Within-Gene Shine–Dalgarno Sequences Are Not Selected for Function

Adam J. Hockenberry,*[,1] Michael C. Jewett,[2,3,4,5,6] Luís A.N. Amaral,[2,7] and Claus O. Wilke[1]

[1]Department of Integrative Biology, The University of Texas at Austin, Austin, TX

[2]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL

[3]Chemistry of Life Processes Institute, Northwestern University, Evanston, IL

[4]Center for Synthetic Biology, Northwestern University, Evanston, IL

[5]Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL

[6]Simpson Querrey Institute, Northwestern University, Evanston, IL

[7]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL

***Corresponding author:** E-mail: adam.hockenberry@utexas.edu.

Associate editor: Deepa Agashe

Abstract

The Shine–Dalgarno (SD) sequence motif facilitates translation initiation and is frequently found upstream of bacterial start codons. However, thousands of instances of this motif occur throughout the middle of protein coding genes in a typical bacterial genome. Here, we use comparative evolutionary analysis to test whether SD sequences located within genes are functionally constrained. We measure the conservation of SD sequences across Enterobacteriales, and find that they are significantly less conserved than expected. Further, the strongest SD sequences are the least conserved whereas we find evidence of conservation for the weakest possible SD sequences given amino acid constraints. Our findings indicate that most SD sequences within genes are likely to be deleterious and removed via selection. To illustrate the origin of these deleterious costs, we show that ATG start codons are significantly depleted downstream of SD sequences within genes, highlighting the constraint that these sequences impose on the surrounding nucleotides to minimize the potential for erroneous translation initiation.

Key words: **translational control, sequence conservation, translation initiation, translational regulation.**

Introduction

The Shine–Dalgarno (SD) sequence is a short motif that facilitates translation initiation via direct base pairing with the anti-Shine–Dalgarno (aSD) sequence on the 16S ribosomal RNA (Shine and Dalgarno 1974). Several previous studies have shown that SD sequences are significantly depleted from *within* the protein coding genes of many bacterial species (Diwan and Agashe 2016; Umu et al. 2016; Yang et al. 2016). Although the depletion of SD sequences within protein coding genes is highly *statistically* significant, many prokaryotic genomes nevertheless contain tens of thousands of these sequences (Itzkovitz et al. 2010; Diwan and Agashe 2016; Yang et al. 2016). While SD sequences and their effect on translation initiation have been studied for decades (de Smit and van Duin 1990; Barrick et al. 1994; Chen et al. 1994), the role of these SD sequences within protein coding genes—hereafter referred to as SD-like sequences—is relatively unknown.

SD-like sequences may promote spurious internal translation initiation resulting in the production of truncated or frame-shifted protein products that are likely to be deleterious (Whitaker et al. 2015). These sequences are also known to promote ribosomal frame-shifting during translation elongation, which can have a beneficial regulatory function in specific cases (Larsen et al. 1994; Devaraj and Fredrick 2010; Chen et al. 2014). More recently, Li et al. (2012) have suggested a general role for SD-like sequences in regulating the local rate of translation elongation by directly binding to the translating ribosome and thus inhibiting ribosomal progression to the next codon. The appearance of SD-like sequence-induced translational pauses with ribosomal dwell times 10–100× baseline levels is supported by ribosome profiling studies in several bacterial species (Li et al. 2012; Liu et al. 2013; Subramaniam et al. 2013; Schrader et al. 2014) as well as experimental studies using a variety of techniques (Wen et al. 2008; Takahashi et al. 2012; Chen et al. 2014; Fluman et al. 2014; Vasquez et al. 2016; Frumkin et al. 2017). If SD-like sequences regulate elongation rates, many of the observed SD-like sequences within genes may actually be beneficial for cells; translational slowdown and pausing has been shown to facilitate proper protein folding in a number of different contexts (Saunders and Deane 2010; Ugrinov and Clark 2010; Spencer et al. 2012; Pechmann and Frydman 2013; Kim et al. 2015; Zhou et al. 2015; Chaney et al. 2017; Jacobs and Shakhnovich 2017; Sharma and O'Brien 2017).

However, other researchers have hypothesized that the experimental evidence for an association between SD-like sequences and translational pausing in ribosome profiling data may be an experimental artifact rather than a true

biological effect (Martens et al. 2015; Mohammad et al. 2016). Using a variety of different experimental techniques, other studies have failed to observe an association between the appearance of SD-like sequences and ribosomal pausing events (Elgamal et al. 2014; Borg and Ehrenberg 2015; Sohmen et al. 2015; Agashe et al. 2016; Chadani et al. 2016; Mohammad et al. 2016).

Taken together, the experimental evidence for whether SD-like sequences regulate translation elongation rates is mixed; these sequences may induce ribosomal pauses that dominate local elongation rate variation within genes, or they may not affect elongation rates at all. Further, if this mechanism of translational pausing is real, we still would not know whether organisms rely on the presence of SD-like sequences to regulate the rate of translation elongation. Just as plausibly, the cellular costs related to frame-shifting and spurious initiation may outweigh any benefits that would arise from employing this regulatory strategy. Determining the balance of these various effects is important for recombinant protein production applications that could use knowledge of SD-like sequences to tune elongation rates and encourage the production of properly folded proteins (Fluman et al. 2014; Vasquez et al. 2016).

Here, we apply comparative evolutionary analysis to determine whether SD-like sequences in the genome of *Escherichia coli* are deleterious, neutral, or beneficial. Evidence for conservation of these sequences would indicate that they are beneficial (Cooper et al. 2008; Kellis et al. 2014; Ashkenazy et al. 2016), perhaps due to a role in facilitating proper protein folding via modulation of local translation elongation rates. By contrast, our results show that 4-fold redundant nucleotides within SD-like sequences have significantly *higher* substitution rates than expected according to two different null model controls. These findings hold across a number of attempts to isolate a pool of functionally constrained sites (including analysis of highly expressed proteins and regions surrounding protein domain boundaries) and strongly suggest that SD-like sequences are weakly deleterious throughout the *E. coli* genome. We find that start codons are significantly depleted downstream of existing SD-like sequences, which provides evidence for the deleterious effects related to internal translation initiation that these sequences may promote. Our findings cast doubt on the role of SD-like sequences as a potential regulator of translation elongation rates in native genes, and urge caution when employing methods that use these sequences to tune translation elongation in recombinant designs.

Results

Assessing the Conservation Status of Shine–Dalgarno-Like Sequence Motifs within Protein Coding Genes

To investigate whether SD-like sequence motifs that occur within protein coding genes have a functional role, we searched for signatures of evolutionary conservation of these sites across related species. Under the hypothesis that some fraction of the SD-like sequence motifs that are present in any genome may be playing an important functional role, we would expect to observe significantly lower rates of nucleotide substitution within these sequence motifs relative to control sites. Conversely, if these sequences perform no such functional role and are instead generally deleterious to organismal fitness, we should observe significantly higher rates of substitution within these sequence motifs.

We assembled a data set of 1394 homologous protein families from 61 species in the order *Enterobacterales* and quantified nucleotide-level substitution rates across the coding sequences from this data set using LEISR (Kosakovsky Pond et al. 2005; Spielman and Kosakovsky Pond 2018). Each resulting substitution rate is a dimensionless number since we normalize the rate at each site within a gene by dividing by the mean rate across the gene (i.e., a substitution rate value of 2 an individual position refers to a site with twice the average substitution rate for that gene). As expected, we observed variability in substitution rates according to nucleotide identity within codons (1st, 2nd, 3rd positions) as well as across positions within genes (supplementary fig. S1, Supplementary Material online).

We used *E. coli* as a reference organism to identify the location of all SD-like sequence motifs that contain 4-fold redundant nucleotide sites in conserved amino acid positions (allowing up to one amino acid difference across all species for a given site). Additionally, we ignored all sites at the 5′ and 3′ gene ends to control for differences in substitution rates that occur at the end of genes to rule out the possibility of analyzing SD sequences of known function (see Materials and Methods). We note that canonical SD sequences are often not perfect complements to the highly conserved anti-SD sequence (Ma et al. 2002; Nakagawa et al. 2010; Hockenberry, Stern, et al. 2017), and in this article, unless specified otherwise, we used a binding energy threshold of -4.5 kcal/mol to define SD-like sequences. According to this threshold, 1,998 out of 4,127 *E. coli* protein coding genes are preceded by SD sequences, significantly more than expected by chance (Expectation: 638.57, z-test: $P < 10^{-16}$). By the same definition, all *E. coli* protein coding genes contain 25,001 SD-like sequences, significantly fewer than expected by chance alone (Expectation: 30,397.57, z-test: $P < 10^{-16}$) but far more than the number of known SD sequences that function in translation initiation (supplementary table S1, Supplementary Material online).

We adopted a paired-control strategy to compare substitution rates between nucleotide sites that fall within SD-like sequence motifs to control sites selected from the same gene that *do not* occur within SD-like sequences. Throughout the remainder of this article, we use the nomenclature of "codon" and "context" controls to refer to two different methods for selecting control nucleotides. In codon controls, after identifying a 4-fold redundant codon *within* a SD-like sequence, we find another occurrence of the same codon within the same gene to use as a control. Similarly, in context controls, we find the same trinucleotide site (at the -1, 0, and $+1$ positions, where a 4-fold redundant position is at position 0) within the same gene to use as a paired control (fig. 1A). The codon control is necessary because altered synonymous codon preferences across species could affect the apparent selection on

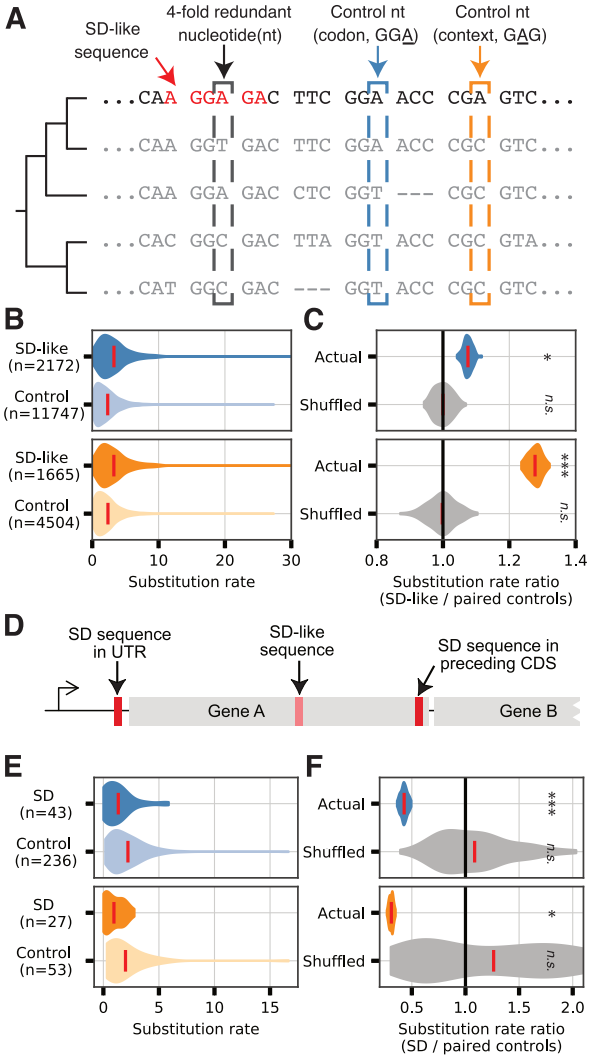

Fig. 1. SD-like sequences have elevated rates of nucleotide substitution. (*A*) Graphical illustration of methodology for identifying 4-fold redundant sites within SD-like and control sites. (*B*) Relative substitution rates for all SD-like sites and control sites. Top (blue) and bottom (orange) panels depict results for codon and context controls, respectively. Red lines in violin plots depict category means. (*C*) The ratio of the average substitution rates between SD-like and control categories based on a gene-specific bootstrapped approach discussed in main text ($P = 0.003$, 0.48, 1.2×10^{-14}, 0.53, top to bottom). (*D*) A diagram of genes within an operon illustrating a SD sequence, a SD-like sequence, and a SD sequence that occurs within the 3′ end of a preceding gene. (*E*) As in (*B*), showing scores for putative SD sites within the 3′ end of genes. (*F*) Substitution rate ratios for putative SD sites depicted in (*E*) ($P = 0.001$, 0.48, 0.013, 0.51, top to bottom). (* denotes $P < 0.05$, *** denotes $P < 0.001$).

SD-like sequences. Additionally, the context control is necessary because the mutation rate and spectrum at individual nucleotides is partially governed by the identity of flanking nucleotides. Controlling for these identities ensures that any differences, we see in SD-like sequence substitution rates are not simply due to local sequence effects that either increase or decrease the mutation rate.

Since SD-like sequences are relatively rare (~0–5 sites for the majority of genes), there are frequently many possible control sites within a given gene for each synonymous nucleotide that occurs within a SD-like sequence (fig. 1*B*). We thus randomly sampled single control nucleotide sites (from within the same gene) for each applicable SD-like nucleotide. From the resulting paired list of substitution rates for SD-like and control sites, we calculated the ratio of the average substitution rates between the two categories (SD-like sites divided by control sites) and repeated this sampling procedure 100 times to estimate the overall effect size. Assuming no difference in substitution rates between SD-like and control sequence categories, the ratios should follow a normal distribution centered around a value of 1. If sites within SD-like sequences are more conserved than control sites, we should observe values significantly <1. Finally, if sites within SD-like sequences have elevated substitution rates, indicating that they are generally deleterious, we expect to observe ratios significantly >1. To give these substitution rate ratios some context, we used a similar gene-specific paired-strategy to compare differences in substitution rates between all 4- and 2-fold redundant nucleotide sites (agnostic of their appearance in or outside of SD-like sequences). The resulting ratio

Table 1. Number of Sites Analyzed in Each Bootstrap Replicate Used to Calculate Substitution Rate Ratios.

Data Set	Codon Control	Context Control
Full data set	**1,740**	**1,161**
Putative SD sites	**31**	**15**
Locally strong	**1,137**	**734**
Locally weak	**298**	**234**
Protein abundance, locally strong		
0–20%	**139**	**79**
20–40%	**257**	**160**
40–60%	**297**	**193**
60–80%	**255**	**175**
80–100%	**187**	**127**
Postdomain, locally strong	**69**	**43**
Postdomain, locally weak	**24**	**21**

NOTE.—This number differs from the total number of all SD-like sites identified for a given criteria as some SD-like sites lack suitable control sites and are discarded from further analysis. When the number of SD-like sites in a gene exceeds the number of control sites for a given criteria, pairs of sites are randomly sampled without replacement until no control sites remain for a given bootstrap replicate. Abundance data are from Wang et al. (2015) and domain boundary information from Ciryam et al. (2013).

value of 2.02 shows that 4-fold redundant sites have, on an average, twice the substitution rate of 2-fold redundant sites–a value that is in line with our expectations based on the redundancy of the genetic code.

Regardless of which null model strategy that we used to select control nucleotides, we found that the substitution rates of SD-like sequences are *higher* than that of control sequences with an effect size on the order of ~10–30% (fig. 1C). By contrast, as a negative control to ensure the statistical validity of our methods, we randomly assigned nucleotide sites to SD-like or control categories. We confirmed that our methods gave the proper null result using this shuffled version of our actual data points: substitution rate ratios were centered around the expected null value of 1 ("shuffled" data in fig. 1C). We conservatively determined statistical significance by calculating the Wilcoxon signed-rank test between SD-like and control categories for each bootstrap replicate and report the median P value ($P = 0.003$ and $P = 1.2 \times 10^{-14}$ for codon and context controls, respectively). We note that our statistical procedure for selecting SD-like sequences with paired gene-specific controls limits the overall number of SD-like sites that we can analyze during any individual bootstrap replicate. Table 1 ("Full data set" row) highlights the numbers of sites analyzed per replicate for the analysis presented in figure 1C. These sites, however, are not evenly spread across all 1394 genes in our data set and the median number of analyzed sites per gene was 0 for both controls—reflecting the relative rarity of SD-like sites with suitable controls and preventing us from making any gene-specific claims. A total of 494 and 432 genes contributed to the analysis in figure 1C with no gene contributing >1.6% of these sites (supplementary fig. S2, Supplementary Material online).

The finding of elevated substitution rates within SD-like sequences remains largely unchanged when we used different thresholds to define these sequences (supplementary fig. S3, Supplementary Material online), as well as different organisms (*S. enterica*, *K. pneumoniae*, and *Y. pestis*) to identify the locations of SD-like sequences (supplementary fig. S4, Supplementary Material online). For *Y. pestis*, the most diverged of all organisms that we considered, we fail to observe a significant difference from random for the codon control but all other analyses in all other organisms remain significant.

To ensure that our methodology was capable of predicting conservation of sequence motifs that are *known* to be functionally constrained, we leveraged the fact that some genes in our data set are directly followed by another gene in the 3′ direction. Thus, the true SD sites of certain downstream genes are expected to occur within the 3′ coding sequence of upstream genes (fig. 1D). We therefore repeated our analysis by only considering putatively true SD sites, which we define as sites that occur within the −50 to −1 region (relative to the stop codon) in the subset of genes where a downstream start codon lies within 20 nucleotides (+/−) of the stop codon of interest (while still selecting control sites from the internal regions of the gene). Despite the low number of motifs that met this criterion, 4-fold redundant sites within this restricted set of putative SD sequences had a substitution rate that is roughly 1/3 that of control nucleotides, indicating strong evolutionary conservation of these known SD sites and validating our overall statistical approach (fig. 1E and F). We ensured that this result was not simply an artifact of differential substitution rates at the 3′ end of genes by conducting the same analysis on sites that occur within the 3′ region of genes that *do not* have any annotated genes directly following, and thus are not expected to function as true SD sites. We detected no significant signal of evolutionary conservation for this set of sites (supplementary fig. S5, Supplementary Material online).

Substitution Rates Differ According to Mutational Effects on SD-Like Sequence Strength

In the preceding section, we showed that 4-fold redundant sites within SD-like sequences have significantly higher substitution rates than control sites. This finding provides support for the model of SD-like sequences being deleterious and evolutionarily transient within genes. However, the SD sequence binds facilitates translation initiation by binding directly to the anti-SD(aSD) sequence on the 30S ribosomal subunit, and this binding strength spans a range of values according to the actual SD nucleotide sequence in question. We thus separately investigated SD-like sites according to how many synonymous mutations to the 4-fold redundant nucleotide in question were predicted to increase the strength of binding to the aSD sequence (see fig. 2A for an example). Note that this designation does not refer to the absolute strength of aSD sequence binding, but rather the capacity for *strictly* synonymous mutations to the site in question to either increase or decrease the relative aSD sequence binding. We refer to the "locally strong" and "locally weak" sites hereafter as those where any synonymous mutation is guaranteed to decrease or increase, respectively, the strength of aSD sequence binding.

Based on our previous results, we hypothesized that if SD-like sites are deleterious, we should observe conservation of

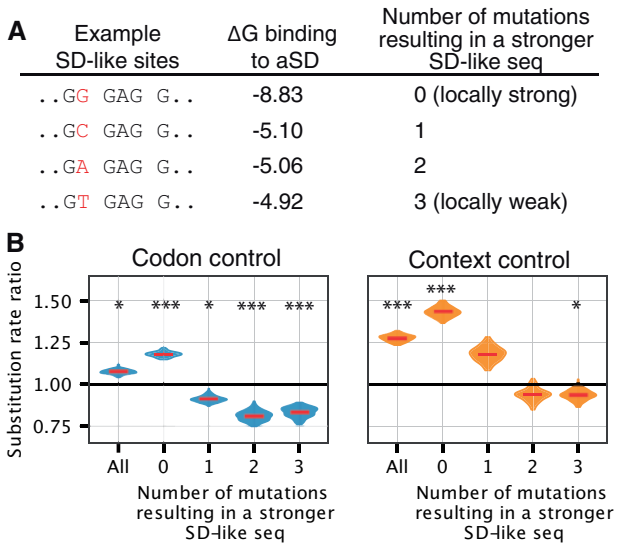

FIG. 2. Local mutational effects on SD strength alter substitution rate patterns. (*A*) Synonymous mutations to SD-like sequences may either increase or decrease SD-like sequence strength depending on the identity of the 4-fold redundant nucleotide. (*B*) Substitution rate ratio results as in figure 1C. Data shown here by stratifying "all" sites into categories that correspond to the expected change in SD strength given a synonymous substitution. Results shown for synonymous codon (left, $P = 0.003$, 2.7×10^{-12}, 0.006, 0.0004, 4.2×10^{-5}) and nucleotide context (right, $P = 1.2 \times 10^{-14}$, 1.3×10^{-22}, 0.17, 0.2, 0.026) controls. (* denotes $P < 0.05$, *** denotes $P < 0.001$).

locally weak sites. For this subset of sites, any synonymous mutation would, by definition, result in an *increased* aSD sequence binding strength. Indeed, substitution rates for this category of sites were significantly lower than expected (substitution rate ratios <1, $P < 0.01$), regardless of our method for selecting control nucleotides (fig. 2*B*). By contrast, when we analyzed the subset of locally strong SD-like sites, where any mutation to the 4-fold redundant position is guaranteed to result in a *weaker* interaction with the aSD sequence, we observed the opposite effect. These sites—which are the majority of identified SD-like sites—had substantially elevated substitution rates compared with paired controls on the order of ~10–40% (see table 1 for the number of data points included in each category, which are highly skewed toward locally strong sites in this analysis).

We stress that these findings are not indicative of conservation of intermediate or weak SD-like sites, but rather the *weakest possible* sites given the amino acid constraints of the sequence. To further address this point, we performed the same analysis on weak SD-like sites, which we define as having aSD sequence binding free energy values between −3.5 and −4.5 kcal/mol. We observed the same pattern of locally strong sites having significantly elevated substitution rates; this is despite the fact that these sites are weaker in absolute terms than all sites depicted in figure 2 (supplementary fig. S6, Supplementary Material online). This nucleotide dependent analysis shows that the magnitude of negative selection acting against SD-like sites is stronger than we initially observed in figure 1. As before, to ensure the robustness of these results, we we used different thresholds to define SD-like sequences (supplementary fig. S7, Supplementary Material online), as well as a different organism (*Y. pestis*) to identify the locations of SD-like sequences and their classifications (supplementary fig. S8, Supplementary Material online). When we used *Y. pestis* to identify all SD-like sites in aggregate, we failed to observe a significant difference in substitution rates for the codon control case (supplementary fig. S4, Supplementary Material online). By contrast, considering locally strong and locally weak mutations restores statistical significance even in this case, highlighting the importance of considering the possible effect of mutations on the interaction between SD-like sequences and the aSD sequence.

Consistent Results across Protein Abundance Bins

While we have thus far shown that SD-like sequences as a whole are less conserved than expected, this does not preclude the possibility that some fraction of SD-like sequences have a functional role and are evolutionary constrained. The SD-like sequences that we have analyzed may actually be a mixture of deleterious and functionally beneficial sites that look weakly deleterious in aggregate. We reasoned that the most highly abundant proteins are most likely to have been purged of deleterious SD-like sequences leaving the SD-like sequences that remain within these genes particularly attractive candidates for functional conservation. Thus, if SD-like sequences are a mixture of effects, we expect to find SD-like sites within highly expressed genes to be *relatively* more conserved than other categories. By contrast, if SD-like sites are a uniform pool in terms of their overall negative effects, we predict that the substitution rates between different gene expression categories will not systematically vary. To test this hypothesis, we separated our data set into quintiles of genes according to their overall protein abundances in *E. coli*, and analyzed the substitution rate ratios of SD-like and control categories as before.

We confirmed that the most highly abundant proteins contain fewer SD-like sequences (fig. 3*A*). Since the the fraction of conserved amino acids per gene varies according to bins of protein abundance (fig. 3*B*), the overall fraction of SD-like sites eligible for analysis is variable between different protein abundance bins (fig. 3*C*). However, we nevertheless observed largely consistent results across all protein abundance bins: locally strong 4-fold redundant nucleotides within SD-like sequences have significantly higher substitution rates than paired controls (fig. 3). These results remained robust to our assumptions with regard to SD-like thresholds (supplementary fig. S9, Supplementary Material online) and species used to identify SD-like sites (supplementary fig. S10, Supplementary Material online)—though we note in the latter case *E. coli* values were still used to classify homologs into protein abundance bins. To ensure that these findings were not affected by gene length differences across bins, we removed the shortest and longest 10% of genes from our data set and observed similar results (supplementary fig. S11, Supplementary Material online). Finally, we also found that the locally weak SD-like sites had significantly lower

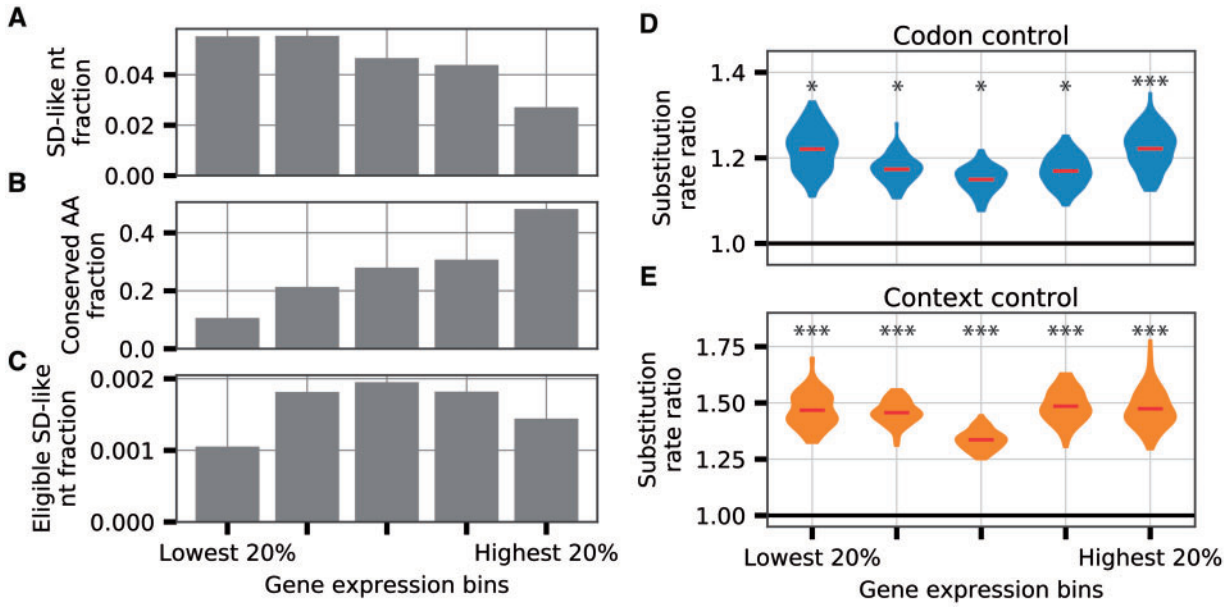

FIG. 3. SD-like sequences have similarly elevated substitution rates across protein abundance bins. (*A*) The most highly abundant proteins contain fewer SD-like sequences. (*B*) Highly abundant proteins are have a higher fraction of conserved amino acids. (*C*) Combined, the effects from (*A*) and (*B*) affect the fraction of SD-like sites within genes that are eligible for our analysis. (*D*) Substitution rate ratios of the locally strong SD-like sequences are elevated across all levels of protein abundance compared with synonymous codon controls ($P = 0.009$, 0.003, 0.004, 0.001, 0.0005). (*E*) As in (*D*), shown according to context controls($P = 0.0001$, 2.3×10^{-6}, 1.9×10^{-5}, 8.4×10^{-7}, 4.0×10^{-5}). (* denotes $P < 0.05$, *** denotes $P < 0.001$).

substitution rates than expected across nearly all protein abundance bins with the only exceptions being for the sites within the very lowest protein abundance bins (supplementary fig. S12, Supplementary Material online).

Importantly for our goal of trying to delineate between competing hypotheses, we found no evidence of a consistent trend that would indicate that sites within highly expressed proteins were more or less likely to show evidence of functional constraint. By contrast, the overall pattern of relative substitution rate ratios across different protein abundance bins is highly similar, casting further doubt on the hypothesis that SD-like sites within a genome are actually composed of a mixture of functionally constrained and deleterious sites.

Consistent Results for Sites Following Protein Domain Boundaries

Most studies that have explored the possible functional benefits resulting from elongation rate variability have focused on the role that slow translation or translational pausing may have in helping to enhance cotranslational protein folding. Past research has indicated that slow translation at domain boundaries may enhance protein solubility by allowing one domain to properly fold before the next domain fully emerges from the ribosome exit tunnel (Ciryam et al. 2013; Sander et al. 2014; Kim et al. 2015; Chaney et al. 2017). The most probable candidates for functional SD-like sites may thus be those sites that occur after protein domains and particularly in multidomain proteins (Zhang et al. 2009).

To test this hypothesis, we relied on previously curated protein domain annotations from Ciryam et al. (2013). After merging data sets, we were left with 415 proteins in our data set with domain annotations. We repeated our analysis within this subset of proteins, while only considering SD-like sites that occur after protein domains. We define this region as the 30–150 nucleotides downstream of 3′ domain boundaries to account for uncertainty in annotations, and maintained our previous restriction of discarding data from the first 100 and the last 50 nucleotides for each gene (effectively discarding domains that occur at the 3′ end of proteins). We specifically looked at the locally strong and locally weak sites, expecting that these categories would show the strongest signal based on our findings in figure 2*B*.

Under the hypothesis that SD-like sites after protein domains may have a functional role, we expected to observe conservation of this subset of SD-like sites (substitution rate ratios <1). A slightly weaker version of this hypothesis is that these SD-like sites should be *relatively* more conserved than SD-like sites in aggregate. If instead SD-like sites following protein domain boundaries do not represent any special category of sites, we should observe results similar to our prior findings where we observed elevated substitution rates in locally strong sites and conservation of locally weak sites.

For both codon and context controls, we found that substitution rates are significantly >1 for locally strong sites following protein domains with no substantial difference between these sites and the aggregated set of all locally strong SD-like sites (fig. 4*A*). Our results for locally weak sites were also consistent with the hypothesis that SD-like sites following protein domains are not obviously a distinct category of SD-like sites (fig. 4*B*). In both cases, we found more heterogeneity

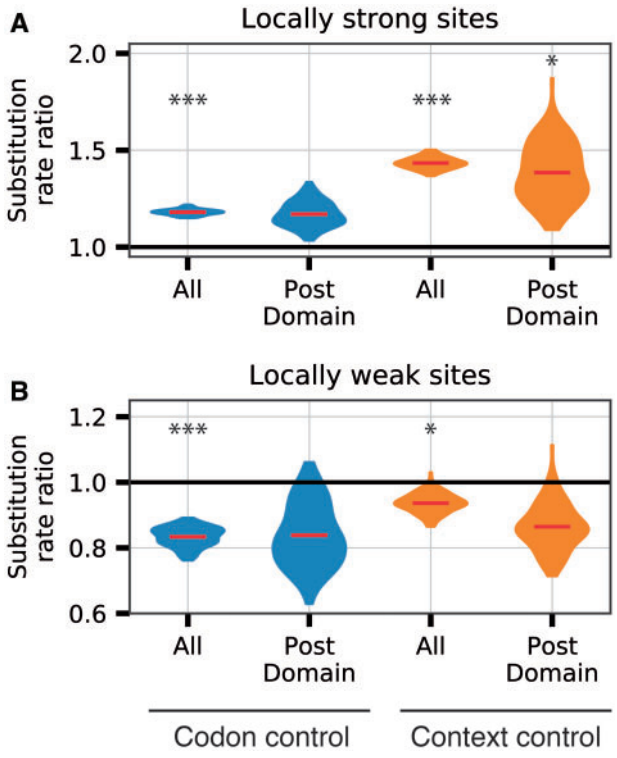

FIG. 4. Consistent results following protein domain boundaries. (*A*) Locally strong SD-like sites downstream of protein domain boundaries exhibit elevated substitution rates similar to all SD-like sites ($P = 2.7 \times 10^{-12}$, 0.2, 1.3×10^{-22}, 0.04). (*B*) Similar results to (*A*) for locally weak SD-like sites following protein domain boundaries ($P = 4.2 \times 10^{-5}$, 0.4, 0.026, 0.3). The greater heterogeneity for postdomain sites in both panels reflects the comparably small number of sites meeting the indicated criteria (see table 1). (* denotes $P < 0.05$, *** denotes $P < 0.001$).

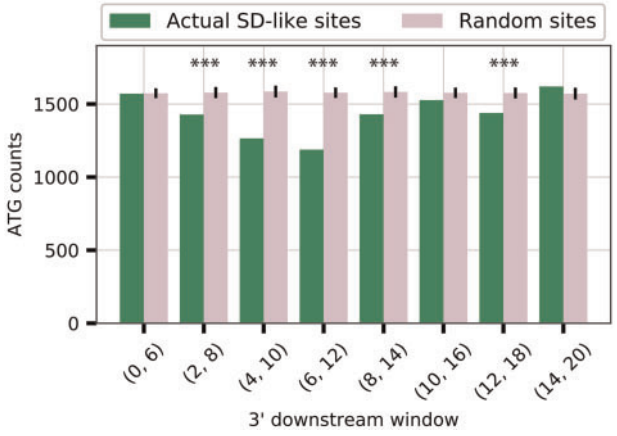

FIG. 5. Start codons are depleted downstream of SD-like sequences. We tallied the number of ATG trinucleotide sequences that occurred within the indicated windows downstream of SD-like sequences throughout the *Escherichia coli* genome. An equivalent number of random sites within each gene were selected as a control to calculate significance ($P = 8.6 \times 10^{-5}$, 3.6×10^{-15}, 5.2×10^{-27}, 9.3×10^{-5}, 0.0003 for comparisons marked as significant). (*** denotes $P < 0.001$).

in the estimates for the mean substitution rate ratios for the postdomain categories, and note that this reflects the comparably small number of SD-like sites that meet the relevant criteria for this analysis (table 1). While this analysis did not consist of exclusively multidomain proteins, over 75% of the analyzed sites occurred within multidomain proteins. Splitting the data further to isolate smaller sets of SD-like sequences was largely prohibited by the small number of sites available for this portion of the analysis (i.e., only investigating multidomain proteins and highly expressed proteins).

SD-Like Sequences and Internal Translation Initiation

All of our results with regard to sequence conservation point to SD-like sequences having elevated rates of substitution indicative of their being largely detrimental to long-term cellular fitness. But exactly what are these detrimental effects? A natural hypothesis is that SD-like sequences may result in erroneous translation initiation, which would produce truncated or frame-shifted protein products. To test whether there is evidence of this effect, we extracted nucleotide sequences downstream of all SD-like sites within the *E. coli* genome ($n = 25{,}001$). For a given downstream window, we asked how many ATG trinucleotide sequences occur (regardless of reading frame). We observed a significant depletion of ATG trinucleotide sequences (fig. 5) within a relatively narrow window downstream of SD-like sites (4–12 nucleotides) that is in line with expectations from the characteristic spacing observed in true SD sites (Hockenberry, Pah, et al. 2017). We calculated random expectation by drawing an equivalent number of random locations per-gene, performing the same analysis, and repeating this procedure 100 times. We observed no qualitative decrease in ATG counts according to this null model at different windows and calculated the significance of each window in the observed data according to this null expectation using a z-test. These results show that coding sequence patterns are constrained as a result of SD-like sequence occurrence to minimize possible translation initiation events. The detrimental effects of such erroneous translation initiation events likely explain at least part of the selection against the occurrence of SD-like sequences within protein coding genes.

Discussion

Several previous studies have shown that SD-like sequences are somewhat depleted within the protein coding genes of bacteria (Itzkovitz et al. 2010; Li et al. 2012; Diwan and Agashe 2016; Umu et al. 2016; Yang et al. 2016). These studies, however, could not comment on whether SD-like sequences are deleterious to organismal fitness or whether they are sparingly used because they serve a potentially important regulatory function. Recently, there has been a debate in the literature as to the possible functional role that SD-like sequences may play in regulating translation elongation rates with different experimental protocols yielding conflicting results (Li et al. 2012; Mohammad et al. 2016). Here, we pursued a complementary approach to investigate the possible function of SD-like sequences within bacterial protein coding genes. We performed a comparative evolutionary analysis and found

that SD-like sequences are weakly deleterious throughout the *E. coli* genome.

Using a relatively strong definition to classify SD sequences, we found that roughly 2,000 of the 4,000 *E. coli* protein coding genes are preceded by an identifiable SD sequence slightly upstream of the start codon (between −20 and −4 nucleotides relative to the start codon, see supplementary table S1, Supplementary Material online). This is substantially more than the ∼600 that would be expected based off the nucleotide composition of UTRs. However, according to this same definition, there are nearly 25,000 SD-like sequences scattered throughout *E. coli* protein coding genes (after excluding the first and last 60 nucleotides). The number of these SD-like sequences is significantly fewer than the ∼30,000 that would be expected based off of codon usage biases and amino acid sequences, but the overall magnitude of depletion is modest in scale. While these exact numbers are subject to change based on various thresholds and definitions, the facts remain that 1) there are >10 times more SD-like sequences inside *E. coli* protein coding genes than there are true SD sequences, 2) the overall depletion of SD-like sequences relative to expectation is highly significant yet small in magnitude, and 3) in the majority of cases, we do not know whether the existing SD-like sequences have any function at all.

Sequence conservation remains one of the gold standards for assessing the functionality of DNA sequences or regions (Cooper et al. 2008; Kellis et al. 2014; Ashkenazy et al. 2016). We therefore looked at the evolutionary conservation of 4-fold redundant sites that occur within SD-like sequences across *E. coli* protein coding genes. We compared the conservation sites within SD-like sequences to gene-specific control sites to determine whether there was any evidence of functional constraint. We failed to find any evidence of evolutionary conservation for the set of all SD-like sequences within our data set of Enterobacteriales, and instead found that these sequences actually have significantly elevated rates of substitution, on the order of ∼10–40% depending on the method used to select control sites. In addition to looking at all SD-like sequences, we performed a number of robustness checks and attempted to isolate subsets of likely functionally constrained SD-like sequences. However, considering sets of SD-like sequences according to 1) their overall strength of binding to the aSD sequence, 2) their occurrence within highly or lowly expressed genes, or 3) their locations relative to known protein domain boundaries did not alter our findings. All of these findings are limited to our choice of organism (*E. coli*) and phylogenetic set (Enterobacteriales) that we chose to analyze, and we cannot claim that these results are universally applicable across all bacteria. However, most previous experimental investigations into the role of SD-like sequences have likewise focused on *E. coli* and this species remains an important model system as well as a commonly used organism for engineering applications. Further experimental and comparative studies into different species and clades are necessary to comment on the universality of SD-like sequence constraints; it may be particularly informative to apply our comparative methods to species where SD-like sequences appear at a frequency equal to or even greater than null expectation (Diwan and Agashe 2016; Yang et al. 2016).

By contrast, we know that some SD-like sequences are functional and we did find that SD-like sequences that are *true* SD sequences for downstream genes in multigene operons are highly conserved. We also found that *locally weak* SD-like sequences are conserved; in these sequences, any mutation to the 4-fold redundant site in question would actually result in an increased SD-like strength. Conservation of nucleotides within these locally weak sites is therefore evidence for avoidance of strong SD-like sequences and supports our conclusion that SD-like sequences are generally deleterious.

Researchers have previously shown that SD-like sequences are capable of promoting internal translation initiation (Whitaker et al. 2015). We therefore hypothesized that the deleterious effects of SD-like sequences may be due to their role in encouraging internal translation initiation which would create truncated and/or frame-shifted protein products. Indeed, we found strong support for this hypothesis by observing that the occurrence of ATG start codons is significantly depleted within a narrow window downstream of existing SD-like sequences in *E. coli*. These data suggest that when SD-like sequences appear, they induce additional downstream constraints on coding sequence evolution and these constraints are consistent with the avoidance of translation initiation sequence features.

Since our analyses were performed on aggregates of SD-like sequences, we could not rule out whether any individual SD-like sequence or any particular set of sequences are highly conserved. In fact, we observed numerous examples of 4-fold redundant sites within SD-like sequences that are entirely conserved across all 61 species. However, the number of these sites is simply no more (and in fact, substantially fewer) than our two different null model controls. Our results do not rule out the possibility that some alternative grouping of particular genes or regions within genes that we did not consider may show increased conservation compared with null expectation. Nevertheless, based on our results and previously identified examples, the numbers of functionally constrained SD-like sequences that are involved in known regulatory processes—such as programed frame-shifting (Larsen et al. 1994; Devaraj and Fredrick 2010; Chen et al. 2014)—appear to be a small minority of all the existing SD-like sequences.

While SD-like sequences may cause spurious internal translation initiation, another possible role they may play is in regulating translational pausing (Li et al. 2012; Mohammad et al. 2016). Many studies have argued that pausing during translation can be beneficial, because it may facilitate proper protein folding (Evans et al. 2008; Zhang et al. 2009; Saunders and Deane 2010; Siller et al. 2010; Ugrinov and Clark 2010; Spencer et al. 2012; Ciryam et al. 2013; Pechmann and Frydman 2013; Fluman et al. 2014; Pechmann et al. 2014; Sander et al. 2014; Kim et al. 2015; Zhou et al. 2015; Sharma et al. 2016; Chaney et al. 2017; Sharma and O'Brien 2017). However, our results here show that the majority of SD-like sequences are deleterious. Based on these findings, we think it is unlikely that SD-like sequences are commonly used as a

means to regulate translation elongation and protein folding in endogenous genes. We cannot, of course, rule out that this effect may exist in a limited number of cases. Additionally, we cannot rule out that SD-like sequences may induce pausing but that pausing itself is mostly deleterious to organismal fitness. Further experiments are necessary to delineate whether SD-like sequence induced pausing is a real effect, and if so whether (and under what precise conditions/contexts) pausing itself may be beneficial for helping to produce properly folded native proteins.

Previous researchers showed that the usage of individual genome-wide sequence motifs is constrained according to the regulatory role of certain sequences (Hahn et al. 2003; Itzkovitz et al. 2010; Tuğrul et al. 2015; Diwan and Agashe 2016; Qian and Kussell 2016; Yang et al. 2016). However, the statistical depletion of particular motifs may result from the regulated usage of these sequences (such as transcription factor binding sequences) or because certain motifs are simply deleterious. Comparing the statistical frequencies of sequences in a genome to a null model cannot delineate between these possibilities, which is why we have taken a comparative evolutionary approach to study the conservation status of the SD-like sequences that *do* appear within genes—outside of their known regulatory context. Our findings show that these SD-like sequences tend to be either purged from closely related genomes or maintained in their weakest possible state given amino acid sequence constraints. The appearance of so many SD-like sequences throughout bacterial genomes may be explained by a combination of factors that are not intended to be exhaustive or mutually exclusive. First, SD-like sequences will continually emerge due to mutation and their ultimate numbers within genomes will be governed by mutation–selection balance. Second, some amino acid pairs and triplets are difficult to encode without strong SD-like sequences and if these amino acids are necessary for protein function then SD-like sequences may be unavoidable. Finally, the overall selective pressures acting to remove these sequences appears to be fairly weak based on our findings, and other evolutionary processes—such as genetic drift and clonal interference—may simply limit the power of natural selection to remove these sequences. Practically speaking, our findings suggest that SD-like sequences should be avoided in the design of recombinant protein expression applications until more is known about their possible deleterious effects to cellular fitness.

Materials and Methods

Data Set Compilation

We assembled a data set of 1394 homologous proteins from 61 genomes within the order *Enterobacteriales*, unique at the individual species level (see supplementary table S2, Supplementary Material online, for a complete list of analyzed genomes). We chose this set of species as a balance between identifying relatively large numbers of homologous proteins for comparative analysis (which becomes progressively more difficult with more highly diverged species) while minimizing the confounding effects of population-level polymorphisms that may occur when analyzing multiple members of a single species. We selected species based off of their inclusion in either the PATRIC "reference" or "representative" species designations (Wattam et al. 2014) and used PATRIC-derived gene annotations since these annotations derive from a consistent pipeline. For each genome, we extracted all amino acid sequences and performed a reciprocal USEARCH (Edgar 2010) comparison against *E. coli* amino acid sequences to find 1:1 best hits (using a 70% identity threshold and a strict e-value cutoff of 10^{-10}). We included all homologs that appeared in at least 45 species.

We next individually aligned the amino acid sequences of each homolog family using MUSCLE (Edgar 2004) and used RAxML (GTR model, 100 bootstrap and 20 maximum likelihood replicates) (Stamatakis 2014) to create a phylogenetic tree on the concatenated amino acid sequences of 108 genes identified in all species with the fewest number of insertions/deletions. With this tree topology, we next calculated relative nucleotide substitution rates at each position by generating aligned nucleotide sequences (based off the known nucleotide sequences of each species and the amino acid alignments) and running LEISR (Kosakovsky Pond et al. 2005; Spielman and Kosakovsky Pond 2018) under a GTR model to estimate position-specific substitution rates within each gene. We trimmed any 5′ and 3′ extensions based on the *E. coli* reference sequence annotations and then normalized each nucleotide substitution rate according to the mean of each gene.

We confirmed that the overall accuracy of relative substitution rate scores by performing several tests. We show via a meta-gene analysis that median substitution rates at 3rd positions of codons are significantly higher than 1st or 2nd positions and that substitution rates at the 5′ end of genes are lower than internal positions reflecting selection on mRNA structure surrounding the start codon (supplementary fig. S1, Supplementary Material online).

Quantifying Substitution Rate Differences between Motifs

To assess the conservation status of longer sequence motifs while controlling for amino acid and gene-specific effects, we focused exclusively on 4-fold redundant nucleotide sites (the 3rd nucleotide position of all amino acids that are encoded for by four codons and the 4-fold redundant box of amino acids that are encoded for by six codons). We opted-for this strategy to remove any artifacts that may arise from selection for or against particular amino acid sequences at different sites, but note that our analysis could in principle be applied to 2-fold redundant sites as well. We identified SD-like sites according to the computationally predicted hybridization energies between all sequential six nucleotide motifs within each gene and a putative anti-Shine–Dalgarno sequence (5′-CCUCCU-3′) using the ViennaRNA (Hofacker 2003) cofold method with default parameters. We used a threshold of −4.5 kcal/mol based-off of the distributions of true SD sequences in the *E. coli* genome (supplementary fig. S13, Supplementary Material online) to classify sequences as SD-like.

For each SD-like sequence motif that we identified, we assessed whether there are any 4-fold redundant nucleotide sites present within that subsequence. We excluded the terminal sites from our analysis (those that appear at positions 1 and 6 within the six nucleotide subsequence) since substitutions to these terminal nucleotides are much less likely to alter the actual hybridization energy between the SD-like sequence and the anti-Shine–Dalgarno sequence. If we identified a 4-fold redundant nucleotide in the central portion of the subsequence (positions 2–5), and if the amino acid site was almost entirely conserved (our selection process allowed for one possible amino acid difference across the species set) we next found all occurrences of the same synonymous codon within the same gene (so long as it does not occur within a SD-like motif) subject to the same near-perfect conservation constraint. We use these 3rd position nucleotides, where 3rd position here refers to the position within the codon (i.e., the 4-fold redundant site) as controls for the 4-fold redundant nucleotide site that occurs within the SD-like sequence. For both categories (SD-like and matched controls) we excluded nucleotides from our analysis if they fell within 100 nucleotides downstream from the *E. coli* annotated start codon or 50 nucleotides upstream from the stop codon to avoid potentially confounding effects related to translation or termination.

Additionally, we conducted a separate analysis that relied on nucleotide context for selecting control nucleotides. After finding a 4-fold redundant nucleotide in a conserved amino acid site within a SD-like sequence motif as before, we searched for another occurrence within the same gene where there is a 4-fold redundant site with the same nucleotide identity and having the same flanking nucleotides at both the +1 and −1 positions, regardless of whether the synonymous codon is the same (i.e., the −2 position). The rest of the calculation proceeded as above, with the exception that we introduced a further constraint here by requiring the +1 nucleotide to be almost perfectly conserved (less than one substitution) in addition to the amino acid under investigation.

To conservatively estimate the effect size and assess statistical significance between SD-like nucleotides and controls (given their nonnormal distribution and unequal *n*'s), we adopted a paired approach as described in the text. For each gene, we randomly selected one of the SD-like nucleotide values and one paired-control value (without replacement) until there are either no more SD-like nucleotides or no suitable control nucleotides for the given gene. We then repeated this procedure across all genes in the data set. This paired analysis method controls for gene-specific effects and creates equally sized categories, which allowed us to estimate the effect size as the ratio between the average relative substitution rates for the SD-like and control site categories. We repeated this sampling procedure 100 times to get a distribution of these ratios and assessed the significance of each bootstrap by performing a Wilcoxon signed-rank test, reporting the median observed *P* value across all replicates.

Further analyses described in text were performed following the same basic procedure as above, by either stratifying all SD-like sites into categories based on their local mutational effects, their positions within genes, or by classifying sites separately according to different gene sets.

Protein Abundance Data

We downloaded protein abundance measurements from the PaxDB database (integrated data set, accessed 07/2017) (Wang et al. 2015) and matched gene ids to the PATRIC genome annotation of *E. coli*. We were able to unambiguously map 1,386 of the 1,394 coding sequences in our complete data set to protein abundance measurements. We split these into equally sized quintile bins (each containing ~277 coding sequences) and analyzed SD-like sequence conservation separately within each set.

Protein Structural Data

Protein domain annotations were downloaded from Ciryam et al. (2013). We cross referenced annotations between our data set and theirs, and for each annotated domain analyzed SD-like sites that occurred within 150 nucleotides downstream of the domain end (while maintaining previous restrictions on 5′ and 3′ gene ends). Control sites were selected from anywhere within the same gene (outside of SD-like sequences).

Data Availability and Computer Code

Data are provided as a supplementary file and all custom scripts and code that is sufficient to perform the analysis can be found at: https://github.com/adamhockenberry/Internal_SD_conservation.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors wish to thank Stephanie J. Spielman for help with nucleotide substitution rate calculations using LEISR. This work was supported by National Institutes of Health grant R01 GM088344, National Science Foundation (Cooperative agreement no. DBI-0939454, BEACON Center), and Army Research Office (ARO, http://www.arl.army.mil/) grant W911NF-12-1-0390. L.A.N.A. and M.C.J. acknowledge a gift from Leslie and John McQuown. M.C.J. acknowledges further support from the David and Lucile Packard Foundation, and the Camille-Dreyfus Teacher-Scholar Program.

Author Contributions

A.J.H., L.A.N.A., M.C.J., and C.O.W. designed the research. A.J.H. performed the research. A.J.H. and C.O.W. wrote the article.

References

Agashe D, Sane M, Phalnikar K, Diwan GD, Habibullah A, Martinez-Gomez NC, Sahasrabuddhe V, Polachek W, Wang J, Chubiz LM, Marx CJ. 2016. Large-effect beneficial synonymous mutations mediate rapid and parallel adaptation in a bacterium. *Mol Biol Evol*. 33(6):1542–1553.

Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, Ben-Tal N. 2016. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* 44(W1):W344–W350.

Barrick D, Villanueba K, Childs J, Kalil R, Schneider TD, Lawrence CE, Gold L, Stormo GD. 1994. Quantitative analysis of ribosome binding sites in *E. coli*. *Nucleic Acids Res.* 22(7):1287–1295.

Borg A, Ehrenberg M. 2015. Determinants of the rate of mRNA translocation in bacterial protein synthesis. *J Mol Biol.* 427(9):1835–1847.

Chadani Y, Niwa T, Chiba S, Taguchi H, Ito K. 2016. Integrated in vivo and in vitro nascent chain profiling reveals widespread translational pausing. *Proc Natl Acad Sci U S A.* 113(7):E829–E838.

Chaney JL, Steele A, Carmichael R, Rodriguez A, Specht AT, Ngo K, Li J, Emrich S, Clark PL. 2017. Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLoS Comput Biol.* 13(5):e1005531–e1005519.

Chen H, Bjerknes M, Kumar R, Jay E. 1994. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs. *Nucleic Acids Res.* 22(23):4953–4957.

Chen J, Petrov A, Johansson M, Tsai A, O'Leary SE, Puglisi JD. 2014. Dynamic pathways of −1 translational frameshifting. *Nature* 512(7514):328–332.

Ciryam P, Morimoto RI, Vendruscolo M, Dobson CM, O'Brien EP, O'Brien EP. 2013. In vivo translation rates can substantially delay the cotranslational folding of the *Escherichia coli* cytosolic proteome. *Proc Natl Acad Sci U S A.* 110(2):E132–E140.

Cooper GM, Brown CD, Mcgaughey DM, Vinton RM, Huynh J. 2008. Qualifying the relationship between sequence conservation and molecular function. *Genome Res.* 18(2):201–205.

de Smit MH, van Duin J. 1990. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc Natl Acad Sci U S A.* 87(19):7668–7672.

Devaraj A, Fredrick K. 2010. Short spacing between the Shine-Dalgarno sequence and P codon destabilizes codon-anticodon pairing in the P site to promote +1 programmed frameshifting. *Mol Microbiol.* 78(6):1500–1509.

Diwan GD, Agashe D. 2016. The frequency of internal Shine-Dalgarno – like motifs in prokaryotes. *Genome Biol Evol.* 8(6):1722–1733.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32(5):1792–1797.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460–2461.

Elgamal S, Katz A, Hersch SJ, Newsom D, White P, Navarre WW, Ibba M. 2014. EF-P dependent pauses integrate proximal and distal signals during translation. *PLoS Genet.* 10(8):e1004553.

Evans MS, Sander IM, Clark PL. 2008. Cotranslational folding promotes β-helix formation and avoids aggregation in vivo. *J Mol Biol.* 383(3):683–692.

Fluman N, Navon S, Bibi E, Pilpel Y. 2014. mRNA-programmed translation pauses in the targeting of *E. coli* membrane proteins. *eLife* 3:1–19.

Frumkin I, Schirman D, Rotman A, Li F, Liron Z, Mordret E, Asraf O, Wu S, Levy SF, Pilpel Y. 2017. Gene architectures that minimize cost of gene expression. *Mol Cell* 65(1):142–153.

Hahn MW, Stajich JE, Wray GA. 2003. The effects of selection against spurious transcription factor binding sites. *Mol Biol Evol.* 20(6):901–906.

Hockenberry AJ, Pah AR, Jewett MC, Amaral LAN. 2017. Leveraging genome-wide datasets to quantify the functional role of the anti-Shine–Dalgarno sequence in regulating translation efficiency. *Open Biol.* 7(1):160239.

Hockenberry AJ, Stern AJ, Amaral LAN, Jewett MC. 2018. Diversity of Translation Initiation Mechanisms across Bacterial Species Is Driven by Environmental Conditions and Growth Demands. *Mol Bio Evol.* 35(3):582–592.

Hofacker IL. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res.* 31(13):3429–3431.

Itzkovitz S, Hodis E, Segal E. 2010. Overlapping codes within protein-coding sequences. *Genome Res.* 20(11):1582–1589.

Jacobs WM, Shakhnovich EI. 2017. Evidence of evolutionary selection for co-translational folding. *Proc Natl Acad Sci U S A.* 114(43):11434–11439.

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al. 2014. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A.* 111(17):6131–6138.

Kim SJ, Yoon JS, Shishido H, Yang Z, Rooney LA, Barral JM, Skach WR. 2015. Translational tuning optimizes nascent protein folding in cells. *Science* 348(6233):444–448.

Kosakovsky Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21(5):676–679.

Larsen B, Wills NM, Gesteland RF, Atkins JF. 1994. rRNA-mRNA base pairing stimulates a programmed −1 ribosomal frameshift. *J Bacteriol.* 176(22):6842–6851.

Li G-W, Oh E, Weissman JS. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484(7395):538–541.

Liu X, Jiang H, Gu Z, Roberts JW. 2013. High-resolution view of bacteriophage lambda gene expression by ribosome profiling. *Proc Natl Acad Sci U S A.* 110(29):11928–11933.

Ma J, Campbell A, Karlin S. 2002. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol.* 184(20):5733–5745.

Martens AT, Taylor J, Hilser VJ. 2015. Ribosome A and P sites revealed by length analysis of ribosome profiling data. *Nucleic Acids Res.* 43(7):3680–3687.

Mohammad F, Woolstenhulme CJ, Green R, Buskirk AR. 2016. Clarifying the translational pausing landscape in bacteria by ribosome profiling. *Cell Rep.* 14(4):686–694.

Nakagawa S, Niimura Y, Miura K-I, Gojobori T. 2010. Dynamic evolution of translation initiation mechanisms in prokaryotes. *Proc Natl Acad Sci U S A.* 107(14):6382–6387.

Pechmann S, Chartron JW, Frydman J. 2014. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. *Nature Structural & Molecular Biology* 21(12):1100–1105.

Pechmann S, Frydman J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat Struct Mol Biol.* 20(2):237–243.

Qian L, Kussell E. 2016. Genome-wide motif statistics are shaped by DNA binding proteins over evolutionary time scales. *Phys Rev X* 6(4):

Sander IM, Chaney JL, Clark PL. 2014. Expanding Anfinsen's principle: contributions of synonymous codon selection to rational protein design. *J Am Chem Soc.* 136(3):858–861.

Saunders R, Deane CM. 2010. Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Res.* 38(19):6719–6728.

Schrader JM, Zhou B, Li G-W, Lasker K, Childers WS, Williams B, Long T, Crosson S, McAdams HH, Weissman JS, Shapiro L. 2014. The coding and noncoding architecture of the *Caulobacter crescentus* genome. *PLoS Genet.* 10(7):e1004463.

Sharma AK, Bukau B, O'Brien EP. 2016. Physical origins of codon positions that strongly influence cotranslational folding: a framework for controlling nascent-protein folding. *J Am Chem Soc.* 138(4):1180–1195.

Sharma AK, O'Brien EP. 2017. Increasing protein production rates can decrease the rate at which functional protein is produced and their steady-state levels. *J Phys Chem B* 121(28):6775–6784.

Shine J, Dalgarno L. 1974. The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A.* 71(4):1342–1346.

Siller E, DeZwaan DC, Anderson JF, Freeman BC, Barral JM. 2010. Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. *J Mol Biol.* 396(5):1310–1318.

Sohmen D, Chiba S, Shimokawa-Chiba N, Innis CA, Berninghausen O, Beckmann R, Ito K, Wilson DN. 2015. Structure of the *Bacillus subtilis*

70S ribosome reveals the basis for species-specific stalling. *Nat Commun*. 6(1):6941.
Spencer PS, Siller E, Anderson JF, Barral JM. 2012. Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. *J Mol Biol*. 422(3):328–335.
Spielman SJ, Kosakovsky Pond SL. 2018. Relative evolutionary rate inference in HyPhy with LEISR. *PeerJ* 6:e4339.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
Subramaniam AR, DeLoughery A, Bradshaw N, Chen Y, O'Shea E, Losick R, Chai Y. 2013. A serine sensor for multicellularity in a bacterium. *eLife* 2:1–17.
Takahashi S, Tsuji K, Ueda T, Okahata Y. 2012. Traveling time of a translating ribosome along messenger RNA monitored directly on a quartz crystal microbalance. *J Am Chem Soc*. 134(15):6793–6800.
Tuğrul M, Paixão T, Barton NH, Tkačik G. 2015. Dynamics of transcription factor binding site evolution. *PLoS Genet*. 11(11):e1005639.
Ugrinov KG, Clark PL. 2010. Cotranslational folding increases GFP folding yield. *Biophys J*. 98(7):1312–1320.
Umu SU, Poole AM, Dobson RCJ, Gardner PP. 2016. Avoidance of stochastic RNA interactions can be harnessed to control protein expression levels in bacteria and archaea. *eLife* 5:1–16.
Vasquez KA, Hatridge TA, Curtis NC, Contreras LM. 2016. Slowing translation between protein domains by increasing affinity between mRNAs and the ribosomal anti-Shine-Dalgarno sequence improves solubility. *ACS Synth Biol*. 5(2):133–145.
Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. 2015. Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* 15(18):3163–3168.
Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, et al. 2014. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res*. 42(D1):D581–D591.
Wen J-D, Lancaster L, Hodges C, Zeri A-C, Yoshimura SH, Noller HF, Bustamante C, Tinoco I. 2008. Following translation by single ribosomes one codon at a time. *Nature* 452(7187):598–603.
Whitaker WR, Lee H, Arkin AP, Dueber JE. 2015. Avoidance of truncated proteins from unintended ribosome binding sites within heterologous protein coding sequences. *ACS Synth Biol*. 4(3):249–257.
Yang C, Hockenberry AJ, Jewett MC, Amaral LAN. 2016. Depletion of Shine-Dalgarno sequences within bacterial coding regions is expression dependent. *G3 (Bethesda)* 6:3467–3474.
Zhang G, Hubalewska M, Ignatova Z. 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat Struct Mol Biol*. 16(3):274–280.
Zhou M, Wang T, Fu J, Xiao G, Liu Y. 2015. Nonoptimal codon usage influences protein structure in intrinsically disordered regions. *Mol Microbiol*. 97(5):974–987.