SCIENTIFIC REPORTS

OPEN

Received: 18 August 2017
Accepted: 4 December 2017
Published online: 15 December 2017

Elucidating the 16S rRNA 3′ boundaries and defining optimal SD/aSD pairing in *Escherichia coli* and *Bacillus subtilis* using RNA-Seq data

Yulong Wei[1], Jordan R. Silke[1] & Xuhua Xia[1,2]

Bacterial translation initiation is influenced by base pairing between the Shine-Dalgarno (SD) sequence in the 5′ UTR of mRNA and the anti-SD (aSD) sequence at the free 3′ end of the 16S rRNA (3′ TAIL) due to: 1) the SD/aSD sequence binding location and 2) SD/aSD binding affinity. In order to understand what makes an SD/aSD interaction optimal, we must define: 1) terminus of the 3′ TAIL and 2) extent of the core aSD sequence within the 3′ TAIL. Our approach to characterize these components in *Escherichia coli* and *Bacillus subtilis* involves 1) mapping the 3′ boundary of the mature 16S rRNA using high-throughput RNA sequencing (RNA-Seq), and 2) identifying the segment within the 3′ TAIL that is strongly preferred in SD/aSD pairing. Using RNA-Seq data, we resolve previous discrepancies in the reported 3′ TAIL in *B. subtilis* and recovered the established 3′ TAIL in *E. coli*. Furthermore, we extend previous studies to suggest that both highly and lowly expressed genes favor SD sequences with intermediate binding affinity, but this trend is exclusive to SD sequences that complement the core aSD sequences defined herein.

Protein production is a highly controlled and optimized process in bacterial species[1], and translation initiation is often recognized as the rate-limiting step of the translation process[2–4]. As such, finding ways to overcome this bottleneck in efficiency is important for using bacteria in transgenic biosynthesis of important pharmaceutical compounds such as insulin[5]. Translation initiation efficiency in bacteria is strongly influenced by the binding affinity between the Shine-Dalgarno (SD) sequence upstream of the start codon on mRNA and the anti-SD (aSD) sequence located at the free 3′ end of the 16S rRNA (3′ TAIL)[6,7]. Furthermore, the location of the SD/aSD interaction relative to the start codon must also be considered to ensure that the pairing positions the ribosomal P-site at the start codon[6–9].

A recent model of SD/aSD interaction[10,11] (Fig. 1) suggests that optimal SD/aSD pairing may depend on three factors: 1) $D_{toStart}$ (Fig. 1) which specifies the distance, in nucleotides, between the 16S rRNA 3′ terminus and the start codon, 2) SD/aSD binding affinity (Figs 1b), and 3) "leash" distance measured by D_1 and D_2 (Fig. 1). $D_{toStart}$ is strongly constrained within a narrow range. Intra-strand secondary structure that embeds the SD sequence is also known to affect SD/aSD function in localizing translation initiation codon[11,12]. Characterizing these features demands the precise terminus of the 16S rRNA which is often unclear, as is the case for *Bacillus subtilis*.

RNA-seq data as a novel approach to define the 3′ TAIL in *E. coli* and *B. subtilis*. The 3′ TAIL was previously reported to be 5′-CCUCCUUUCU-3′[13] based on personal communication between the authors and Carl Woese, although no explicit data to substantiate the terminus of the 3′ TAIL in *B. subtilis* was published. Acceptance of the 5′-CCUCCUUUCU-3′ end[13,14] arose because Woese and colleagues published the details of their RNA sequencing method[15] as well as the 3′ TAILs in a number of bacterial species[16]. Since then, alternative

[1]Department of Biology, University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A, Ottawa, Ontario, Canada. [2]Ottawa Institute of Systems Biology, Ottawa, Ontario, K1H 8M5, Canada. Yulong Wei and Jordan R. Silke contributed equally to this work. Correspondence and requests for materials should be addressed to X.X. (email: xxia@uottawa.ca)

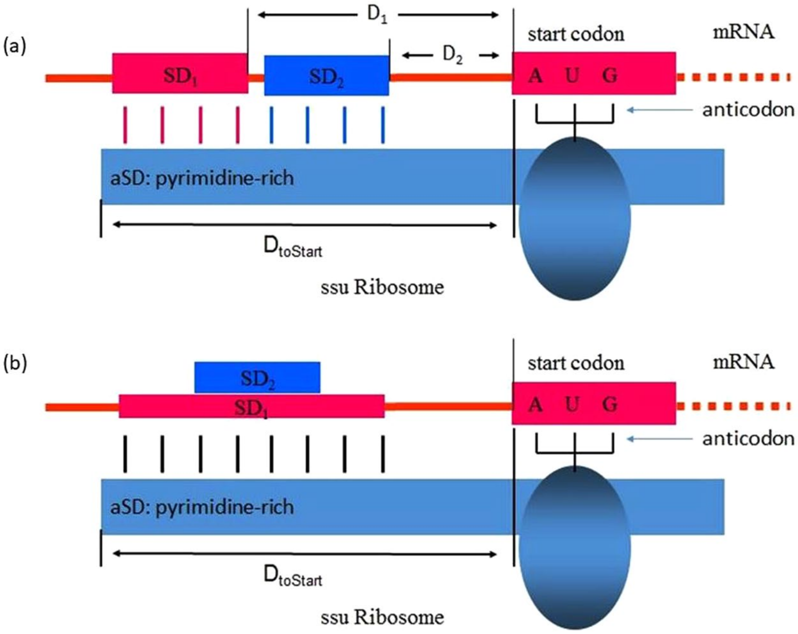


Figure 1. Schematic model of SD/aSD interaction, illustrating $D_{toStart}$ (**a** and **b**), difference in the two "leash" distances (D1 and D2) and in binding affinity (**b**) between two SD/aSD interactions involving SD_1 and SD_2.

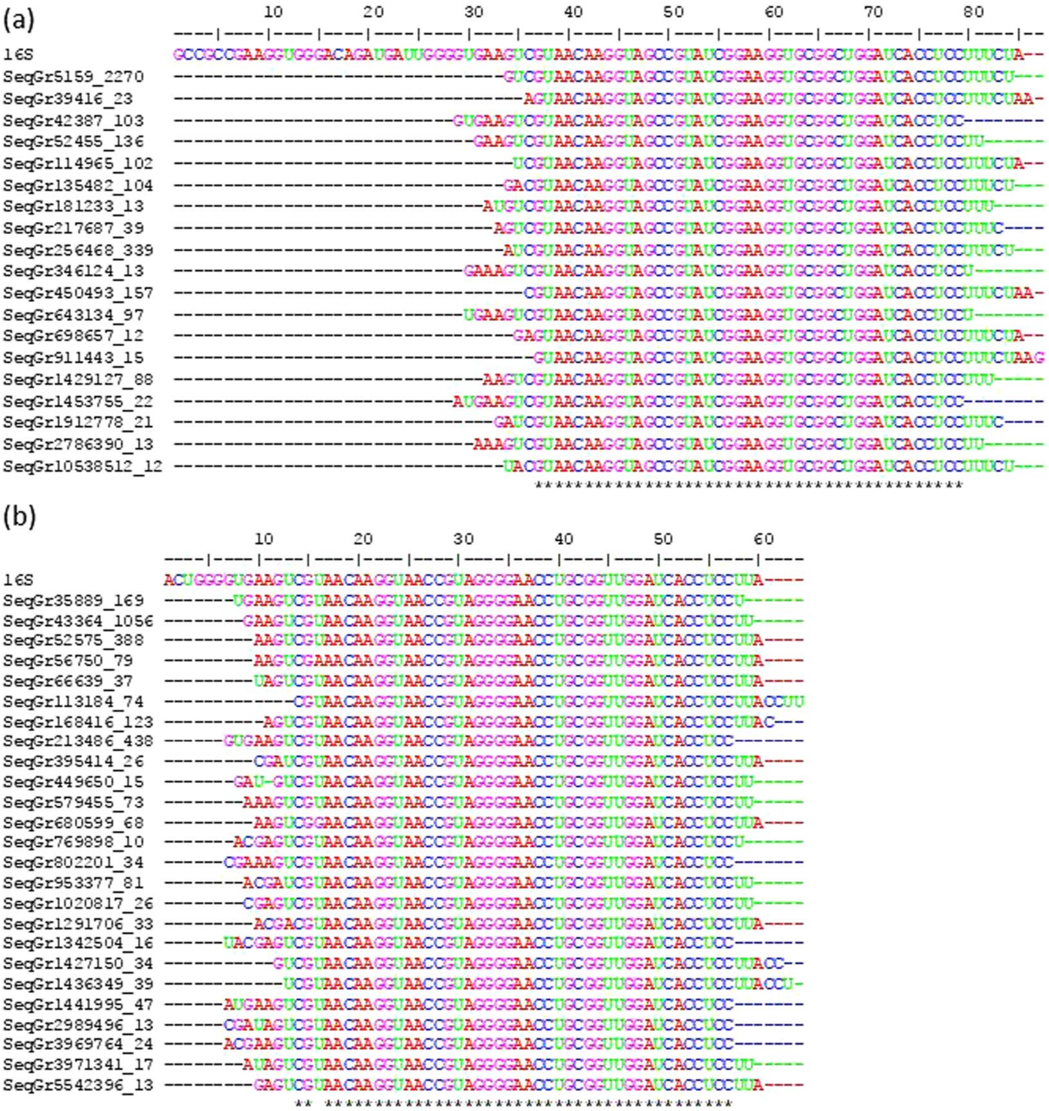
rDNA annotations of the *B. subtilis* 3′ TAIL have emerged, including 5′-CCUCCUUUCUA-3′ (NC_000964)[17] and 5′-CCUCCUUUCUAA-3′ (NZ_CP010052) which have been used in recent studies on *B. subtilis* 16S rRNA[9,18]. Discrepancies in these reported 3′ TAILs likely arose due to the fact that multiple exoribonucleases participate in the maturation process of the 3′ TAIL[19]. These include *RNase II*, *RNase R*, *PNPase* and *RNase PH*[20], as well as *YbeY*[21]; hence, the 3′ TAIL is continuously degraded.

Resolving the terminus of the mature 3′ TAIL in *B. subtilis* is the first objective of our study. To this end, we employ high-throughput RNA sequencing (RNA-seq) data. Recent advances in RNA-Seq technologies[22–24] offer a novel way to identify the 3′ TAIL in the cell by mapping millions of short RNA reads onto the annotated sequence. However, one issue with using RNA-Seq data to analyze the 3′ TAIL is that rRNAs are often removed in the experiments with the use of kits such as RiboMinus from Invitrogen or Ribo-Zero from Epicenter. To circumvent this challenge, we employ publically available datasets for *E. coli* and *B. subtilis* that have not undergone ribo-depletion. We predict that our findings will corroborate the mature 3′ TAIL previously reported[13]. To ensure the fidelity of our method, we analyze *E. coli* data from the same experiment with the expectation of recovering the widely accepted 5′-GAUCACCUCCUUA-3′ reported before[6].

Determining the non-volatile 3′ end of mature 16S rRNA is crucial to establish 1) correct and meaningful $D_{toStart}$ positioning of the SD/aSD interaction and 2) which nucleotides should be considered when determining the complement SD sequences. Achieving these goals will lead to our second objective: to assess the effects of SD/aSD binding affinity on initiation efficiency while controlling for the optimal $D_{toStart}$ range.

Determining the optimal SD/aSD interaction that maximizes initiation efficiency. It was generally believed that high SD/aSD binding affinity facilitated translation initiation[25–28]; accordingly, the core aSD motif (CCUCC) was characterized based on its high binding affinity (most negative change in Gibbs free energy [ΔG]). Furthermore, CCUCC is conserved in 277 prokaryotic species using multiple sequence alignment in MAFFT[29]. In practice, putative SD sequences are determined based on their complementarity with an extended sequence at the 3′ TAIL[8–11,30,31]: the inclusion of the core motif CCUCC is canonical, but what constitutes the full extent of the core aSD sequence remains unclear[9].

The set of identified SD sequences varies depending on the choice of the aSD sequence. A poor set of SD sequences will not provide much insight on initiation efficiency. For example, a recent study[1] uses 5′-CACCUCC-3′ as the *E. coli* aSD sequence to find putative SD sequences, but observes no correlation between SD binding affinity and translation efficiency. This finding leads to the surprising conclusion that SD/aSD pairing potential has little predictive power over gene expression[30]. A similar study[9] uses extended aSD sequences (e.g. 5′-ACCUCCUUA-3′ in *E. coli*), and found that intermediate levels of SD/aSD binding maximize translation efficiency, not high binding affinities. This discovery corroborates previous reports[8,32] showing that SD sequences with intermediate levels of aSD (5′-ACCUCCUU-3′) binding occur most frequently in *E. coli* genes[8] and that six SD/aSD base pairs lead to more efficient translation and growth than shorter or longer SD/aSD pairs[32]. Taken together, these studies suggest that intermediate levels of SD/aSD binding facilitate the recruitment of the ribosome to the mRNA, but high SD/aSD binding inhibits the transition from initiation to elongation leading to ribosome stalling.



```
(a)
                         10        20        30        40        50        60        70        80
                 ----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|--
16S              GCCGCCGAAGGUGGGACAGAUGAUUGGGGUGAAGUCGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUUUCUA--
SeqGr5159_2270   ---------------------------------GUCGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUUUCU---
SeqGr39416_23    -----------------------------------AGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUUUCUAA-
SeqGr42387_103   ----------------------------GUGAAGUCGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCC--------
SeqGr52455_136   ------------------------------GAAGUCGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUU------
SeqGr114965_102  ----------------------------------UCGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUUUCUA--
SeqGr135482_104  ---------------------------------GACGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUUUCU---
SeqGr181233_13   -------------------------------AUGUCGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUUU-----
SeqGr217687_39   --------------------------------AGUCGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUUUC----
SeqGr256468_339  ---------------------------------AUCGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUUUCU---
SeqGr346124_13   -----------------------------GAAAGUCGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCU-------
SeqGr450493_157  -----------------------------------CGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUUUCUAA-
SeqGr643134_97   -----------------------------UGAAGUCGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCU-------
SeqGr698657_12   ----------------------------------GAGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUUUCUA--
SeqGr911443_15   ------------------------------------GUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUUUCUAAG
SeqGr1429127_88  -------------------------------AAGUCGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUUU-----
SeqGr1453755_22  ----------------------------AUGAAGUCGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCC--------
SeqGr1912778_21  --------------------------------GAUCGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUUUC----
SeqGr2786390_13  ------------------------------AAAGUCGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUU------
SeqGr10538512_12 ---------------------------------UACGUAACAAGGUAGCCGUAUCGGAAGGUGCGGCUGGAUCACCUCCUUUCU---
                                                     *******************************************

(b)
                         10        20        30        40        50        60
                 ----|----|----|----|----|----|----|----|----|----|----|----|----
16S              ACUGGGGUGAAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUUA----
SeqGr35889_169   -------UGAAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCU------
SeqGr43364_1056  --------GAAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUU-----
SeqGr52575_388   ---------AAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUUA----
SeqGr56750_79    ---------AAGUCGAAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUUA----
SeqGr66639_37    ---------UAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUUA----
SeqGr113184_74   -------------CGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUUACCUU
SeqGr168416_123  ----------AGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUUAC---
SeqGr213486_438  ------GUGAAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCC-------
SeqGr395414_26   ---------CGAUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUUA----
SeqGr449650_15   -------GAU-GUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUU-----
SeqGr579455_73   --------AAAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUU-----
SeqGr680599_68   ---------AAGUCGGAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUUA----
SeqGr769898_10   -------ACGAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCU------
SeqGr802201_34   ------CGAAAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCC-------
SeqGr953377_81   --------ACGAUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUU-----
SeqGr1020817_26  --------CGAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUU-----
SeqGr1291706_33  ---------ACGACGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUUA----
SeqGr1342504_16  ------UACGAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCC-------
SeqGr1427150_34  -----------GUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUUACC--
SeqGr1436349_39  ------------UCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUUACCU-
SeqGr1441995_47  ------AUGAAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCC-------
SeqGr2989496_13  ------CGAUAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCC-------
SeqGr3969764_24  ------ACGAAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCC-------
SeqGr3971341_17  --------AUAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUU-----
SeqGr5542396_13  ---------GAGUCGUAACAAGGUAACCGUAGGGGAACCUGCGGUUGGAUCACCUCCUUA----
                              ** *****************************************
```

Figure 2. Multiple sequence alignment of reads in FASTA+ format (with sequence ID in the form of 'ID_##' where'##' represents the number of reads that are identical to the represented fragment) matching the 3′ TAIL in (**a**) *B. subtilis* and (**b**) *E. coli*. The top sequence in each panel corresponds to the annotated 3′ TAIL rDNA reference used in BLAST searches from (**a**) NC_000964 and (**b**) NC_000913. Hits were only included in the alignment if they extended to or beyond the 3′ CCUCC motif without base calling errors, and had at least 10 identical matches (accounting for 97.5% of reads in *B. subtilis* and 94% of reads in *E. coli* that fit our criteria).

It remains controversial as to what constitutes the core aSD, i.e., the aSD embedded in 3′TAIL that is most frequently involved in functional SD/aSD interactions. We operationally define the core aSD as the sequence motif within 3′TAIL most frequently involved in SD/aSD interactions within optimal $D_{toStart}$ ranges. Although previous studies suggested CCUCC as the core aSD[25–28], the corroborative reasoning that CCUCC is conserved among bacterial species is a weak one, as 5′-GAUCACCUCCU-3′ is highly conserved among 249 bacterial species[29], not just CCUCC.

Results and Discussion

Elucidating the mature 16S rRNA 3′ tail using RNA-Seq data. We identify the 3′ TAIL in *E. coli* and *B. subtilis* using RNA-Seq data. To this end, we BLASTed *B. subtilis* single reads from RNA-Seq run SRR1232437 against 85 nt at the 3′ terminus of the annotated *B. subtilis* 16S rDNA sequence (Fig. 2a, entry labelled 16S, NC_000964). This procedure was also repeated for *E. coli* single reads (SRR1232430) with 60 nt at the 3′ terminus of annotated *E. coli* 16S rDNA sequence (Fig. 2b, entry labelled 16S, NC_000913). We then eliminated BLAST hits that did not extend to encompass the conserved core CCUCC motif of the 3′ TAIL. From the remaining hits, we generated a distribution that indicates the prevalence of 3′ termini (Fig. 2) in both species.

We expect to recover the universally accepted 3′ terminus reported for *E. coli*[6] and, at minimum, the 5′-CCUCCUUUCU-3′ end reported for *B. subtilis*[13]. In keeping with expectations, the data shows dominant

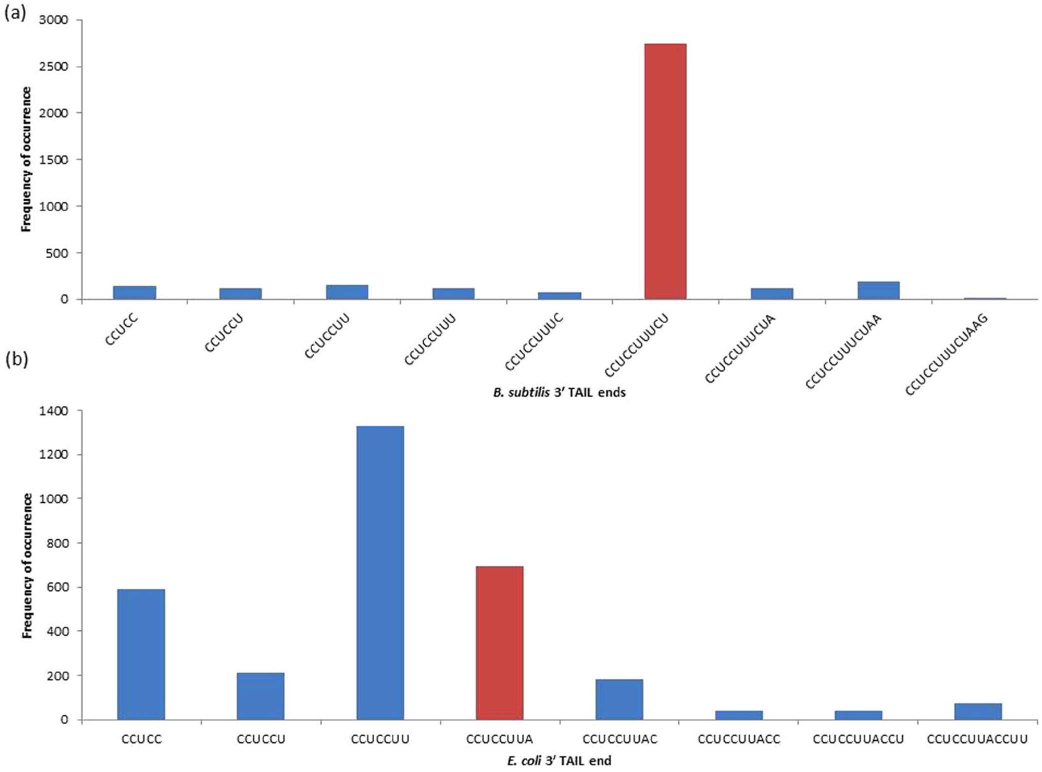


Figure 3. The distribution of hits corresponding to specific 16S rRNA 3′ ends in (**a**) *B. subtilis* and (**b**) *E. coli*. The frequencies of terminal nucleotides for each 3′ TAIL BLAST hit extending to or beyond CCUCC are depicted. Red bars represent the frequencies associated with the first reported 3′ ends in each species.

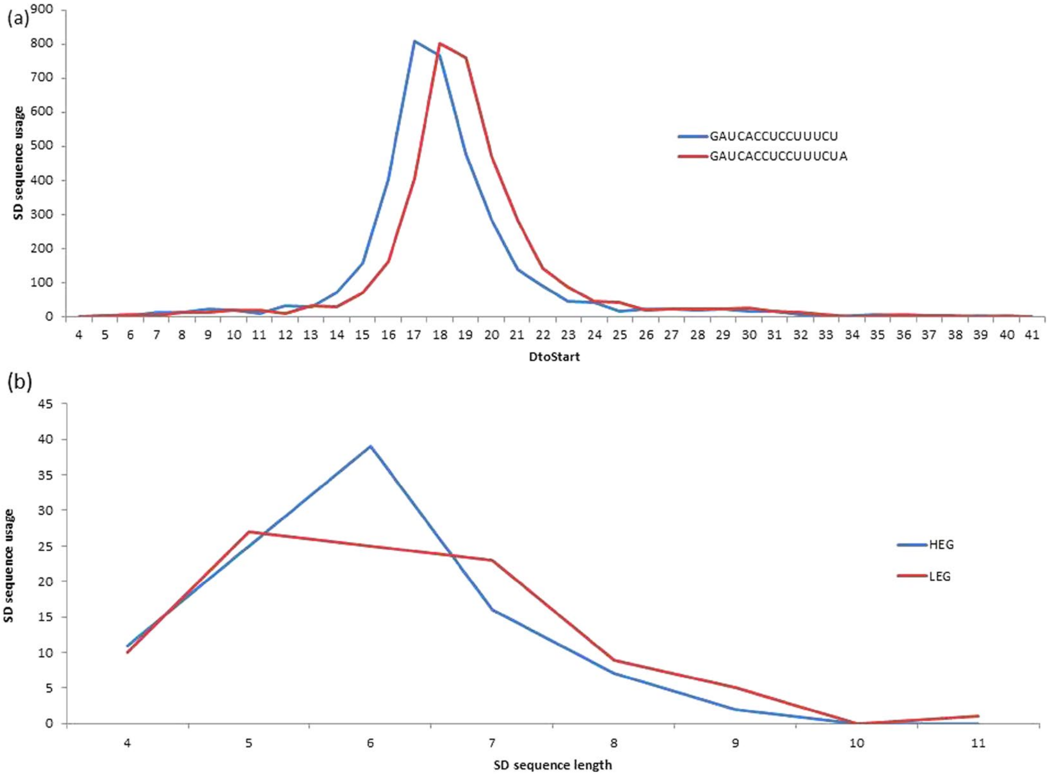


Figure 4. (**a**) D_{toStart} is constrained to a narrow range in all *B. subtilis* putative SD sequences, but the optimal range varies depending on the terminus of the 3′ TAIL. (**b**) Difference in motif length preference of SD sequences with $D_{\text{toStart}} = 17$ in HEGs and LEGs.

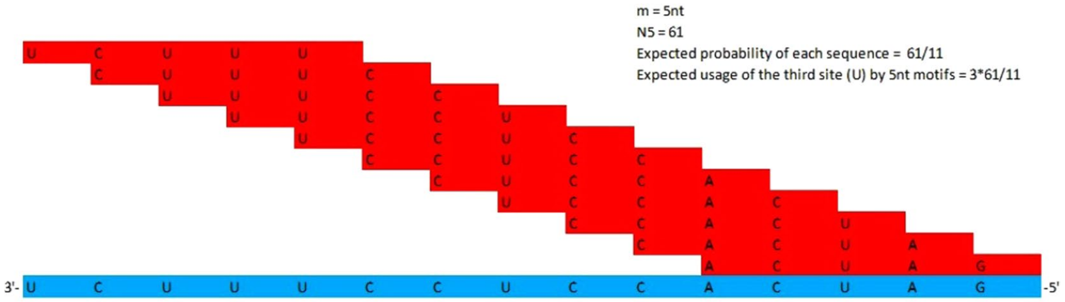

m = 5nt
N5 = 61
Expected probability of each sequence = 61/11
Expected usage of the third site (U) by 5nt motifs = 3*61/11
U C U U U
C U U U C
U U U C C
U U C C U
U C C U C
C C U C C
C U C C A
U C C A C
C C A C U
C A C U A
A C U A G
3'- U C U U U C C U C C A C U A G -5'

Figure 5. The matching scheme illustrating the expected site specific usage at each aSD site by 5 nt SD sequences (e.g. 61 observed 5 nt SD sequences). Each aSD site (blue) is equally likely to participate in SD/aSD binding with an individual SD sequence (red) assuming there are no site specific selection biases. The aSD site specific expected usage is location-dependent, varying based on displacements of 61 sequences.

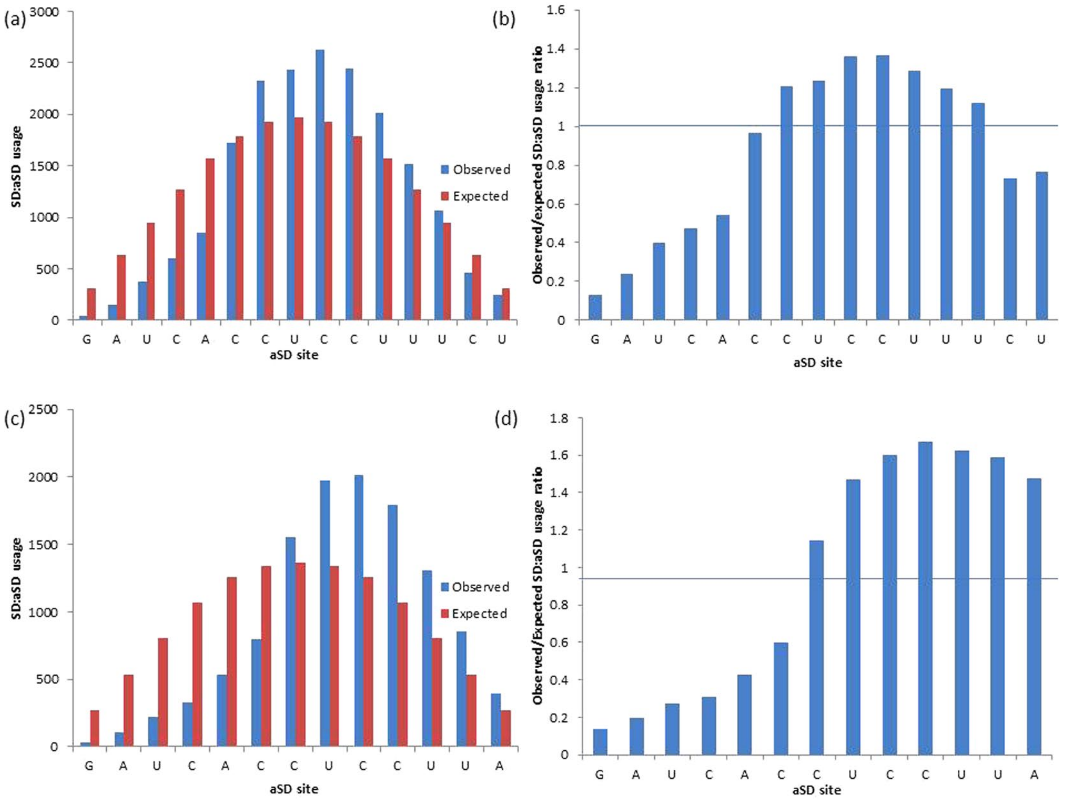


Figure 6. The observed and expected usages and observed/expected usage ratios of aSD sites in (**a** and **b**) *B. subtilis* (5′-GAUCACCUCCUUUCU-3′) and (**c** and **d**) *E. coli* (5′-GAUCACCUCCUUA-3′) by all putative SD sequences. Putative SD sequences (4 to 12 nt) are determined at optimal D_{toStart} locations (15 to 21 in *B. subtilis*, 10 to 21 in *E. coli*).

usage of the originally reported 5′-CCUCCUUUCU-3′ end in *B. subtilis* (Fig. 3a), and provides no basis for the inclusion of downstream nucleotides such as A[18] (NC_000964) or AA (NZ_CP010052) in the mature 3′ TAIL. In contrast, our data suggests characterization of the mature 3′ TAIL in *E. coli* may be less straightforward than previously reported[6]. Figure 3b presents three major 3′ TAIL termini, the longest of which is the widely accepted 5′-CCUCCUUA-3′. Unexpectedly, we also observe high frequencies of reads ending with CCUCC and 5′-CCUCCUU-3′, which suggests that there may be up to three distinct termini for the mature 3′ TAIL in *E. coli*. Importantly, we do recover the expected 3′ end, which indicates that our method works as intended. These observations show that RNA-Seq data is reasonably accurate and can be used to define rRNA termini in the absence of ribo-depletion. Moreover, we propose that the methodology which we apply herein to map the 3′ termini of 16S rRNAs can be extended not only to other species, but also to mapping the termini of other RNA molecules. The

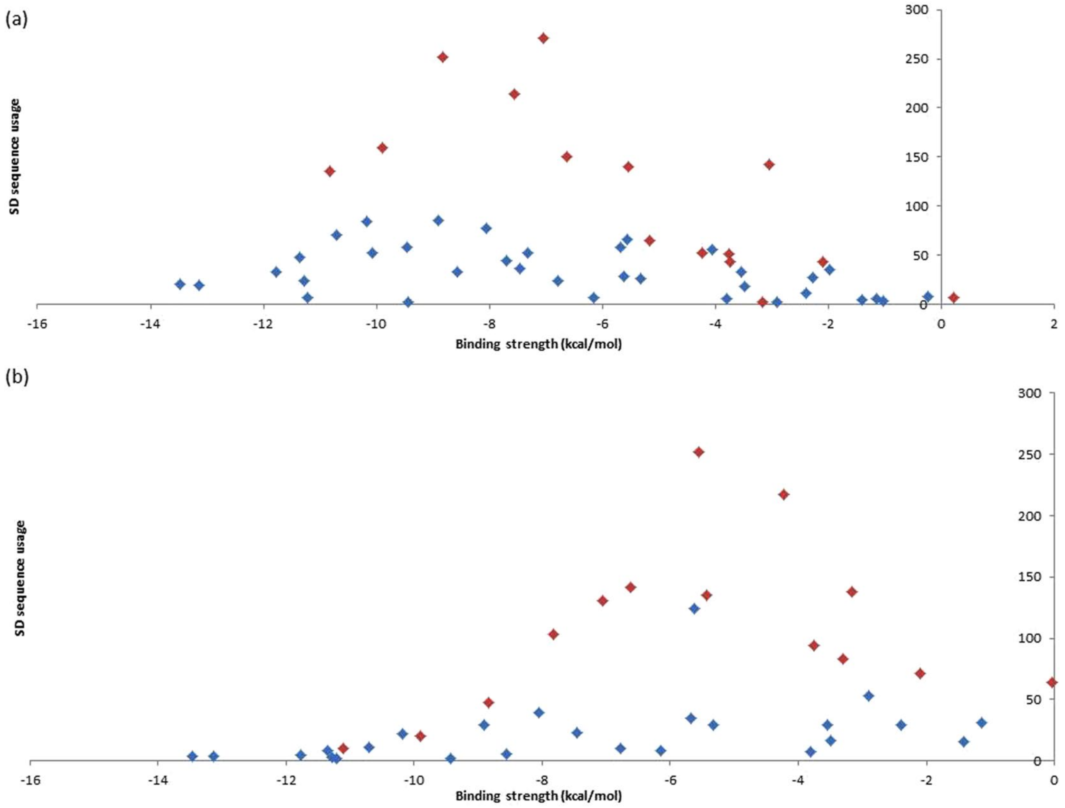


Figure 7. Usages of 4 to 8 nt putative SD sequences and their aSD binding affinity in (**a**) *B. subtilis* and (**b**) *E. coli*. SD sequences with complementarity to the extended core aSD sequences 5′-CCUCCUUU-3′ (*B. subtilis*) and 5′-CCUCCUUA-3′ (*E. coli*) are highlighted red, all other SD sequences are highlighted blue.

RNA-Seq data can be potentially used to characterize transcription start and termination sites as well, paving the way for accurate determination of operons.

It is worth mentioning that the quality of rRNA identification may vary depending on the RNA-Seq protocol used. For instance, when ribo-depletion is employed, although reads mapping to rRNAs may be recovered, the sequence quality is generally poor (Supplementary Fig. S1). Other factors affecting sequence quality include the average read length sequenced in the experiment (sequence quality tends to depreciate towards the end of longer reads), and whether single or paired-end reads are assessed.

Our characterization of the discrete terminus of the mature 3′ TAIL in *B. subtilis* emphasizes that the common practice of approximating the 16S rRNA terminus based on sequence similarity[29,33] is inadequate. The underlying issue surrounding these instances of poor annotation is the ease with which they are propagated in automated annotation[34,35]. Using such annotations may potentially skew conclusions in studies on translation initiation. For example, investigations considering the *B. subtilis* 3′ TAIL 5′-GAUCACCUCCUUUCUA-3′[1,8–10] will inherently include a subset of SD/aSD interactions that may detract from the clarity of existing patterns because there can be no translation-mediated selection affecting nucleotides that are absent at the RNA level (the 3′ A). This motivates us to reanalyze optimal SD and aSD sequences in *E. coli* and *B. subtilis* using 3′ TAILs determined by the RNA-Seq data herein.

The effect of SD/aSD pairing location on initiation efficiency. The mature 3′ TAIL in *B. subtilis* identified here (5′-GAUCACCUCCUUUCU-3′) requires that the optimal range for $D_{toStart}$ positions, described in a previous study as 15–25 using 5′-GAUCACCUCCUUUCUA-3′[10], to be redefined. In order to accomplish this, we determined all putative SD sequences between the lengths of 4 and 12 nt (see Materials and Methods for more detail) that complement the mature 3′ TAIL 5′-GAUCACCUCCUUUCU-3′ determined herein. We redefined the optimal range of $D_{toStart}$ distances as 15 to 21 nt in *B. subtilis* based on the optimal range shown in Fig. 4a. As for *E. coli*, the previously reported $D_{toStart}$ range of 10 to 21 nt[10] was preserved because the same mature 3′ TAIL (5′-GAUCACCUCCUUA-3′) was used.

To clearly highlight the effect of binding affinity on initiation efficiency and show that positioning alone is insufficient to determine optimal SD/aSD pairings, we examine *B. subtilis* putative SD sequences occurring at the most frequently observed distance (Fig. 4a; $D_{toStart} = 17$). We show a high preference for the usage of six nt motifs in highly expressed genes (HEGs), but not in lowly expressed genes (LEGs) (Fig. 4b). The SD/aSD pairing length is directly associated with binding affinity (longer sequences have higher binding affinity than short sequences), but this association alone is inadequate to capture the heterogeneity intrinsic to a given pair length. For instance, 5′-CCUUU-3′ and CCUCC are both five nt SD sequences that are complementary to the aSD in *B.*

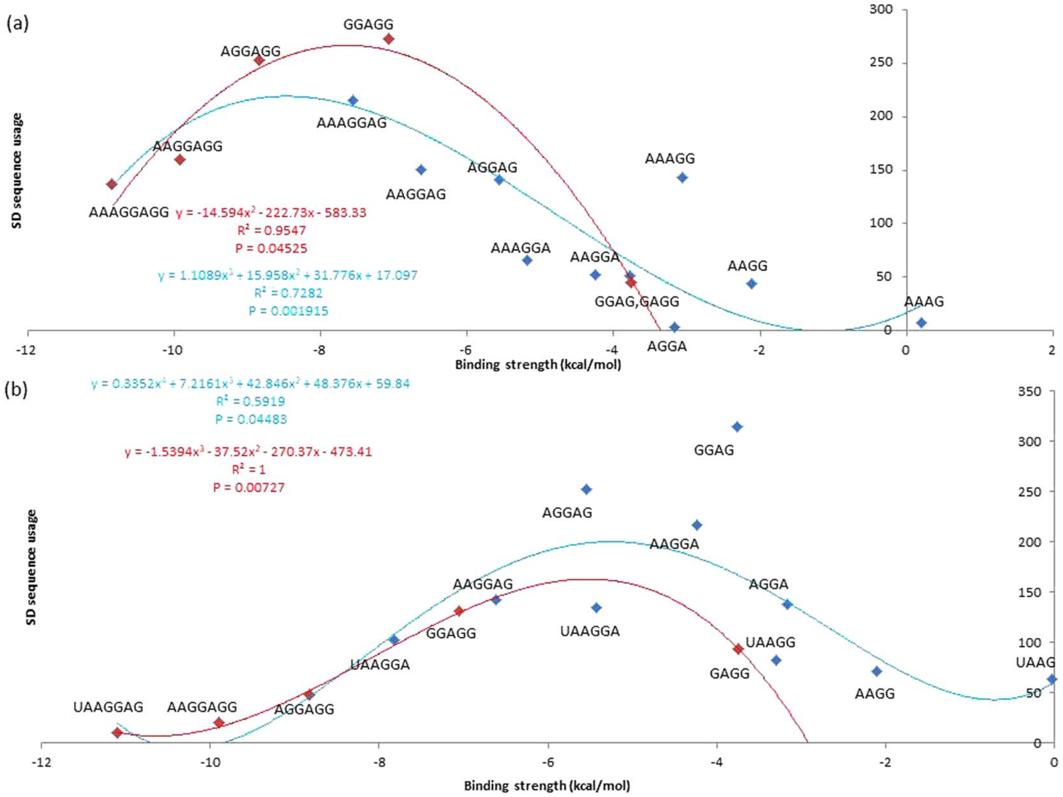


Figure 8. Relationship between the usages of 4 to 8 nt putative SD sequences and their aSD binding affinity in (**a**) *B. subtilis* and (**b**) *E. coli*. All SD sequences have complementarity with the aSD sequences 5′-CCUCCUUU-3′ (*B. subtilis*) and 5′-CCUCCUUA-3′ (*E. coli*). Highlighted in blue are SD sequences that complement only to the un-extended 5′-CUCCUUU-3′ (*B. subtilis*) and 5′-CUCCUUA-3′ (*B. subtilis*). Highlighted in red are SD sequences that were identified after the core aSD sequences were extended to encompass CCUCC.

subtilis; however, the binding affinity in the former is −3.04 kcal/mol while it is −7.05 kcal/mol in the latter (based on RNAcofold[36]). This implies that, despite having the same pairing length, SD/aSD pairings with CCUCC are substantially more stable than those with 5′-CCUUU-3′. It is for this reason that we explicitly consider binding affinity in the next section.

Determining the core aSD sequence based on SD/aSD pairing preference. To determine the extent of the core aSD sequence for both species, we examined the observed and expected usages for each site of the 3′ TAIL in base pairing with all putative SD sequences. To control for the influence of SD/aSD binding location, we only considered putative SD sequences that are located within optimal $D_{toStart}$ ranges discussed previously. For *B. subtilis*, we define bases within the 3′ TAIL 5′-GAUCACCUCCUUUCU-3′ as aSD sites. The expected aSD site usage is estimated assuming that a given SD sequence between four and 12 nt has an equal chance to pair with any given segment within the 3′ TAIL (Fig. 5; See Materials and Methods for more detail). Determining the observed and expected aSD site usages is an important step in examining SD sequence preference. Bases toward the middle of the aSD sequence are more predisposed to pairing with SD sequences than those towards the ends, as illustrated in Fig. 5. Since CCUCC constitutes the middle segment of the 3′ TAIL: 5′-GAUCACCUCCUUA-3′[6] and 5′-GAUCACCUCCUUUCU-3′[13] in *E. coli* and *B. subtilis*, respectively, it is unsurprising that the expected usage of this motif is the highest, as illustrated in Fig. 6. Consequently, one must contrast between observed and expected usages of aSD sites to determine their preference and avoidance of SD sequences. In this respect, the shortcoming of Osterman *et al*. (2013) is that they did not contrast the observed and expected SD sequence usages when contrasting sequence occurrences by binding affinity.

We characterize an aSD site to be favorably selected if it pairs with putative SD sequences more frequently than expected (Fig. 6a and c), or has an observed/expected usage ratio >1 (Fig. 6b and d). In *B. subtilis*, aSD sites 5′-CUCCUUU-3′ are favorably selected (Fig. 6a and b), and in *E. coli*, aSD sites 5′-CUCCUUA-3′ were found to be favorably selected (Fig. 6c and d). These results suggest that these sequences make up the extent of the core aSD sequence in the two species. We extended these sequences to 5′-CCUCCUUU-3′ in *B. subtilis* and 5′-CCUCCUUA-3′ in *E. coli* in order to examine the necessity of including the core aSD motif CCUCC in core aSD sequences.

To investigate whether our core aSD sequences are ideal for translation initiation, we consider their complementary SD sequences. We predicted that 1) putative SD sequences that complement the aforementioned core aSD sequences are favorably selected and constitute the majority of observed SD sequences used

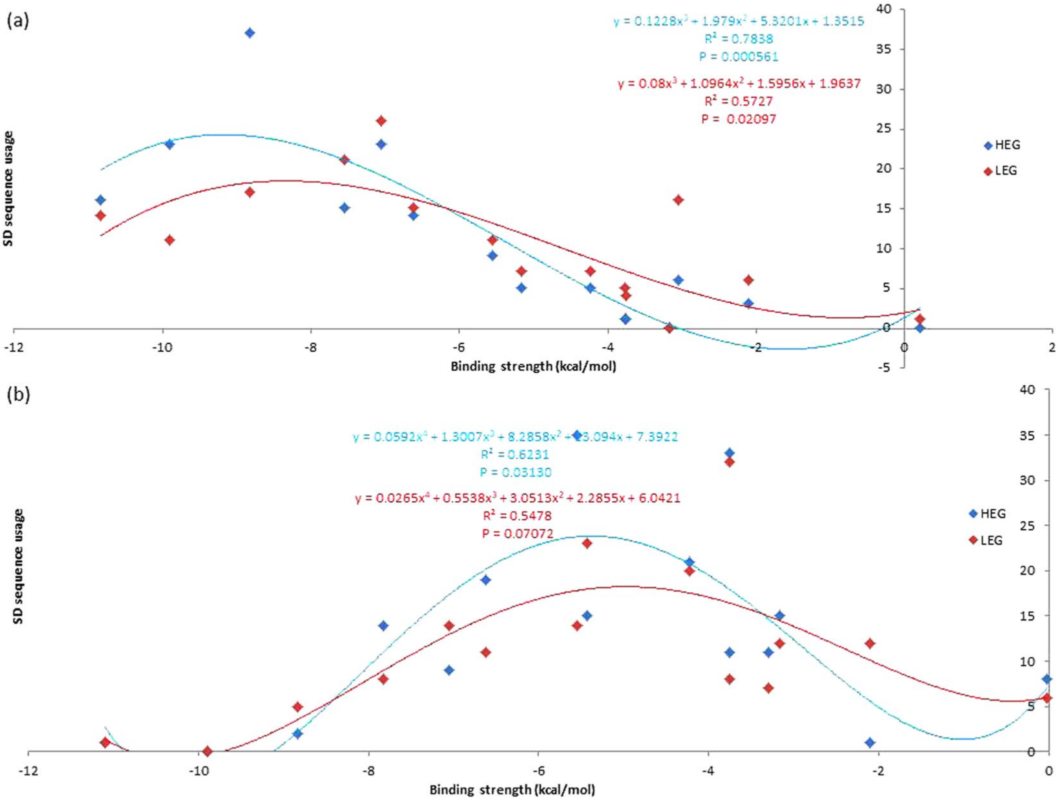


Figure 9. The association between SD sequence usage and binding affinity is more pronounced in HEGs than LEGs in (**a**) *B. subtilis* and (**b**) *E. coli*. All 4 to 8 nt SD sequences are complementary to aSD sequences CCUCCUUU-3′ (*B. subtilis*) and 5′-CCUCCUUA-3′ (*E. coli*).

by protein-coding genes, and 2) the usage of these SD sequences can be explained by their binding affinities. We found that the most abundant SD sequences used by protein-coding genes are among the four to eight nt putative SD sequences that complement 5′-CCUCCUUU-3′ in *B. subtilis* (Figs. 7a), and 5′-CCUCCUUA-3′ in *E. coli* (Fig. 7b). Furthermore, usages of these SD sequences that complement our core aSD sequences can be explained by their binding affinities (ΔG for heterodimer binding). Specifically, highly used SD sequences have relatively intermediate levels of binding affinities in *B. subtilis* (Fig. 8a: approximately −9 kcal/mol to −7 kcal/mol, $P = 0.001915$, $R^2 = 0.7282$) and in *E. coli* (Fig. 8b: approximately −6 kcal/mol to −4 kcal/mol, $P = 0.04483$, $R^2 = 0.5919$). However, usages of other SD sequences are minimal and cannot be explained by binding affinity. Thus, not all SD sequences with intermediate levels of aSD binding affinities maximize translation efficiency, only the ones that complement the core aSD sequence.

The inclusion of CCUCC in the core aSD sequence depends on the species specific preferred SD/aSD binding affinity; it is not necessarily encompassed by the core aSD sequence of all species. For example, *E. coli* has a lower preferred SD/aSD binding affinity relative to *B. subtilis*, hence SD sequences that complement CCUCC (−7.05 kcal/mol) are less selected for in the former than the latter due to the high binding affinity of the motif (Fig. 8b). Based on these observations, we suggest that the core aSD sequence is extended to 5′-CCUCCUUU-3′ in *B. subtilis*. It should also be noted that the observed/expected ratio at the 5′ C is very close to one (the base is not avoided by SD sequences; Fig. 6B).

We expect the association between SD sequence usage and binding affinity to be more pronounced in HEGs than LEGs. Indeed, SD sequences with relatively intermediate levels of binding affinity are more preferred in HEGs than LEGs in both *B. subtilis* and *E. coli* (Fig. 9). This contrast further emphasizes the importance of SD binding affinity in translation efficiency because HEGs are under greater selective pressure to evolve towards high translation efficiency than LEGs. This finding complements the claim made by Hockenberry *et al.*[9] that translation efficiency is maximized at intermediate levels of SD/aSD binding affinity, and extends their conclusion to suggest that intermediate SD binding affinities are preferred in both HEGs and LEGs.

We suggest that optimal SD sequences are 5′-AGGAGG-3′ and 5′-AAAGGAG-3′ in *B. subtilis*, and 5′-AGGAG-3′ and 5′-GGAG-3′ in *E. coli* (Fig. 9), based on their 1) high usages, especially in HEGs, 2) intermediate binding affinity to core aSD sequences (5′-CCUCCUUU-3′ in *B. subtilis*, and 5′-CUCCUUA-3′ in *E. coli*), and 3) occurrences at optimal $D_{toStart}$ locations. Elucidating the full extent of the core aSD sequence is important to identify the complete set of optimal SD/aSD pairs. For example, one would not be able to detect the highly preferred SD sequences 5′-AGGAG-3′, 5′-AAGGA-3′ and 5′-GGAG-3′ in *E. coli* using the aSD sequence 5′-CACCUCC-3′. This explains why no correlation was observed between SD binding affinity and translation efficiency in a previous study[1]. On the other hand, one will overestimate the amount of different SD sequences by extending past the core aSD sequence at either end. The usages of such SD sequences are not preferred and cannot be explained by binding affinity; they are likely poor motifs for translation initiation. Lastly, we acknowledge that there is considerable flexibility in the SD sequence (perfect complementarity is not necessary between SD and aSD bases). We speculate that this is due to the fact that intermediate levels of binding affinity are preferable.

Materials and Methods

Processing the genome and RNA-Seq data. The annotated genomes of *B. subtilis 168* (accession number: NC_000964) and *E. coli* K12 (NC_000913) in GenBank formats were retrieved from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov). Two FASTQ files in BioProject PRJNA244362 (*B. subtilis 168* wild type, experiment SRX515181, sequencing length ~ 51 nt) and (*E. coli K12* wild type experiment SRX515174, sequencing length ~ 51 nt) were downloaded from NCBI and converted into FASTA files using seqtk (https://github.com/lh3/seqtk), then subsequently into FASTA+ format using ARSDA[37] (http://dambe.bio.uottawa.ca/Include/software.aspx). The site specific qualities of RNA-Seq reads were visualized in ARSDA via the 'Get.FASTQ Info′ from the FASTQ files.

Aligning RNA-Seq reads to annotated rRNA sequences. The FASTA+ files were converted into BLAST databases using the "Create BLAST DB" function in ARSDA. Annotated segments of the 3′ 16S rDNA were used as the query sequences (the final 85 nt of 16S rRNA in *B. subtilis* accession NC_000964 and the final 60 nt of the 16S rRNA in accession NC_000913) for BLAST alignments against the generated BLAST databases (both using specified e-value cutoffs of 10^{-17} and word length = 20). The resulting hits were retrieved from the FASTA+ files using DAMBE and aligned by multiple sequence alignment (using the Clustal Omega algorithm implemented in DAMBE, default parameters) against the corresponding 16S segment for each organism. Reads were retained if they extended to at least the final C in the canonical CCUCC motif and had no errors in base calling towards the 3′ ends. All reads that match these criteria were used in generating the distributions shown in Fig. 2.

Classifying genes according to gene expression. We used protein abundance (ppm) data as proxies of gene expression. The integrated datasets were downloaded from PaxDB[38] for *E. coli* and *B. subtilis*. The *B. subtilis* protein IDs (224308-paxdb_uniprot.txt) were mapped to Gene IDs in NC_000964 using UniProt Retrieve/ID mapping http://www.uniprot.org/uploadlists/. The *E. coli* protein IDs were in the same format as the Gene IDs in NC_000913. The genes were ranked by protein abundance values, and the top and bottom 10% of the genes were classified as HEGs and LEGs, respectively. Only genes with non-zero protein abundance values were selected in this study.

Determining putative SD sequences based on pairing potential, location, and binding affinities. The 3′ TAILs 5′-GAUCACCUCCUUUCU-3′ (*B. subtilis*) and 5′-GAUCACCUCCUUA-3′ (*E. coli*) were used in identifying putative SD sequences using DAMBE[39], following the methods used in two previous studies[10,11]: 30 nt upstream of start codon of all CDSs were extracted and matched against the annotated 3′ TAIL with 'Analyzing 5′UTR' in DAMBE, with minimum SD length = 4 nt and maximum SD length = 12 nt. The SD/aSD binding affinities (ΔG for heterodimer binding) were calculated using RNAcofold with default settings[36].

Only SD sequences occurring at optimal distances relative to the start codon were analyzed in this study. The optimal distances for SD sequences were determined to be 10 to 21 $D_{toStart}$ bases in *E. coli*[10] and 15 to 21 $D_{toStart}$ bases in *B. subtilis*. $D_{toStart}$ denotes the distance between the 16S rRNA 3′ end and the start codon during SD/aSD binding.

Calculating the SD/aSD observed and expected site specific usage. The observed usage of each *B. subtilis* SD site represents the total number of times the base is observed in all putative *B. subtilis* SD sequences of protein-coding genes and of highly and lowly expressed subsets of genes. The expected usage of each SD site represents the total number of times the base is expected to occur in putative SD sequences, assuming each SD site is equally likely to be used by all SD sequences of lengths 4 nt to 12 nt (no selection bias). Thus, the expected number of SD/aSD binding at the first aSD site is represented by equation (1), with N_m denoting N observed number of SD sequences of length m:

$$\sum_{m=4}^{12} \frac{N_m}{15 - m + 1} \tag{1}$$

While the expected frequency at the sixth aSD site is represented by equation (2):

$$4 \times \frac{N_4}{12} + 5 \times \frac{N_5}{11} + 6 \times \frac{N_6}{10} + 6 \times \frac{N_7}{9} + 6 \times \frac{N_8}{8} + 6 \times \frac{N_9}{7} + 6 \times \frac{N_{10}}{6} + 6 \times \frac{N_{11}}{5} + 5 \times \frac{N_{12}}{4} \tag{2}$$

The same methodology is applied to measure usage of *E.* coli SD sequences. These computations are implemented in DAMBE[39,40] under the 'Analyze 5UTR' function.

Data availability. All data used in our analyses are publicly available in the file Supplementary Dataset 1. Raw data are extracted from the NCBI GEO DataSets database (https://www.ncbi.nlm.nih.gov/gds). The runs used for *B. subtilis* (SRR1232437) and *E. coli* (SRR1232430) are both included under accession GSE56720. The integrated protein abundance data are available at PaxDB (https://pax-db.org/).

References

1. Li, G. W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157**, 624–635, https://doi.org/10.1016/j.cell.2014.02.033 (2014).
2. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science* **324**, 255–258, https://doi.org/10.1126/science.1170160 (2009).
3. Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. USA* **107**, 3645–3650, https://doi.org/10.1073/pnas.0909910107 (2010).
4. Xia, X. A Major Controversy in Codon-Anticodon Adaptation Resolved by a New Codon Usage Index. *Genetics* **199**, 573–579 (2015).
5. Walsh, G. Therapeutic insulins and their large-scale manufacture. *Appl. Microbiol. Biotechnol.* **67**, 151–159, https://doi.org/10.1007/s00253-004-1809-x (2005).
6. Shine, J. & Dalgarno, L. The 3′-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci. USA* **71**, 1342–1346 (1974).
7. Hui, A. & de Boer, H. A. Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in Escherichia coli. *Proc. Natl. Acad. Sci. USA* **84**, 4762–4766 (1987).
8. Osterman, I. A., Evfratov, S. A., Sergiev, P. V. & Dontsova, O. A. Comparison of mRNA features affecting translation initiation and reinitiation. *Nucleic Acids Res* **41**, 474–486, https://doi.org/10.1093/nar/gks989 (2013).
9. Hockenberry, A. J., Pah, A. R., Jewett, M. C. & Amaral, L. A. Leveraging genome-wide datasets to quantify the functional role of the anti-Shine-Dalgarno sequence in regulating translation efficiency. *Open biology* **7**, https://doi.org/10.1098/rsob.160239 (2017).
10. Abolbaghaei, A., Silke, J. R. & Xia, X. How Changes in Anti-SD Sequences Would Affect SD Sequences in *Escherichia coli* and Bacillus subtilis. *G3 (Bethesda, Md.)* **7**, 1607–1615, https://doi.org/10.1534/g3.117.039305 (2017).
11. Prabhakaran, R., Chithambaram, S. & Xia, X. Escherichia coli and Staphylococcus phages: effect of translation initiation efficiency on differential codon adaptation mediated by virulent and temperate lifestyles. *J Gen Virol* **96**, 1169–1179, https://doi.org/10.1099/vir.0.000050 (2015).
12. de Smit, M. H. & van Duin, J. Translational initiation on structured messengers. Another role for the Shine-Dalgarno interaction. *J Mol Biol* **235**, 173–184 (1994).
13. Murray, C. L. & Rabinowitz, J. C. Nucleotide sequences of transcription and translation initiation regions in Bacillus phage phi 29 early genes. *J. Biol. Chem.* **257**, 1053–1062 (1982).
14. Green, C. J., Stewart, G. C., Hollis, M. A., Vold, B. S. & Bott, K. F. Nucleotide sequence of the Bacillus subtilis ribosomal RNA operon, rrnB. *Gene* **37**, 261–266 (1985).
15. Uchida, T. *et al.* The use of ribonuclease U2 in RNA sequence determination. Some corrections in the catalog of oligomers produced by ribonuclease T1 digestion of Escherichia coli 16S ribosomal RNA. *J. Mol. Evol.* **3**, 63–77 (1974).
16. Woese, C. R. *et al.* Conservation of primary structure in 16S ribosomal RNA. *Nature* **254**, 83–86 (1975).
17. Barbe, V. *et al.* From a consortium sequence to a unified sequence: the Bacillus subtilis 168 reference genome a decade later. *Microbiology* **155**, 1758–1775, https://doi.org/10.1099/mic.0.027839-0 (2009).
18. Sohmen, D. *et al.* Structure of the Bacillus subtilis 70S ribosome reveals the basis for species-specific stalling. *Nature communications* **6**, 6941, https://doi.org/10.1038/ncomms7941 (2015).
19. Deutscher, M. P. Twenty years of bacterial RNases and RNA processing: how we've matured. *RNA* **21**, 597–600, https://doi.org/10.1261/rna.049692.115 (2015).
20. Sulthana, S. & Deutscher, M. P. Multiple exoribonucleases catalyze maturation of the 3′ terminus of 16S ribosomal RNA (rRNA). *J. Biol. Chem.* **288**, 12574–12579, https://doi.org/10.1074/jbc.C113.459172 (2013).
21. Jacob, A. I., Kohrer, C., Davies, B. W., RajBhandary, U. L. & Walker, G. C. Conserved bacterial RNase YbeY plays key roles in 70S ribosome quality control and 16S rRNA maturation. *Mol Cell* **49**, 427–438, https://doi.org/10.1016/j.molcel.2012.11.025 (2013).
22. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63, https://doi.org/10.1038/nrg2484 (2009).
23. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621–628, https://doi.org/10.1038/nmeth.1226 (2008).
24. Li, S., Dong, X. & Su, Z. Directional RNA-seq reveals highly complex condition-dependent transcriptomes in E. coli K12 through accurate full-length transcripts assembling. *BMC Genomics* **14**, 520, https://doi.org/10.1186/1471-2164-14-520 (2013).
25. Lim, K., Furuta, Y. & Kobayashi, I. Large variations in bacterial ribosomal RNA genes. *Mol Biol Evol* **29**, 2937–2948, https://doi.org/10.1093/molbev/mss101 (2012).
26. Ma, J., Campbell, A. & Karlin, S. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol* **184**, 5733–5745 (2002).
27. Schurr, T., Nadir, E. & Margalit, H. Identification and characterization of E.coli ribosomal binding sites by free energy computation. *Nucleic Acids Res* **21**, 4019–4023 (1993).
28. Starmer, J., Stomp, A., Vouk, M. & Bitzer, D. Predicting Shine-Dalgarno Sequence Locations Exposes Genome Annotation Errors. *PLoS Comput Biol* **2**, e57 (2006).
29. Nakagawa, S., Niimura, Y., Miura, K. & Gojobori, T. Dynamic evolution of translation initiation mechanisms in prokaryotes. *Proc. Natl. Acad. Sci. USA* **107**, 6382–6387, https://doi.org/10.1073/pnas.1002036107 (2010).
30. Li, G. W. How do bacteria tune translation efficiency? *Curr Opin Microbiol* **24C**, 66–71, https://doi.org/10.1016/j.mib.2015.01.001 (2015).
31. Li, G. W., Oh, E. & Weissman, J. S. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484**, 538–541, https://doi.org/10.1038/nature10965 (2012).
32. Vimberg, V., Tats, A., Remm, M. & Tenson, T. Translation initiation region sequence preferences in *Escherichia coli*. *BMC Mol Biol* **8**, 100, https://doi.org/10.1186/1471-2199-8-100 (2007).
33. Lin, Y. H., Chang, B. C., Chiang, P. W. & Tang, S. L. Questionable 16S ribosomal RNA gene annotations are frequent in completed microbial genomes. *Gene* **416**, 44–47, https://doi.org/10.1016/j.gene.2008.02.023 (2008).
34. Jones, C. E., Brown, A. L. & Baumann, U. Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics* **8**, 170, https://doi.org/10.1186/1471-2105-8-170 (2007).
35. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100–3108, https://doi.org/10.1093/nar/gkm160 (2007).
36. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26, https://doi.org/10.1186/1748-7188-6-26 (2011).
37. Xia, X. ARSDA: A New Approach for Storing, Transmitting and Analyzing Transcriptomic Data. *G3: Genes|Genomes|Genetics*. https://doi.org/10.1534/g3.117.300271 (2017).
38. Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D. & von Mering, C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* **15**, 3163–3168, https://doi.org/10.1002/pmic.201400441 (2015).
39. Xia, X. DAMBE6: New Tools for Microbial Genomics, Phylogenetics, and Molecular Evolution. *J Hered* **108**, 431–437, https://doi.org/10.1093/jhered/esx033 (2017).
40. Xia, X. DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. *Mol Biol Evol* **30**, 1720–1728 (2013).

Acknowledgements

This study was supported by Discovery Grants of the Natural Science and Engineering Research Council of Canada to X.X. (NSERC, RGPIN/261252-2013).

Author Contributions

Y.W., J.R.S. and X.X. contributed equally to the design of the study. Y.W. and J.R.S. wrote the main manuscript text and prepared Figs 1–5. X.X. was responsible for devising the SD usage algorithm and computational analyses in ARSDA. All authors reviewed the manuscript. X.X. supervised the project.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-17918-6.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.