. (2008) Chromatin- and transcription-related factors repress transcription from within coding regions throughout the *Saccharomyces cerevisiae* genome. *PLoS Biol. 6*, e277.
(4) Baird, S. D., Turcotte, M., Korneluk, R. G., and Holcik, M. (2006) Searching for IRES. *RNA 12*, 1755−1785.
(5) Ray, P. S., Grover, R., and Das, S. (2006) Two internal ribosome entry sites mediate the translation of p53 isoforms. *EMBO Rep. 7*, 404−10.
(6) Cot, S. S.-W., So, A. K.-C., and Espie, G. S. (2008) A multiprotein bicarbonate dehydration complex essential to carboxysome function in cyanobacteria. *J. Bacteriol. 190*, 936−45.
(7) Kofoid, E. C., and Parkinson, J. S. (1991) Tandem translation starts in the cheA locus of *Escherichia coli*. *J. Bacteriol. 173*, 2116−9.
(8) Matsumura, P., Silverman, M., and Simon, M. (1977) Synthesis of mot and che gene products of *Escherichia coli* programmed by hybrid ColE1 plasmids in minicells. *J. Bacteriol. 132*, 996−1002.
(9) Thomas, J.-C., Ughy, B., Lagoutte, B., and Ajlani, G. (2006) A second isoform of the ferredoxin:NADP oxidoreductase generated by an in-frame initiation of translation. *Proc. Natl. Acad. Sci. U.S.A. 103*, 18368−73.
(10) Hahn, M. W., Stajich, J. E., and Wray, G. A. (2003) The effects of selection against spurious transcription factor binding sites. *Mol. Biol. Evol. 20*, 901−6.
(11) Itzkovitz, S., Hodis, E., and Segal, E. (2010) Overlapping codes within protein-coding sequences. *Genome Res. 20*, 1582−9.
(12) Cabantous, S., Pédelacq, J.-D., Mark, B. L., Naranjo, C., Terwilliger, T. C., and Waldo, G. S. (2005) Recent advances in GFP folding reporter and split-GFP solubility reporter technologies. Application to improving the folding and solubility of recalcitrant proteins from *Mycobacterium tuberculosis*. *J. Struct. Funct. Genomics 6*, 113−9.
(13) Sachadyn, P., Stanisławska-Sachadyn, A., Kabat, E. M., Zielińska, A., and Kur, J. (2009) A cryptic ribosome binding site, false signals in reporter systems and avoidance of protein translation chaos. *J. Biotechnol. 143*, 169−72.
(14) Dueber, J. E., Yeh, B. J., Bhattacharyya, R. P., and Lim, W. A. (2004) Rewiring cell signaling: The logic and plasticity of eukaryotic protein circuitry. *Curr. Opin. Struct. Biol. 14*, 690−699.
(15) 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*, 538−41.
(16) Mutalik, V. K., Guimaraes, J. C., Cambray, G., Lam, C., Christoffersen, M. J., Mai, Q.-A., Tran, A. B., Paull, M., Keasling, J. D., Arkin, A. P., and Endy, D. (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods 10*, 354−60.
(17) Operon Calculator. https://salis.psu.edu/software/OperonCalculator_OptimizeCDSOnly (accessed May 9, 2013).
(18) Moon, T. S., Clarke, E. J., Groban, E. S., Tamsir, A., Clark, R. M., Eames, M., Kortemme, T., and Voigt, C. A. (2011) Construction of a genetic multiplexer to toggle between chemosensory pathways in *Escherichia coli*. *J. Mol. Biol. 406*, 215−27.
(19) Tabor, J. J., Levskaya, A., and Voigt, C. A. (2011) Multichromatic control of gene expression in *Escherichia coli*. *J. Mol. Biol. 405*, 315−24.
(20) Whitaker, W. R., Davis, S. A., Arkin, A. P., and Dueber, J. E. (2012) Engineering robust control of two-component system phosphotransfer using modular scaffolds. *Proc. Natl. Acad. Sci. U.S.A. 109*, 18090−5.
(21) Dueber, J. E., Wu, G. C., Malmirchegini, G. R., Moon, T. S., Petzold, C. J., Ullal, A. V., Prather, K. L. J., and Keasling, J. D. (2009) Synthetic protein scaffolds provide modular control over metabolic flux. *Nat. Biotechnol. 27*, 753−759.
(22) Moon, T. S., Dueber, J. E., Shiue, E., and Prather, K. L. J. (2010) Use of modular, synthetic scaffolds for improved production of glucaric acid in engineered *E. coli*. *Metab. Eng. 12*, 298−305.
(23) Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002) Stochastic gene expression in a single cell. *Science 297*, 1183−1186.
(24) Swain, P. S., Elowitz, M. B., and Siggia, E. D. (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. U.S.A. 99*, 12795−800.
(25) Chang, B., Halgamuge, S., and Tang, S.-L. (2006) Analysis of SD sequences in completed microbial genomes: Non-SD-led genes are as common as SD-led genes. *Gene 373*, 90−9.
(26) Campbell, R. E., Tour, O., Palmer, A. E., Steinbach, P. A., Baird, G. S., Zacharias, D. A., and Tsien, R. Y. (2002) A monomeric red fluorescent protein. *Proc. Natl. Acad. Sci. U.S.A. 99*, 7877−7882.
(27) Schultz, J., Hoffmüller, U., Krause, G., Ashurst, J., Macias, M. J., Schmieder, P., Schneider-Mergener, J., and Oschkinat, H. (1998) Specific interactions between the syntrophin PDZ domain and voltage-gated sodium channels. *Nat. Struct. Biol. 5*, 19−24.
(28) Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T. C., and Waldo, G. S. (2006) Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol. 24*, 79−88.
(29) Park, S.-H., Zarrinpar, A., and Lim, W. a. (2003) Rewiring MAP kinase pathways using alternative scaffold assembly mechanisms. *Science 299*, 1061−4.
(30) McAdams, H. H., and Arkin, A. (1997) Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. U.S.A. 94*, 814−819.
(31) Khlebnikov, A., Datsenko, K. A., Skaug, T., Wanner, B. L., and Keasling, J. D. (2001) Homogeneous expression of the PBAD promoter in *Escherichia coli* by constitutive expression of the low-affinity high-capacity AraE transporter. *Microbiology 147*, 3241−3247.
(32) Anderson, J. C., Dueber, J. E., Leguia, M., Wu, G. C., Goler, J. A., Arkin, A. P., and Keasling, J. D. (2010) BglBricks: A flexible standard for biological part assembly. *J. Biol. Eng. 4*, 1.
(33) Datsenko, K. A., and Wanner, B. L. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. U.S.A. 97*, 6640−5.
(34) Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., and Hofacker, I. L. (2008) The Vienna RNA websuite. *Nucleic Acids Res. 36*, W70−4.