# Diversity of Translation Initiation Mechanisms across Bacterial Species Is Driven by Environmental Conditions and Growth Demands

Adam J. Hockenberry,[1,2] Aaron J. Stern,[1] Luís A.N. Amaral,*[,1,3,4] and Michael C. Jewett*[,1,3,5,6]

[1]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA
[2]Interdisciplinary Program in Biological Sciences, Northwestern University, Evanston, IL, USA
[3]Northwestern Institute for Complex Systems, Northwestern University, Evanston, IL, USA
[4]Department of Physics and Astronomy, Northwestern University, Evanston, IL, USA
[5]Center for Synthetic Biology, Northwestern University, Evanston, IL, USA
[6]Simpson Querrey Institute for BioNanotechnology, Northwestern University, Evanston, IL, USA

***Corresponding authors:** E-mails: adam.hockenberry@utexas.edu; m-jewett@northwestern.edu.

**Associate editor:** Deepa Agashe

## Abstract

**The Shine–Dalgarno (SD) sequence motif is frequently found upstream of protein coding genes and is thought to be the dominant mechanism of translation initiation used by bacteria. Experimental studies have shown that the SD sequence facilitates start codon recognition and enhances translation initiation by directly interacting with the highly conserved anti-SD sequence on the 30S ribosomal subunit. However, the proportion of SD-led genes within a genome varies across species and the factors governing this variation in translation initiation mechanisms remain largely unknown. Here, we conduct a phylogenetically informed analysis and find that species capable of rapid growth contain a higher proportion of SD-led genes throughout their genomes. We show that SD sequence utilization covaries with a suite of genomic features that are important for efficient translation initiation and elongation. In addition to these endogenous genomic factors, we further show that exogenous environmental factors may influence the evolution of translation initiation mechanisms by finding that thermophilic species contain significantly more SD-led genes than mesophiles. Our results demonstrate that variation in translation initiation mechanisms across bacterial species is predictable and is a consequence of differential life-history strategies related to maximum growth rate and environmental-specific constraints.**

***Key words:*** **translation initiation, Shine–Dalgarno sequence, bacterial growth, genome evolution.**

Article

## Introduction

Translation of a given messenger-RNA (mRNA) into functional protein relies on the ability of the translational apparatus to recognize the proper start codon. Bacteria have evolved several distinct mechanisms to discriminate between potential start codons, with the Shine–Dalgarno (SD) mechanism being the most well-studied (Shine and Dalgarno 1974; Nakagawa et al. 2010). Variants of the SD sequence are frequently found upstream of bacterial start codons where they facilitate translation initiation by hybridizing with the complementary anti-SD (aSD) sequence on the 16S rRNA of the small ribosomal subunit (fig. 1A).

For a given gene within an organism, the structural accessibility of the SD sequence, the thermodynamic binding potential between the SD sequence and the aSD sequence, and the exact positioning of the SD sequence relative to the start codon, are all features that can be predictably tuned in order to modulate the translation initiation rate of downstream genes (Barrick et al. 1994; de Smit and van Duin 1994; Vimberg et al. 2007; Salis et al. 2009; Devaraj and Fredrick 2010; Na et al. 2010; Kosuri et al. 2013; Espah Borujeni et al. 2014; Bonde et al. 2016; Espah Borujeni and Salis 2016; Hockenberry et al. 2017). In particular, researchers have shown that transcripts with a strong SD sequence are translated at higher rates resulting in more protein being produced per mRNA, a fact which is particularly important for the design of recombinant protein expression systems (Salis et al. 2009; Na et al. 2010; Bonde et al. 2016). Nevertheless, there are several SD sequence-independent mechanisms that operate in bacteria including leaderless translation and RPS1-mediated translation of unstructured mRNA sequences (Komarova et al. 2005; Chang et al. 2006; Gu et al. 2010; Scharff et al. 2011; Zheng et al. 2011; Keller et al. 2012; Barendt et al. 2013; Cortes et al. 2013; Duval et al. 2013; Kramer et al. 2014; Shell et al. 2015). Recent research also suggests that mechanisms traditionally associated with eukaryotic species such as translational scanning and internal ribosome entry sites may operate in bacterial systems (Colussi et al. 2015; Yamamoto et al. 2016).

The aSD sequence is highly conserved throughout the bacterial domain (though notable exceptions exist) and

© The Author 2017. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

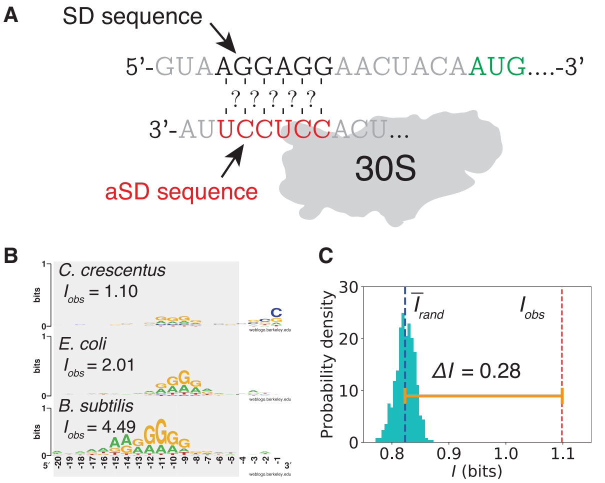

Fig. 1. Sequence entropy quantifies genome-wide SD sequence utilization. (*A*) Illustration of the anti-Shine–Dalgarno(aSD)::Shine–Dalgarno(SD) sequence mechanism of translation initiation. (*B*) Representative sequence logos of the 5′ upstream region of all annotated coding sequences for individual genomes displays heterogeneity in sequence entropy within and between species. (*C*) Illustration of the Δ*I* metric for *Caulobacter crescentus* as an example.

stronger SD sequence interactions are associated with increased translation efficiency across a wide array of species (Chen et al. 1994; Osada et al. 1999; Sakai et al. 2001; Ma et al. 2002; Starmer et al. 2006; Lim et al. 2012; Omotajo et al. 2015; Hockenberry et al. 2017). We therefore might expect that most species would utilize the SD sequence mechanism to a similar degree. However, there is a wide-diversity of SD sequence utilization between different species. For instance, roughly 90% of *Bacillus subtilis* genes are preceded by a SD sequence, whereas for *Caulobacter crescentus*, the comparable number is closer to 50% (Chang et al. 2006; Starmer et al. 2006; Nakagawa et al. 2010; Schrader et al. 2014). SD-like sequence motifs are also underrepresented within the coding sequences of most bacteria—possibly reflecting their role in translational pausing and/or erroneous initiation—and like the diversity of SD sequence utilization, the degree of this underrepresentation varies from species-to-species (Li et al. 2012; Diwan and Agashe 2016; Mohammad et al. 2016; Yang et al. 2016).

Cross-species variation in translation initiation mechanisms may impact genetic isolation and transfer of genetic material, and quantifying the source and extent of variation may prove useful in identifying important genes in a genome or microbial community(Krisko et al. 2014; Omotajo et al. 2015). Further, the synthetic biology community is increasingly targeting both translation-system engineering and biotechnology applications involving less well-studied microbial species (Tauer et al. 2014; Markley et al. 2015; Orelle et al. 2015; Guiziou et al. 2016; Weinstock et al. 2016; Yi et al. 2016). A better understanding of the factors shaping the utilization of different translation initiation mechanisms may ultimately aid in the design of synthetic gene constructs.

Here, we conduct a phylogenetic comparative analysis in order to isolate independent evolutionary events and determine whether any endogenous or exogenous factors are predictive of genome-wide SD sequence utilization. We develop a metric that captures position-dependent sequence preferences within the translation initiation region of a given genome, and demonstrate a strong link between SD sequence utilization and minimum doubling times for 187 species. Furthermore, in a database of 613 phylogenetically diverse bacterial species, we show that genome-wide variation in SD sequence utilization covaries along-side a number of genomic features that are indicative of rapid-growth and efficient translation. Finally, we investigate exogenous environmental constraints and show that SD sequence utilization varies according to optimal growth temperatures.

## Results

### Quantifying Genome-Wide SD Sequence Utilization

Several techniques have been previously developed to quantify the overall utilization of the aSD::SD mechanism within a given species. In motif-based methods, researchers predefine subsequences closely related to the canonical SD sequence and search a sequence window upstream of each protein coding gene within a given genome to determine the fraction of genes that are preceded by a SD motif (Chang et al. 2006; Omotajo et al. 2015). Similarly, in aSD sequence complementarity based methods, researchers predefine a window upstream of the start codon to consider for each gene, a putative aSD sequence, and a hybridization energy threshold for determining whether a gene is SD-led or not (Osada et al. 1999; Sakai et al. 2001; Ma et al. 2002; Starmer et al. 2006; Nakagawa et al. 2010).

Both of these metrics rely on critical assumptions that may not hold when applied across large sets of phylogenetically diverse organisms. First, they carry an implicit assumption that a SD sequence, regardless of its location relative to the start codon, has the same impact on translation initiation. However, experimental approaches have shown that spacing between the SD sequence and start codon can have dramatic effects on translation initiation rates (Vimberg et al. 2007; Salis et al. 2009; Devaraj and Fredrick 2010; Hockenberry et al. 2017). Second, both methods rest on a dichotomy between SD-led and non-SD-led genes. Although this simplification is useful for *describing* the phenomenon, an abundance of research has shown that a spectrum of sequence complementarity affects translation initiation in a continuous manner (Vimberg et al. 2007; Salis et al. 2009). Third, bacterial genomes span a range of GC contents, and previous research has shown that it is possible to account for this bias by comparing the proportion of SD-led genes in a genome to an appropriate null model expectation (Nakagawa et al. 2010). We define the following term to summarize SD sequence utilization using the SD-motif based method by:

$$\Delta f_{SD} = f_{SD,obs} - \overline{f}_{SD,rand} \quad (1)$$

where $f_{SD}$ is the fraction of genes within a genome classified as SD-led and $\overline{f}_{SD,rand}$ is the expected fraction of SD-led genes derived from repeating this calculation for 500 nucleotide shuffled “genomes” (where 30nts of the 5′-UTR for each gene are randomly permuted, see Materials and Methods). We similarly define $\Delta f_{aSD<-4.5}$, where SD-led genes are defined via hybridization of the putative aSD sequence using a threshold binding energy value of −4.5 kcal/mol (see Materials and Methods).

## Sequence Entropy and Its Relationship with SD Sequence Utilization

We sought a complementary approach that would allow us to investigate hundreds of diverse genomes without having to a priori define either an aSD sequence or SD sequence motifs. To measure arbitrary position-specific sequence signals, we extract the 5′ upstream sequences from all annotated protein coding sequences (see Materials and Methods) and sum the "information content" at each position within the region where SD motifs are expected to occur (−20 to −4 relative to the start codon):

$$I_{obs} = \sum_{i=-20}^{-4} \left( \log_2 4 - \sum_{k \in \{A,T,G,C\}} p_{i,k} \log_2 p_{i,k} \right) \qquad (2)$$

where $p_{i,\,k}$ is the empirical frequency of base $k$ at position $i$. The right side of this equation is the entropy (calculated from empirical base frequencies at each position) while the $\log_2 4$ component scales the data in line with what is typically represented in sequence logo plots such that individual nucleotide positions have a maximum value of 2 when fully conserved and a minimum of 0 when fully random. We again repeat this process for 500 shuffled "genomes" (as described above) and compare the sequence information from the actual genome to this null expectation:

$$\Delta I = I_{obs} - \bar{I}_{rand} \qquad (3)$$

For a more intuitive interpretation, figure 1B shows sequence logo plots of the 5′-UTRs for three genomes to highlight the variation in sequence preferences between species. $I_{\text{obs}}$ is simply the sum across each position of these nucleotide "stacks." To account for uneven nucleotide usage, we calculate $\Delta I$ by comparing this value to nucleotide shuffled controls (fig. 1C). In supplementary figure S1, Supplementary Material online, we simulate sequences to show that $\Delta I$ takes larger values when (all else being equal): 1) a larger fraction of simulated genes contain a particular sequence motif, 2) the enriched sequence motif is longer, and 3) the enriched sequence motif is more strictly defined in its position.

We compiled a data set of 613 bacterial species, unique at the genus level, with annotated genome-sequences as well as a previously constructed high-quality phylogenetic tree describing their relatedness (Hug et al. 2016) (see Materials and Methods). In figure 2A, we show that while summary methods based on SD-motif and aSD sequence complementarity ($\Delta f_{\text{SD}}$ and $\Delta f_{aSD<-4.5}$, respectively) are strongly related for a large set of diverse species, this association is not perfect and there is a distinct change in the slope that occurs for the Firmicutes phylum.

In principle, $\Delta I$ may quantify a variety of position-specific sequence signals that may or may not be related to the SD sequences or translation initiation. In practice, we observe that this metric correlates strongly with both methods traditionally used to describe SD sequence utilization (fig. 2A and supplementary fig. S2, Supplementary Material online). However, in the Bacteroidetes phylum, we detect significant variation in $\Delta I$ without any apparent variation in either of the other two metrics. These findings are consistent with prior research that identified changes in the aSD sequence region of the 16 S rRNA sequence within this phylum (Lim et al. 2012), and indicate that Bacteroidetes may nevertheless contain position-specific translation initiation sequences. The sequence logos for representative Bacteroidetes species showed A/T rich UTRs (supplementary fig. S3, Supplementary Material online), possibly indicative of translation initiation mediated by ribosomal protein S1. However, we stress that general nucleotide bias is removed in our calculation of $\Delta I$ via shuffling and the A/T richness in these species instead has a fairly strong position-specific signal. The fact that the $\Delta I$ metric quantifies position-specific translation initiation signals for Bacteroidetes allows us to incorporate them into future analyses (fig. 2A, red data points) though we note that much remains to be understood about possible changes to the identity of the SD and aSD sequences in this phyla.

Due to the strong relationship between $\Delta I$ and explicit SD sequence methods, we refer to $\Delta I$ as quantifying SD sequence utilization throughout the remainder of this manuscript and make explicit note of differences between SD sequence quantification methods when necessary. Consistent with prior research (Nakagawa et al. 2010), we show that SD sequence utilization according to the $\Delta I$ metric varies considerably across species while showing broad phyla-specific patterns that should be accounted for when performing statistical tests (fig. 2B, bar heights). We additionally found that the UTRs from ribosomal protein coding genes contained more position-dependent sequence information—using either $I_{\text{obs}}$ or $\Delta I$, for nearly all species in our data set—compared with genome-wide UTRs (supplementary fig. S4, Supplementary Material online). This observation is consistent with theory emphasizing the overall importance of efficient translation for the most highly expressed proteins. We also tested whether $\Delta I$ varies according to genome size and found no significant association with either genome length or the number of annotated protein coding genes ($P = 0.54$ and 0.84, respectively).

## Translation Initiation and Organismal Growth Demands

In prior research, Vieira-Silva et al. (2010) curated a list of minimum doubling times from the literature for a large number of bacterial species (Roller et al. 2016). Organisms that are capable of rapid growth have high protein production demands during these periods and there are a number of regulatory points that can be bottlenecks for this process. Meeting high translational demands associated with rapid growth requires coordination of a number of processes, and Vieira-Silva et al. (2010) showed that increasing numbers of rRNA and tRNA genes, and increasing codon usage biases among ribosomal mRNAs in individual genomes were all

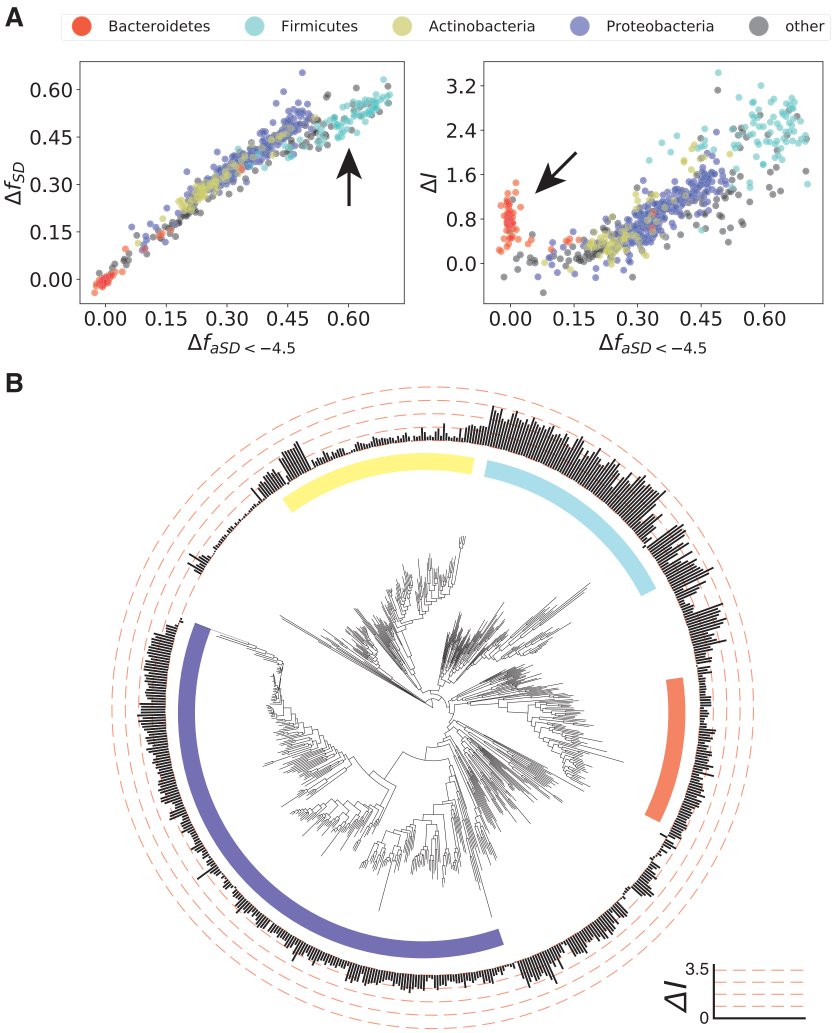

Fig. 2. Relationship between $\Delta I$ and existing metrics of SD sequence utilization. (*A*) Comparison between different ways of summarizing SD sequence utilization; each data point represents a single genome. On the left, we show the relationship between SD motif and aSD sequence complementarity based methods ($\Delta f_{\mathrm{SD}}$ and $\Delta f_{aSD<-4.5}$). On the right, we compare $\Delta I$ and $\Delta f_{aSD<-4.5}$. The four largest phyla are color-coded according to the legend. Arrows highlight phyla with "anomalous" patterns. (*B*) Phylogenetic tree illustrating variation in SD sequence utilization across species according to the $\Delta I$ metric (indicated by bar plots on concentric rings).

partially predictive of the minimum doubling times of individual species. Subsequently, Yang et al. (2016) showed that the extent of anti-SD sequence binding across coding sequences was also predictive of minimum doubling times reflecting the possible influence of internal SD-like sequences on translational pausing and/or spurious translation initiation events (Yang et al. 2016).

At the individual gene level, translation initiation is an important control point and likely to be rate-limiting for the production of most proteins. Experimental evidence largely supports this conjecture, with alterations to SD sequence utilization, mRNA structural availability, and start codon identities predictably tuning protein production (Kudla et al. 2009; Salis et al. 2009; Goodman et al. 2013; Kosuri et al. 2013; Espah Borujeni et al. 2014, 2016; Espah Borujeni and Salis 2016; Hecht et al. 2017). We thus reasoned that efficient translation initiation across all or subsets of genes may similarly play an important role in meeting protein production demands imposed by rapid growth rates. In addition to genome-wide SD sequence utilization, we also quantified the percentage "ATG" start codons in a genome as well as the average difference in mRNA folding energy surrounding the start codons relative to a control region internal to the gene (see Materials and Methods).

We first replicated several of the findings of Vieira-Silva et al. (2010) and Yang et al. (2016) using Phylogenetically Generalized Least Squares regression (Revell 2010) to account for the lack of independence in species (see Materials and Methods). We verified that rRNA gene counts, tRNA gene counts, a metric of relative codon usage bias (based off the "effective number of codons" [$\Delta ENC$], see Materials and Methods), and a metric of internal SD-like sequence binding (see Materials and Methods) are all significantly predictive of minimum doubling times after controlling for phylogenetic effects (*F*-test, $P < 0.002$ for all cases, table 1).

Next, we turned to metrics related to translation initiation. We found that $\Delta I$ calculated over all genes within a genome significantly correlates with minimum doubling times in this

**Table 1.** Contribution of Several Factors for Predicting Minimum Doubling Times.

| Model for Min. Doubling Time | $R^2$ | Pagel's $\lambda$ [95% CI] | $\|\Delta R^2\|$ |
|---|---|---|---|
| Full model | 0.35*** | 0.91 [0.80, 0.96] | – |
| $\Delta ENC$ | 0.17*** | 0.96 [0.92, 0.99] | 0.15 |
| $\Delta I$ | 0.11*** | 0.97 [0.93, 0.99] | 0.12 |
| mRNA folding | 0.08*** | 0.98 [0.95, 0.99] | 0.04 |
| Internal SD-like | 0.08*** | 0.98 [0.95, 0.99] | 0.03 |
| 16S gene counts | 0.06** | 0.98 [0.95, 0.99] | 0.01 |
| tRNA gene counts | 0.06*** | 0.98 [0.95, 0.99] | 0.01 |
| ATG start % | 0.02 | 0.98 [0.95, 0.99] | <0.01 |

NOTE.—The left column indicates individual variables that we considered for predicting minimum doubling times with the full multivariate model listed at the top. $R^2$ column illustrates the overall goodness-of-fit for individual factors (*** indicates $P < 0.001$, ** indicates $P < 0.01$). Pagel's $\lambda$ is the fitted phylogenetic signal parameter, which we show with 95% confidence intervals in brackets. The right column illustrates the change in goodness-of-fit from a model that includes all predictors to one that excludes only the variable in the given row.

set of species ($P < 10^{-5}$), showing the second strongest correlation of any individual trait that we considered (table 1 and fig. 3A). Our metric of mRNA structural stability surrounding the start codons—"mRNA folding" in table 1—similarly showed a significant relationship such that genomes that, on an average, have the weakest mRNA structure surrounding the start codon are more capable of rapid growth. By contrast, the proportion of protein coding genes containing an ATG start codon is not significantly correlated with minimum doubling times ($P = 0.056$), but showed the expected slope indicative of faster growth for genomes with more ATG start codons.

In order to test the robustness of these findings and to assess *overall* predictability of minimum doubling times from these features, we constructed a multivariable Phylogenetic Generalized Least Squares regression model that combines all of the above factors, and found that only relative codon usage biases ($\Delta ENC$), SD sequence utilization ($\Delta I$), and mRNA folding energy had statistically significant coefficients in this full model ($P < 0.01$). Overall, a model containing all factors resulted in $R^2 = 0.35$ ($P < 10^{-13}$, fig. 3B), whereas a more parsimonious model containing only the three factors with statistically significant coefficients resulted in $R^2 = 0.31$ ($P < 10^{-14}$). Removing either of these features from the full model reduces its predictive power as illustrated in the right column of table 1, which further emphasizes the large contribution of $\Delta ENC$ and $\Delta I$ in particular.

In order to compare our work with prior research, we also conducted a phylogenetically *agnostic* linear regression model using all of these factors, which yielded $R^2 = 0.6$ ($P < 10^{-15}$)—though we caution that ignoring the effects of shared ancestry will substantially bias statistical analyses. We further generated the same data as in table 1 using $\Delta f_{aSD < -4.5}$ as a metric of SD sequence utilization and found largely similar results with slightly less predictive power overall: $R^2 = 0.29$ for the full model and $R^2 = 0.09$ as an individual predictor compared with 0.35 and 0.11 using $\Delta I$ (supplementary table S1, Supplementary Material online). We repeated these analyses excluding species from Bacteroidetes phylum, but this restriction removed only five species and correlations for both $\Delta I$ and $\Delta f_{aSD < -4.5}$ independently and in the full model remained essentially unchanged.

Interestingly, we looked at the effect of using only using ribosomal protein coding genes to calculate $\Delta I$ and found that this metric performed worse at predicting observed minimum doubling times individually ($R^2 = 0.04$) and as part of the full model ($R^2 = 0.27$) (supplementary table S2, Supplementary Material online). The metrics of codon usage bias and internal SD-like sequence usage that we used here are both calculated from the *difference* between values for ribosomal protein coding genes and all genes within a genome, thus representing relative differences in selection strength acting specifically on ribosomal protein coding genes. However, we observed little predictive power when using a relative difference metric between ribosomal protein coding genes and all genes within a genome—using either $I_{obs}$ or $\Delta I$ to predict minimum doubling times ($R^2 \approx 0$). Instead, genome-wide calculations of $\Delta I$ consistently provided the highest predictive power.

As a final caveat for this data, we do note that the list of species compiled by Vieira-Silva et al. (2010) is phylogenetically biased toward Proteobacteria, contains only culturable species, and is distinct in its composition from the data set analyzed in figure 2B.

## Relationship between SD Sequence Utilization and Other Translation Efficiency-Associated Traits

Since a coordinated effort between multiple processes is required to maximize protein production in experimental systems, we reasoned that the various genome-wide traits associated with efficient translation are likely to covary with one another across species. In order to test this hypothesis, we assessed the correlation between different definitions of SD sequence utilization and all of the alternative traits listed in table 1 via Phylogenetic Generalized Least Squares regression. In figure 4A, we show the results of this analysis, finding that in all cases where a pair of traits is significantly correlated, the correlation is positive (note that the color bar corresponds to $R$ and not $R^2$). Increasing SD sequence utilization is thus significantly associated with an increasing fraction of ATG start codons, less structured mRNA folding around the start codons, increased 16S rRNA and tRNA gene counts, and increasing codon usage bias/avoidance of internal SD-like sequences within ribosomal protein coding genes.

We next tested the overall robustness and universality of these results by independently analyzing these relationships *within* individual phyla. We specifically looked at the four largest phyla in our data set—Proteobacteria, Firmicutes, Actinobacteria, and Bacteroidetes—and repeated the analysis from figure 4A using independent model fits. Again, we observed that nearly every significant correlation is in the positive direction (supplementary fig. S5, Supplementary Material online) with the exception of mRNA structure and SD sequence utilization in the Firmicutes phylum. When looking at relationships between variable SD sequence utilization in the Bacteroidetes phylum, $\Delta I$ has a significantly positive

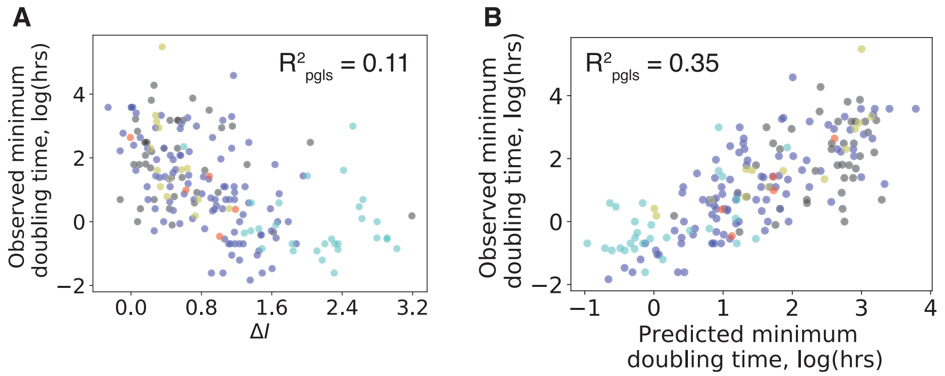

**FIG. 3.** Relationship between SD sequence utilization and organismal growth. (*A*) Δ*I* is significantly correlated with minimum observed doubling times for 187 bacterial species. (*B*) Visualization of the full model listed in table 1 depicting a strong relationship between observed and predicted minimum doubling times. In both plots, individual species data points are colored according to phyla as in figure 2*A*.

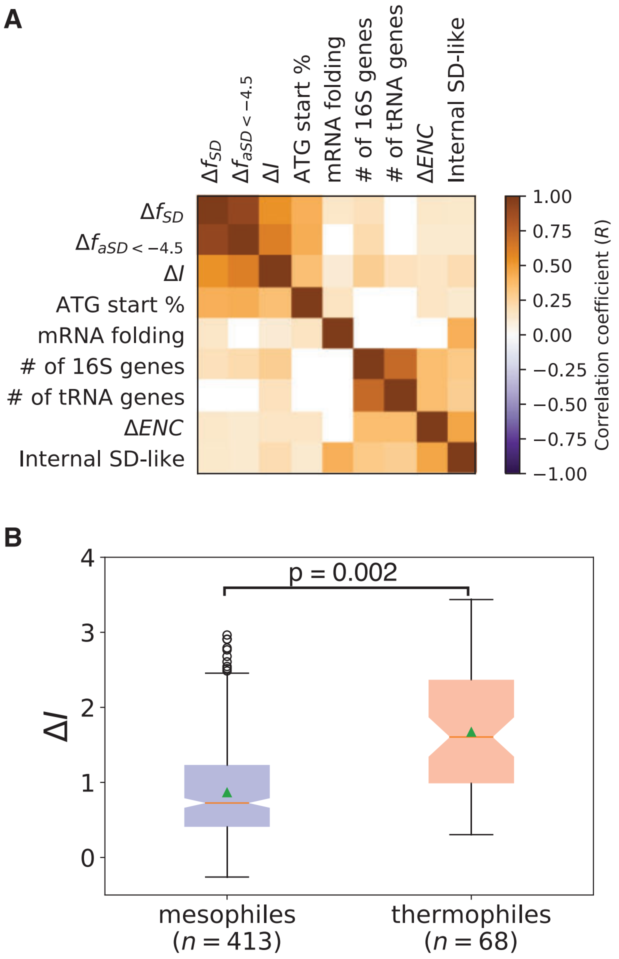

**FIG. 4.** SD sequence utilization covaries alongside a suite of translation-related traits and according to optimal growth temperatures. (*A*) Correlation matrix between listed variables used in table 1 for a set of 613 diverse bacterial species. In all instances of significant correlation, the features covary with one another in the positive direction. (*B*) SD sequence utilization, quantified using Δ*I* is significantly higher in thermophiles than in mesophiles. Box limits show 25th and 75th percentiles of the data, whiskers extend to 5th and 95th percentiles, triangles depict the means of each category and red lines highlight the median.

relationship with several other variables whereas $\Delta f_{SD}$ and $\Delta f_{aSD<-4.5}$ show no significant relationships apart from with one-another.

## Relationship between Translation Initiation Mechanisms and Ecological Factors

Having established that genome-scale SD sequence utilization is part of a suite of traits related to differential organismal growth strategies, we last wanted to assess whether ecological factors pertaining to an organisms habitat may constrain the evolution of SD sequence utilization. The aSD:: SD sequence interaction is thought to increase translation rates by stabilizing the assembly of the translation initiation complex. Since this interaction is based on RNA base-pairing, pairing between longer sequences may be necessary in order to get an equivalent level of stabilization at higher temperatures. As we show in supplementary figure S1, Supplementary Material online, Δ*I* takes larger values when longer sequences are preferred so we reasoned that there might be an association between this metric and optimal growth temperatures.

Nakagawa et al. (2010) investigated this possibility, but found no association (Nakagawa et al. 2010). By contrast, our phylogenetically informed modeling approach applied to this larger data set (481 of the 613 species in our data set have high-confidence growth temperature annotations) finds that temperature constrains genome-wide SD sequence utilization. Specifically, the genomes of thermophilic species display significantly larger values of Δ*I* than mesophilic species (fig. 4*B*, *F*-test $P = 0.002$ using temperature as a fixed-effect in Phylogenetically Generalized Least Squares modeling) in line with our hypothesis. Although optimal growth temperature is a continuous variable, in practice, measurements for the vast majority of species fall into discrete categories (30 °C and 37 °C). This fact, coupled with the availability of a large database of "mesophile" and "thermophile" annotations motivated our decision to study this effect using discrete temperature categories.

Our finding here illustrates the role that ecological factors relating to growth conditions places on the evolution of genome architectures. We further tested whether there are any systematic differences in SD sequence utilization by looking at free-living/host-associated species, or pathogenic/nonpathogenic species, and found no significant effects for either ($P = 0.91$ and 0.44, respectively). Nevertheless, it is possible that other ecological and life-history traits that we did not consider here may play an important role in constraining the evolution and utilization of different translation initiation mechanisms.

## Discussion

Our study shows a relationship between bacterial translation initiation mechanisms, life-history strategies, and environmental demands faced by individual species. Although we do not specifically address causality, we found that minimum observed doubling times and SD sequence utilization at the genome-scale are significantly correlated (fig. 3A). In a larger and more diverse data set of 613 species, we further showed that SD sequence utilization predictably covaries with several other genomic and environmental features, including the number of rRNA genes and optimal growth temperatures. Taken together, our findings demonstrate that organisms with greater translational demands are likely to coevolve a common suite of genomic features that help to maximize translation during periods of rapid growth, and that SD sequence utilization is an important component of this shared genome architecture.

Our analysis throughout is performed in a manner that corrects for the confounding effects of shared ancestry between species, and our phyla specific results illustrate several critical points. First, the sign on the relationships between individual features that we observe is extremely robust, regardless of the phylum or SD sequence utilization summary statistic under consideration (fig. 4A and supplementary fig. S5, Supplementary Material online). Increasing 16S/tRNA gene counts, increasing codon usage biases/avoidance of internal SD-like sequences in ribosomal protein genes, higher ATG start codon usage, and weak mRNA structure surrounding the start codon are associated with increasing SD sequence utilization. Second, we developed a novel metric ($\Delta I$) to measure position-specific translation initiation region sequence preferences in a manner that is independent of the anti-SD or SD sequence for a given species. We found that $\Delta I$ is highly correlated with prior estimates of SD sequence utilization, but that this relationship breaks down in the Bacteroidetes phylum, which may reflect novel position-specific sequence preferences in this lineage. Future research applying our methodology to larger data sets may allow researchers to uncover branches within phylogenetic trees where mechanistic differences in the translational apparatus—resulting in differences in the slope and/or sign on the relationships between different features—have evolved.

Despite the overall robustness of our results, there are several caveats to our study that we wish to explicitly highlight for readers. First, the lack of available growth rate data for most phyla prohibits us from looking at the relationships between genomic traits and minimum doubling times on a phyla specific level. The species for which we do have minimum doubling time estimates are biased phylogenetically and represent only a fraction of the known bacterial universe. Second, our findings throughout rely on genome annotations, which frequently rely on the presence of SD sequences to identify the 5′ ends of genes. Such annotations are known to be imperfect, and as many as ≈ 10% of genes in a given genome may have misannotated start codons (Schrader et al. 2014). However, these biases are likely to be uniform—or at least randomly distributed—throughout our data set making it unlikely that systematic differences in start codon annotation between genomes would produce spurious results of the order that we report here. Nevertheless, continued improvement in start codon annotations is an important issue and potential limitation of our study.

Our results add to the body of knowledge showing that a small number of genomic traits—that includes utilization of the SD sequence mechanism—can be used to predict variation in minimum doubling times. However, a critical question that remains unanswered is: what is the causal relationship between different genome-scale phenomena and growth rate control? We emphasize that our study is correlative, and that truly establishing causality on genome-scale patterns will require large-scale experiments that, for instance, systematically alter the SD sequences for hundreds of different genes at the same time and measure the resulting growth rates. Prior research has shown detrimental growth effects from deleting rRNA operons and tRNA genes in different microbial species (Stevenson and Schmidt 2004; Yano et al. 2013; Bloom-Ackermann et al. 2014; Samhita et al. 2014). Additionally, genome-scale engineering efforts to alter codon usage bias patterns have likewise showed increasing doubling times in these strains (Napolitano et al. 2016; Ostrov et al. 2016). Rather than suggest a single limiting feature for bacterial growth control, these results highlight that a number of features are likely necessary but insufficient on their own to increase bacterial growth rates.

We do not wish to suggest that increasing the translation initiation rates for all genes within a genome would result in a faster growing species, but we do hypothesize that *decreasing* the translation initiation rates for a large number of randomly chosen genes is likely to produce substantially longer doubling times. Why our results showed the strongest correlations with minimum doubling times when assessing genome-wide SD sequence utilization rather than only looking at ribosomal protein coding genes is unknown. However, we speculate that under periods of rapid growth, hundreds of genes (that include ribosomal proteins) are likely required and capable of bottle-necking bacterial growth. Calculating SD sequence utilization on a subset of such genes may produce an even stronger relationship with minimum doubling times, but the identities of these genes are likely to be organism and condition-specific.

We found that measurements of SD sequence utilization outperform more commonly known associations such as the number of rRNA genes at predicting minimum doubling times ($R^2 = 0.11$ vs. 0.06, and increasing feature importance in the full model depicted in table 1) (Roller et al. 2016).

We believe that this finding may, in part, be a consequence of the relevant evolutionary time-scales that it takes to alter different traits. Specifically, substantially changing genome-wide SD sequence utilization or codon usage biases would require hundreds or thousands of fixed mutations, respectively. We found that these two traits were the most predictive of minimum doubling times out of all traits that we considered, which is in contrast to traits such as rRNA and tRNA gene counts that can be altered rapidly by deletions/insertions. Thus, the degree of codon usage bias or SD sequence utilization within a genome may better reflect long-term historical forces acting on a species, whereas copy number variations have the potential to better reflect more rapid evolutionary changes.

Finally, we note that—like codon usage biases and in contrast to rRNA and tRNA gene counts—summary statistics based on SD sequence utilization do not require complete genome sequences and therefore may be *estimated* with partial genome fragments. The results and methods that we present here may thus have important applications in our understanding of novel, uncultivated genomes, environmental meta-genomic sequencing efforts, and the relationship between higher order genome traits and growth strategies (Brown et al. 2016).

## Materials and Methods

### Data Assembly

We first assembled a database of prokaryotic genomes from NCBI using the GBProks software (https://github.com/hyattpd/gbproks), including only "complete" genomes in our download and subsequent analysis (accessed on: March 10, 2016). From the annotated GenBank files, we excluded pseudogenes and plasmid based sequences from all subsequent analyses and proceeded to compile a data table with several traits for each genome. In addition to SD sequence utilization summary statistics described below, we applied RNAmmer to each genome in order to compile a list of ribosomal-RNA genes, and tRNAscan-SE to assemble a list of the tRNA genes (Lagesen et al. 2007; Lowe and Chan 2016).

We wrote custom scripts to calculate the fraction of annotated coding sequences that begin with "ATG," as well as the metric of codon usage bias ($\Delta ENC$ as described in Vieira-Silva et al. 2010). For this latter metric, we first parsed the gene annotations to find ribosomal protein coding genes. We next computed the relative differences in codon usage bias between ribosomal protein coding genes and the rest of the genome, whereby:

$$\Delta ENC = \frac{ENC_{all} - ENC_{ribo}}{ENC_{all}} \tag{4}$$

where "all" and "ribo" refer to all protein coding genes and ribosomal protein coding genes, respectively. We altered the method used to calculate the "effective number of codons" or "ENC" from the one originally used by Vieira-Silva et al. (2010) to better control for GC content differences according to recent metric developed in our lab (manuscript submitted). The interpretation is the same, with larger positive values indicating more codon usage bias in ribosomal protein coding genes relative to the rest of the genome.

We calculated the metric of internal SD-like sequence occurrences as reported in in Yang et al. (2016). Briefly, we first calculate the average aSD sequence binding strength for each hexamer within a given coding sequence. As above, the final metric is then calculated as the difference in average aSD binding energies between all protein coding genes and ribosomal protein genes divided by the value for all genes. Larger value indicate fewer SD-like sequences are present in ribosomal protein coding genes relative to all genes.

Our measurement of genome-wide "mRNA folding," is calculated as follows. For each gene greater than 150nts in length, we extract a 60-nt long region centered on the start codon ($-30$ to $+30$) and calculate the minimum free energy of this segment under ViennaRNA defaults. We next do the same thing for an internal 60-nt segment (which we chose to standardize as nucleotides $+90$ to $+150$ for all genes). For each gene, we calculate the difference of these two folding energies and for each genome, we calculate the average of these differences across all genes. Negative values indicate that regions surrounding the start codon are more structured than internal regions, and as values get more positive the region surrounding the start codon is increasingly less structured compared with internal sites.

For data on minimum doubling times, we downloaded the data table from Vieira-Silva et al. (2010), and paired each bacterial species with a complete genome from our database resulting in 187 matched species. To control for shared ancestry in subsequent analyses, we constructed a phylogenetic tree based off the rRNA sequences for this set of species. We first used RNAmmer to extract a randomly chosen 16S and 23S rRNA sequence from each genome, followed by MUSCLE (v3.8.31) on each individual rRNA to produce a multiple-sequence alignments (Edgar 2004). These were concatenated together and we conducted a partitioned analysis using RAxML to construct a final tree. We performed 100 rapid Bootstrap searches, 20 ML searches and selected the best ML tree for subsequent analysis (Stamatakis 2014).

For the larger data set, we instead relied on a previously computed high-quality phylogenetic tree published by Hug et al. (2016) (Hug et al. 2016). We used custom scripts to match entries in this tree with genomes from our complete-genome database, and pruned away all species without a high-quality match resulting in 613 bacterial species in our final data set that were used for subsequent analyses. For temperature annotations, we matched this set of 613 species to the ProTraits database using custom scripts, and restricted our analysis to species with temperature annotations exceeding a precision of 0.9 (equivalent to a FDR $< 0.1$) (Brbić et al. 2016).

### Calculating Summary Statistics of SD Sequence Utilization

The calculation of $\Delta I$ is illustrated mathematically in the main text. Here, we only add that the calculation of the randomized sequences for all SD summary statistics is performed by first shuffling the upstream region of each gene between the region −30 to 0 (where +1 is the first base of the start codon). Having shuffled each gene in this manner, we then performed the analysis as discussed in the main text for this shuffled "genome" and repeat this calculation 500 times in order to derive null expectation for $f_{SD}$, $f_{aSD<-4.5}$ and $I_{obs}$.

Next, we elaborate on our calculation of the other two methods for calculating SD sequence utilization. For each genome, we extract the −20 to −4 region upstream of the start codon for each gene. For $f_{SD}$, we consider a gene as being SD-led if, in this defined region, any of the following motifs appear: "GGAA," "GGAG," "GAGG," "AGGA," or "AAGG." We repeat this same process for 500 randomized "genomes" where a randomized genome is defined as noted above (with the nucleotide region from −30 to 0 for each gene shuffled on a per-gene basis) prior to motif search.

For $f_{aSD<-4.5}$, we perform a nearly identical procedure to the one listed above with the major difference being that instead of searching the upstream region of genes for particular motifs, we evaluate the hybridization energy between each eight nucleotide segment contained within the −20 to −4 region and the putative aSD sequence defined as 5′-ACCUCCUU-3′ using the "cofold" method of the ViennaRNA software package with default parameters. If any sequence binds at a threshold of −4.5 kcal/mol or stronger (i.e., more negative $\Delta G$ values), we consider this gene to be SD-led.

### Phylogenetically Generalized Least Squares

Throughout this manuscript, we utilize Phylogenetically Generalized Least Squares regression in order to mitigate the effects that arise from shared ancestry in statistical analyses. Our Phylogenetically Generalized Least Squares analysis relies on the most common null model, which assumes a Brownian motion model of trait evolution. For all statistical analyses presented in the paper, we use the R package "caper" and perform a simultaneous maximum-likelihood estimate of Pagel's $\lambda$, a branch length transformation, alongside the coefficients for independent variables of interest. All *P* values that we report come from the *F*-test according to these results. For temperature analysis, we assigned "mesophiles" and "thermophiles" a value of 0 and 1, respectively, and performed the equivalent fixed-effect analysis with $\Delta I$ as the dependent variable.

### Data Availability and Computer Code

Data are provided as a supplementary file, Supplementary Material online and all custom scripts and code that is sufficient to perform the analysis can be found at https://github.com/adamhockenberry/SD-evolution-publication.

## Supplementary Material

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

## Acknowledgments

The authors wish to thank Thomas Stoeger for helpful discussions and critical reading of the manuscript, and Helio Tejedor for general computational support. AJH acknowledges financial support from the Northwestern University Presidential Fellowship. LANA acknowledges a gift from Leslie and John Mac McQuown. MCJ acknowledges the Army Research Office (W911NF-16-1-0372), the National Science Foundation (MCB 1413563, MCB-1716766), the David and Lucile Packard Foundation, and the Camille-Dreyfus Teacher-Scholar Program.

## Author Contributions

A.J.H., M.C.J., and L.A.N.A. conceived and designed the study. A.J.H. collected the data and performed analysis. A.J.S. contributed important preliminary results. A.J.H., M.C.J., and L.A.N.A. provided interpretation, and wrote the manuscript.

## Materials and Correspondence

Correspondence and materials request should be addressed to adam.hockenberry@utexas.edu.

## References

Barendt PA, Shah NA, Barendt GA, Kothari PA, Sarkar CA. 2013. Evidence for context-dependent complementarity of non-Shine-Dalgarno ribosome binding sites to *Escherichia coli* rRNA. *ACS Chem Biol*. 8(5):958–966.

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.

Bloom-Ackermann Z, Navon S, Gingold H, Towers R, Pilpel Y, Dahan O, Copenhaver GP. 2014. A comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool. *PLoS Genet*. 10(1):e1004084.

Bonde MT, Pedersen M, Klausen MS, Jensen SI, Wulff T, Harrison S, Nielsen AT, Herrgård MJ, Sommer MOA. 2016. Predictable tuning of protein expression in bacteria. *Nat Methods* 13(3):2230–2226.

Brbić M, Piškorec M, Vidulin V, Kriško A, Šmuc T, Supek F. 2016. The landscape of microbial phenotypic traits and associated genes. *Nucleic Acids Res*. 44(21):10074–10090.

Brown CT, Olm MR, Thomas BC, Banfield JF. 2016. Measurement of bacterial replication rates in microbial communities. *Nat Biotechnol*. 34(12):057992.

Chang B, Halgamuge S, Tang SL. 2006. Analysis of SD sequences in completed microbial genomes: non-SD-led genes are as common as SD-led genes. *Gene* 373(1–2):90–99.

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.

Colussi TM, Costantino D. a, Zhu J, Donohue JP, Korostelev A. a, Jaafar Z. a, Plank T-d. M, Noller HF, Kieft JS. 2015. Initiation of translation in bacteria by a structured eukaryotic IRES RNA. *Nature* 519(7541):110–113.

Cortes T, Schubert OT, Rose G, Arnvig KB, Comas I, Aebersold R, Young DB. 2013. Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*. *Cell Rep*. 5(4):1121–1131.

de Smit MH, van Duin J. 1994. Translation initiation on structured messengers: another role for the Shine-Dalgarno interaction. *J Mol Biol*. 235(1):173–184.

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.

Duval M, Korepanov A, Fuchsbauer O, Fechter P, Haller A, Fabbretti A, Choulier L, Micura R, Klaholz BP, Romby P, et al. 2013. *Escherichia coli* ribosomal protein S1 unfolds structured mRNAs onto the ribosome for active translation initiation. *PLoS Biol*. 11(12):e1001731.

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

Espah Borujeni A, Cetnar D, Farasat I, Smith A, Lundgren N, Salis HM. 2016. Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. *Nucleic Acids Res*. 45(9):5437–5448.

Espah Borujeni A, Channarasappa AS, Salis HM. 2014. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res*. 42(4):2646–2659.

Espah Borujeni A, Salis HM. 2016. Translation initiation is controlled by RNA folding kinetics via a ribosome drafting mechanism. *J Am Chem Soc*. 138(22):7016–7023.

Goodman DB, Church GM, Kosuri S. 2013. Causes and effects of N-terminal codon bias in bacterial genes. *Science* 342(6157):475–479.

Gu W, Zhou T, Wilke CO. 2010. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol*. 6(2):e1000664.

Guiziou S, Sauveplane V, Chang H-J, Cler E,C, Declerck N, Jules M, Bonnet J. 2016. A part toolbox to tune genetic expression in *Bacillus subtilis*. *Nucleic Acids Res*. 44(10):7495–7508.

Hecht A, Glasgow J, Jaschke PR, Bawazer LA, Munson MS, Cochran JR, Endy D, Salit M. 2017. Measurements of translation initiation from all 64 codons in *E. coli*. *Nucleic Acids Res*. 45(7):3615–3626.

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

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, et al. 2016. A new view of the tree of life. *Nat Microbiol*. 1(5):16048.

Keller TE, Mis SD, Jia KE, Wilke CO. 2012. Reduced mRNA secondary-structure stability near the start codon indicates functional genes in prokaryotes. *Genome Biol Evol*. 4(2):80–88.

Komarova AV, Tchufistova LS, Dreyfus M, Boni IV. 2005. AU-rich sequences within 5′ untranslated leaders enhance translation and stabilize mRNA in *Escherichia coli*. *J Bacteriol*. 187(4):1344–1349.

Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, Arkin AP, Endy D, Church GM. 2013. Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc Natl Acad Sci U S A*. 110(34):14024–14029.

Kramer P, Gäbel K, Pfeiffer F, Soppa J, de Crécy-Lagard V. 2014. Haloferax volcanii, a prokaryotic species that does not use the Shine Dalgarno mechanism for translation initiation at 5′-UTRs. *PLoS One* 9(4):e94979.

Krisko A, Copic T, Gabaldón T, Lehner B, Supek F. 2014. Inferring gene function from evolutionary change in signatures of translation efficiency. *Genome Biol*. 15(3):R44.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324(5924):255–258.

Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*. 35(9):3100–3108.

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

Lim K, Furuta Y, Kobayashi I. 2012. Large variations in bacterial ribosomal RNA genes. *Mol Biol Evol*. 29(10):2937–2948.

Lowe TM, Chan PP. 2016. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res*. 44(W1):W54–W57.

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.

Markley AL, Begemann MB, Clarke RE, Gordon GC, Pfleger BF. 2015. Synthetic biology toolbox for controlling gene expression in the Cyanobacterium *Synechococcus* sp. strain PCC 7002. *ACS Synth Biol*. 4(5):595–603.

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.

Na D, Lee S, Lee D. 2010. Mathematical modeling of translation initiation for the estimation of its efficiency to computationally design mRNA sequences with desired expression levels in prokaryotes. *BMC Syst Biol*. 4(1):71.

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.

Napolitano MG, Landon M, Gregg CJ, Lajoie MJ, Govindarajan L, Mosberg JA, Kuznetsov G, Goodman DB, Vargas-Rodriguez O, Isaacs FJ, et al. 2016. Emergent rules for codon choice elucidated by editing rare arginine codons in *Escherichia coli*. *Proc Natl Acad Sci U S A*. 113(38):E5588–E5597. pages

Omotajo D, Tate T, Cho H, Choudhary M. 2015. Distribution and diversity of ribosome binding sites in prokaryotic genomes. *BMC Genomics* 16(1):604.

Orelle C, Carlson ED, Szal T, Florin T, Jewett MC, Mankin AS. 2015. Protein synthesis by ribosomes with tethered subunits. *Nature* 524(7563):119–124.

Osada Y, Saito R, Tomita M. 1999. Analysis of base-pairing potentials between 16S rRNA and 5′ UTR for translation initiation in various prokaryotes. *Bioinformatics* 15(1996):578–581.

Ostrov N, Landon M, Guell M, Kuznetsov G, Teramoto J, Cervantes N, Zhou M, Singh K, Napolitano MG, Moosburner M, et al. 2016. Design, synthesis, and testing toward a 57-codon genome. *Science* 353(6301):819–822.

Revell LJ. 2010. Phylogenetic signal and linear regression on species data. *Methods Ecol Evol*. 1(4):319–329.

Roller BRK, Stoddard SF, Schmidt TM. 2016. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. *Nat Microbiol*. 1(11):16160.

Sakai H, Imamura C, Osada Y, Saito R, Washio T, Tomita M. 2001. Correlation between Shine-Dalgarno sequence conservation and codon usage of bacterial genes. *J Mol Evol*. 52(2):164–170.

Salis HM, Mirsky EA, Voigt CA. 2009. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol*. 27(10):946–950.

Samhita L, Nanjundiah V, Varshney U. 2014. How many initiator tRNA genes does *Escherichia coli* need? *J Bacteriol*. 196(14):2607–2615.

Scharff LB, Childs L, Walther D, Bock R, Casadesús J. 2011. Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. *PLoS Genet*. 7(6):e1002155.

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

Shell SS, Wang J, Lapierre P, Mir M, Chase MR, Pyle MM, Gawande R, Ahmad R, Sarracino DA, Ioerger TR, et al. 2015. Leaderless transcripts and small proteins are common features of the Mycobacterial translational landscape. *PLoS Genet*. 11(11):1–31.

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.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Starmer J, Stomp A, Vouk M, Bitzer D. 2006. Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLoS Comput Biol*. 2(5):454–466.

Stevenson BS, Schmidt TM. 2004. Life history implications of rRNA gene copy number in *Escherichia coli*. *Appl Environ Microbiol*. 70(11):6670–6677.

Tauer C, Heinl S, Egger E, Heiss S, Grabherr R. 2014. Tuning constitutive recombinant gene expression in *Lactobacillus plantarum*. *Microbial Cell Factories* 13(1):150.

Vieira-Silva S, Rocha EPC, Moran NA. 2010. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* 6(1):e1000808.

Vimberg V, Tats A, Remm M, Tenson T. 2007. Translation initiation region sequence preferences in *Escherichia coli*. *BMC Mol Biol.* 8(1):100.

Weinstock MT, Hesek ED, Wilson CM, Gibson DG. 2016. *Vibrio natriegens* as a fast-growing host for molecular biology. *Nat Methods* 13(10):1–39.

Yamamoto H, Wittek D, Gupta R, Qin B, Ueda T, Krause R, Yamamoto K, Albrecht R, Pech M, Nierhaus KH. 2016. 70S-scanning initiation is a novel and frequent initiation mode of ribosomal translation in bacteria. *Proc Natl Acad Sci U S A.* 113(9):E1180–E1189.

Yang C, Hockenberry AJ, Jewett MC, Amaral LAN. 2016. Depletion of Shine-Dalgarno sequences within bacterial coding regions is expression dependent. *G3 (Bethesda)* 6(November):3467–3474.

Yano K, Wada T, Suzuki S, Tagami K, Matsumoto T, Shiwa Y, Ishige T, Kawaguchi Y, Masuda K, Akanuma G, et al. 2013. Multiple rRNA operons are essential for efficient cell growth and sporulation as well as outgrowth in *Bacillus subtilis*. *Microbiology* 159(Pt_11):2225–2236.

Yi JS, Kim MW, Kim M, Jeong Y, Kim E-J, Cho B-K, Kim B-G. 2016. A novel approach for gene expression optimization through native promoter and 5' UTR combinations based on RNA-seq, Ribo-seq, and TSS-seq of *Streptomyces coelicolor*. *ACS Synth Biol.* 6(3):555–565.

Zheng X, Hu G-Q, She Z-S, Zhu H. 2011. Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. *BMC Genomics* 12(1):361.