How Changes in Anti-SD Sequences Would Affect SD Sequences in Escherichia coli and Bacillus subtilis

Akram Abolbaghai et al.†, Jordan R. Silke,† and Xuhua Xia*†,1
*Department of Biology, University of Ottawa, Ontario K1N 6N5, Canada and †Ottawa Institute of Systems Biology, Ontario K1H 8M5, Canada
ORCID ID: 0000-0002-3092-7566 (X.X.)

ABSTRACT The 3’ end of the small ribosomal RNAs (ssu rRNA) in bacteria is directly involved in the selection and binding of mRNA transcripts during translation initiation via well-documented interactions between a Shine-Dalgarno (SD) sequence located upstream of the initiation codon and an anti-SD (aSD) sequence at the 3’ end of the ssu RNA. Consequently, the 3’ end of ssu rRNA (3’TAIL) is strongly conserved among bacterial species because a change in the region may impact the translation of many protein-coding genes. Escherichia coli and Bacillus subtilis differ in their 3’ ends of ssu rRNA, being GAUCACCUCUUUA3’ in E. coli and GAUCACCUCUUUUCU3’ or GAUCACCUCUUUCUA3’ in B. subtilis. Such differences in 3’TAIL lead to species-specific SDs (designated SD Ec for E. coli and SD Bs for B. subtilis) that can form strong and well-positioned SD/aSD pairing in one species but not in the other. Selection mediated by the species-specific 3’TAIL is expected to favor SD Bs against SD Ec in B. subtilis, but favor SD Ec against SD Bs in E. coli. Among well-positioned SDs, SD Ec is used more in E. coli than in B. subtilis, and SD Bs more in B. subtilis than in E. coli. Highly expressed genes and genes of high translation efficiency tend to have longer SDs than lowly expressed genes and genes with low translation efficiency in both species, but more so in B. subtilis than in E. coli. Both species overuse SDs matching the bolded part of the 3’TAIL shown above. The 3’TAIL difference contributes to the host specificity of phages.

Many studies suggest that initiation is the principle bottleneck of the translation process in bacteria (Liljenstrom and von Heijne 1987; Bulmer 1991; Xia 2007a; Xia et al. 2007; Kudla et al. 2009; Tuller et al. 2010; Prabhakaran et al. 2015). Successful initiation requires that the ribosome is able to bind to the mRNA template in such a manner that the start codon correctly lines up at the ribosomal P site (Farwell et al. 1992; Komarova et al. 2002; Duval et al. 2013). This translation initiation process in most bacterial species is facilitated by (1) ribosomal protein S1 (RPS1) acting as an RNA chaperone that unfolds secondary structural elements that may otherwise embed the start codon and obscure the start signal (Vellonoweth and Rabinowitz 1992; Duval et al. 2013; Prabhakaran et al. 2015), and (2) the Shine-Dalgarno (SD) sequence located upstream of the start codon (Shine and Dalgarno 1974, 1975; Steitz and Jakes 1975; Dunn et al. 1978; Taniguchi and Weissmann 1978; Eckhardt and Luhrmann 1979; Luhrmann et al. 1981) that base-pairs with anti-SD (aSD) located at the free 3’ end of ssu rRNA (3’TAIL) located at the free 3’ end of the small ribosomal rRNA (ssu rRNA, whose 3’ end will hereafter be referred to as 3’TAIL). A well-positioned SD/aSD pairing and reduced secondary structure in sequences flanking the start codon and SD are the hallmarks of highly expressed genes in Escherichia coli and Staphylococcus aureus, as well as their phages (Prabhakaran et al. 2015).

The SD/aSD pairing offers a simple and elegant solution to start codon recognition in bacteria and their phages (Hui and de Boer 1987; Vimberg et al. 2007; Prabhakaran et al. 2015). Because many protein-coding genes depend on aSD motifs located at 3’TAIL for translation, strong sequence conservation is observed in the 3’TAIL among diverse bacterial species (Woese 1987; Orso et al. 1994; Clarridge 2004; Chakravorty et al. 2007). Conversely, a change in 3’TAIL is expected to result in fundamental changes in SD usage in protein-coding genes.

E. coli, as a representative of the gram-negative bacteria, and Bacillus subtilis, as a representative of gram-positive bacteria, differ in their
3’TAIL in only a minor detail, with the former ending with A and the latter with 3’UCU or 3’AUCU (Table 1). 3’UCU was suggested by early experimental studies (Murray and Rabinowitz 1982; Band and Henner 1984), and annotated in the *B. subtilis* genome database SubtiList (http://genolist.pasteur.fr/SubtiList/). However, 3’AUCU appears in *B. subtilis* genomes annotated in GenBank (e.g., NC_000964). A recent study on *B. subtilis* ribosomal structure (e.g., Solmen et al. 2015) also assumed a 3’AUCU tail in ssu rRNA (D. Wilson, personal communication). Existing evidence suggests heterogeneous “mature” ssu rRNA pool given that mature ssu rRNA in bacterial species results from endoribonuclease digestion from the precursor 30S rRNA followed by exonuclease nibbling (Britton et al. 2007; Yao et al. 2007; Kurata et al. 2015). For example, 3’→ 5’ exoribonucleases such as RNases II, R, and PH, as well as PNPase, all participate in maturation of the 3’TAIL of ssu rRNA (Sulthana and Deutscher 2013), and endoribonuclease YbeY has also been recently shown to participate in the 3’end maturation of ssu rRNA (Davies et al. 2010; Jacob et al. 2013). In *E. coli*, 67% of mature ssu rRNA ends with the 3’TAIL (Kurata et al. 2015). Thus, the trailing 3’UCU and 3’ACUC may both be present in functional ssu rRNA of *B. subtilis*.

The minor difference in 3’TAIL between *E. coli* and *B. subtilis* suggests different sets of permissible SDs between the two species, i.e., some SDs that function well in one species may not function at all in the other. These species-specific SDs (Table 1) include six in *E. coli* (designated SD*a* and 25 in *B. subtilis* (designated SD*b*). Such differences in permissible SDs could contribute to fundamental species differences in translation.

Most *E. coli* mRNAs cannot be efficiently translated in *B. subtilis* (McLaughlin et al. 1981a,b), but most *B. subtilis* mRNAs can be efficiently translated in *E. coli* (Stallcup et al. 1976). Many gram-negative bacteria, including *E. coli*, can even translate poly(U) messages (Nirenberg and Matthaei 1961; Stallcup et al. 1976) but gram-positive bacteria, including *B. subtilis*, cannot translate poly(U) messages (Stallcup et al. 1976). In retrospect, it was indeed good luck that Nirenberg and Matthaei (1961) happened to experiment with *E. coli* instead of *B. subtilis*, otherwise the landmark study would have ended up with nothing to report. It is also known that *E. coli* translation machinery can translate leaderless mRNAs (O’Donnell and Janssen 2002; Krishnan et al. 2010; Vesper et al. 2011; Giliberti et al. 2012), and that its 30S ribosomal subunit can still localize the start codon even when the last 30 nucleotides of ssu rRNA is deleted (Melincon et al. 1990).

The differences in mRNA permissibility between gram-negative and gram-positive bacteria is often attributed to the presence of the six-domain that is highly conserved RPS1 in gram-negative bacteria (Subramanian 1983), but absent or highly variable in gram-positive bacteria with translation specificity (Roberts and Rabinowitz 1989). RPS1 facilitates translation initiation by reducing secondary structure that could otherwise embed the translation initiation region (TIR) which includes SD and start codon (Roberts and Rabinowitz 1989; Farwell et al. 1992; Tzareva et al. 1994). *B. subtilis* has a homologous gene with four domains that are not conserved among gram-positive bacteria, with *Mycoplasma pulmonis* and *Spiroplasma kunkelli* having only one domain with weak homology to any known functional RPS1 (Salah et al. 2009). These findings corroborate earlier experimental evidence (McLaughlin et al. 1981b; Band and Henner 1984) demonstrating that *B. subtilis* requires a more stringent SD region for gene expression than does *E. coli*.

However, the conventional belief that *E. coli* possesses a more permissible translation machinery than *B. subtilis* is not always true. In rare cases, some mRNAs that can be translated efficiently in *B. subtilis* cannot be translated well in *E. coli*, and one such mRNA is gene 6 of the *B. subtilis* phage φ29 (Vellanoweth and Rabinowitz 1992). In particular, such translation specificity can often be traced to the 30S ribosome and the mRNAs, rather than other components of the translation machinery, strongly suggesting SD/aSD pairing as the cause for the translation specificity. Indeed, we will show later, gene 6 of phage

| Table 1 ssu RNA 3’ ends that are free to base-pair with SD motifs in *E. coli* and *B. subtilis* and their compatible motifs |
|-----------------------------------------------|
| **Species and 3’ TAIL Sequence** | **SD Motifs** |
|-----------------------------------------------|
| *E. coli* 3’-AUUCUCCACUAG-5’ | UAAG |
| *E. coli* 3’-AUUCUCCACUAG-5’ | UAAGG |
| *E. coli* 3’-AUUCUCCACUAG-5’ | UAAGGA |
| *E. coli* 3’-AUUCUCCACUAG-5’ | UAAGGAG |
| *E. coli* 3’-AUUCUCCACUAG-5’ | UAAGGAGG |
| *E. coli* 3’-AUUCUCCACUAG-5’ | UAAGGAGGG |
| *E. coli* 3’-AUUCUCCACUAG-5’ | UAAGGAGGGG |
| *E. coli* 3’-AUUCUCCACUAG-5’ | UAAGGAGGGGG |
| *E. coli* 3’-AUUCUCCACUAG-5’ | UAAGGAGGGGGG |
| *B. subtilis* 3’-AUCUUUCCUCCACUAG-5’ | AAAAG |
| *B. subtilis* 3’-AUCUUUCCUCCACUAG-5’ | AAGG |
| *B. subtilis* 3’-AUCUUUCCUCCACUAG-5’ | AAGGA |
| *B. subtilis* 3’-AUCUUUCCUCCACUAG-5’ | AAGGAG |
| *B. subtilis* 3’-AUCUUUCCUCCACUAG-5’ | AAGGAGG |
| *B. subtilis* 3’-AUCUUUCCUCCACUAG-5’ | AAGGAGGG |
| *B. subtilis* 3’-AUCUUUCCUCCACUAG-5’ | AAGGAGGGG |
| *B. subtilis* 3’-AUCUUUCCUCCACUAG-5’ | AAGGAGGGGG |
| *B. subtilis* 3’-AUCUUUCCUCCACUAG-5’ | AAGGAGGGGGG |

* Bolded letters show the differences in the base composition between two species. (*E. coli* ends with A whereas *B. subtilis* ends with UCU or AUCU). The underlined nucleotides denote the alternative 3’-AUCU-5’ TAIL and motifs exclusively compatible with it.

b The SD motifs shown are derived from differences in 3’TAIL (boldface) for both species.
ψ29 can form a well-positioned SD/aSD pair only with the 3′ TAIL of \textit{B. subtilis} but not with that of \textit{E. coli}. Thus, proper SD/aSD pairing of mRNAs may be the key factor in specifying host specificity of phages, in determining whether a horizontally transferred gene will function in the new genetic background of the host cell, and, ultimately, in speciation and diversification of bacterial lineages.

To facilitate the quantification of optimal positioning of SD/aSD base pairing, we adopted a model of SD/aSD interaction proposed recently (Prabhakaran \textit{et al.} 2015), illustrated with $D_{\text{toStart}}$ as a better measure of quantifying the optimal positioning of SD and aSD than the conventional distance from putative SD to start codon. SD1 or SD2, as illustrated, are equally good in positioning the start codon AUG against the anticodon of the initiation tRNA, but they differ in their distances to the start codon. $D_{\text{toStart}}$ is the same for the two SDs. (C, D) $D_{\text{toStart}}$ is constrained to a narrow range in \textit{E. coli} (C) and \textit{B. subtilis} (D); solid blue line denotes SD hits with the UCU-ending TAIL, and the dashed red line shows SD hits with the UCUA-ending TAIL. The y-axis in (C) and (D) represents the percentage of SD motif hits detected. See Materials and Methods section for details.

**Figure 1** A model of SD sequence and aSD interactions. (A) The free 3′ end of SSU rRNA (3′ TAIL) of \textit{E. coli} and \textit{B. subtilis} based on the predicted secondary structure of the 3′ end of the ssu rRNA of \textit{E. coli} and \textit{B. subtilis} from mfold 3.1, adapted from the comparative RNA web site and project (http://www.rna.icmb.utexas.edu). (B) A schematic representation of SD and aSD interaction illustrates $D_{\text{toStart}}$ as a better measure for quantifying the optimal positioning of SD and aSD than the conventional distance from putative SD to start codon. SD1 or SD2, as illustrated, are equally good in positioning the start codon AUG against the anticodon of the initiation tRNA, but they differ in their distances to the start codon. $D_{\text{toStart}}$ is the same for the two SDs. (C, D) $D_{\text{toStart}}$ is constrained to a narrow range in \textit{E. coli} (C) and \textit{B. subtilis} (D); solid blue line denotes SD hits with the UCU-ending TAIL, and the dashed red line shows SD hits with the UCUA-ending TAIL. The y-axis in (C) and (D) represents the percentage of SD motif hits detected. See Materials and Methods section for details.

### MATERIALS AND METHODS

#### Retrieval of genome sequence and protein abundance data

The annotated whole genome sequences for \textit{E. coli} K12 (accession number# NC_000913.3) and \textit{B. subtilis} 168 (accession # NC_000964.3) in GenBank format were downloaded from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov). Excluding 180 sequences annotated as pseudogenes in the \textit{E. coli} genome from the analysis resulted in a final total of 4139 genes from \textit{E. coli} and 4175 from \textit{B. subtilis}.

Protein abundance data were retrieved from PaxDB (Wang \textit{et al.} 2012) at www.pax-db.org. The integrated data sets were downloaded for both \textit{B. subtilis} and \textit{E. coli} in order to maximize coverage and consistency scores. We downloaded the paxdb-uniprot-links file relevant to the species (e.g., 224308-paxdb_uniprot.txt for \textit{B. subtilis}), saved the Uniprot ID (the last column) to a file (e.g., BsUniprotID.txt), and browsed to http://www.uniprot.org/uploadlists (last accessed March 7, 2017) to obtain GeneID. Under “Provide your identifiers,” we uploaded the BsUniprotID.txt file, under “Selection options,” we selected the mapping from “UniProtKB AC/ID” to “Gene name” (or GeneID), and clicked “Go”. The STRING identifiers used for each gene in the protein abundance data sets were converted into Gene IDs using UniProt’s retrieve/ID mapping tool (http://www.uniprot.org/uploadlists/) for use in subsequent analyses. The resulting mapping file was generated with two columns (original input Uniprot IDs and the mapped gene name (or GIs GeneID) corresponding to gene name or other IDs in a GenBank file. Unmapped ID is stored in a separate file, also available for downloading.

#### HEGs and LEGs

Genes were delimited as HEGs or LEGs on the basis of two metrics: steady state protein abundance levels taken from PaxDB, and $I_{\text{TE}}$ (Index of translation elongation) scores computed with DAMBE (Xia 2013)
using the default reference files for *E. coli* and *B. subtilis*, which were included in the DAMBE distribution. $t_{1E}$ is advantageous over the DAMBE adaptation index (CAI Sharp and Li 1987) or its improved form (Xia et al. 2007b) in that it takes background mutation bias into consideration (Xia 2015). DAMBE's $t_{1E}$ function has four settings that differ in their treatment of synonymous codon families, and we selected the option breaking sixfold degenerate codon families into four and twofold families. For *E. coli* and *B. subtilis*, the top and bottom 10% of genes for both of these metrics were designated as HEGs and LEGs, respectively.

**Genes of high translation efficiency (HTE) and low translation efficiency (LTE)**

HEGs and LEGs defined as above may not be the same as HTE genes and LTE genes. HTE and LTE genes may be characterized by regressing protein abundance on mRNA abundance, so that, given genes with the same mRNA level, those producing many proteins are translated more efficiently than those producing few. The former would be HTE genes, and the latter LTE genes. This requires proteomic and transcriptomic studies carried out with similar bacterial strains, and under similar culture and growth conditions. For *E. coli*, we have used proteomic data from Lu et al. (2007) deposited at PDB (Wang et al. 2012), and transcriptomic data in RPKM (reads per kilobase per million matched reads) from the wild-type strain of *E. coli* (BioProject PRJNA257498, Pobre and Arraiano 2015). For *B. subtilis*, the proteomic data are from Chi et al. (2011) deposited in PDB and transcriptomic raw counts for three wild-type replicates were downloaded from BioProject PRJNA319983 (GSM2137056 to SM2137058), and then normalized to RPKM. These two transcriptomic studies ignored reads that match the minimum free energy (MFE) state of the structure that in turn were predicted using mfold version 3.1 (http://unafold.rna.albany.edu/?q=mfold; Zuker 2003), with the resulting free 3' ends shown in Figure 1A.

The sequence of the 3' TAIL used in our analysis for *E. coli* is 3'-AUUUCCUCACUAG-5' (Shine and Dalgarno 1974; Brosius et al. 1978; Gold et al. 1981; Luhrmann et al. 1981; Band and Henner 1984; Tu et al. 2009), because, based on the *E. coli* SSU rRNA secondary structure (Woese et al. 1980; Noah et al. 2000; Yassin et al. 2005; Kitahara et al. 2012; Prabhakaran et al. 2015), these are the 13 nt at the 3' end of the ssu rRNA that are free to base pair with the SD sequence. There are two versions of 3'TAIL for *B. subtilis*: 3'-UCUUCCUCACUAG (Murray and Rabinowitz 1982; Band and Henner 1984), and 3'-AUCUUCCUCACUAG in the genomic annotation. We discussed the possibility of heterogeneous "mature" ssu rRNA pool in the Introduction.

**Identification of anti-SD and SD sequences**

The 3' TAILs for *B. subtilis* and *E. coli* used in this paper were based on early empirical evidence (Shine and Dalgarno 1974; Brosius et al. 1978; Gold et al. 1981; Luhrmann et al. 1981; Murray and Rabinowitz 1982; Band and Henner 1984; Tu et al. 2009), as well as a series of chemical modification and nuclease digestion experiments that aimed to identify the sequence and secondary structure of bacterial ssu rRNAs using *E. coli* and *Bacillus brevis* (Woese et al. 1980). The experimentally derived 3' TAILs for both species are compatible with their corresponding ssu RNA secondary structure schematics from the Comparative RNA Web Site & Project at [www.rna.icmb.utexas.edu](http://www.rna.icmb.utexas.edu), which is curated by the Gutell Lab at the University of Texas at Austin. The schematics include base pairing interactions that are predicted based on the minimum free energy (MFE) state of the structure that in turn were predicted using mfold version 3.1 (http://unafold.rna.albany.edu/?q=mfold; Zuker 2003), with the resulting free 3' ends shown in Figure 1A.
RESULTS AND DISCUSSION

E. coli has 4323 protein-coding genes (CDSs), with 180 annotated as pseudogenes in the genome and excluded from the analysis, resulting in 4144 functional CDSs. B. subtilis has 4175 CDSs with none annotated as pseudogenes. The genomic nucleotide frequencies are 0.2462, 0.2542, 0.2537, and 0.2459, respectively for A, C, G, and T in E. coli. The corresponding values in B. subtilis are 0.2818, 0.2181, 0.2171, and 0.2830, respectively.

SD_Ec and SD_Bs are used more in E. coli and B. subtilis, respectively

As expected, SD_Ec are much more frequent in E. coli than in B. subtilis, with 455 in E. coli, in contrast to 267 in B. subtilis (Table 2). The difference is highly significant, either against the null hypothesis of equal frequencies ($\chi^2 = 48.9529$, $P < 0.0001$), against the expected value based on the relative number of CDSs ($\chi^2 = 50.3648$, $P < 0.0001$; a slightly increased $\chi^2$ is because E. coli has slightly fewer included CDSs than B. subtilis), or against the expected values based on both relative number of CDSs and genomic nucleotide frequencies (e.g., AGAA is proportional to $P_X^3P_G$, AGAAA to $P_X^4P_G$, and so on, where $P_X$ is the genomic frequency of nucleotide X in either E. coli or B. subtilis), with $\chi^2 = 103.07$, $P < 0.0001$.

The relative abundance of different SDs depends on selection favoring an optimal SD length, and mutations disrupting long SDs. In E. coli, the optimal SD length is six (Vimberg et al. 2007). B. subtilis favors longer SDs. In an experiment with B. subtilis with SD lengths of 5, 6, 7, and 12, longer SDs consistently produce more proteins than shorter ones (Band and Henner 1984). This is consistent with the results presented in Table 2, where UAAG is expected to be strongly selected against in B. subtilis because it can form only 3 bp against B. subtilis 3’TAIL. However, the longer SD_Ec is not selected against because an SD_Ec such as UAAGGGG can form 7 bp (except for the first U) against B. subtilis 3’TAIL.

Also as expected, SD_Ec are also more frequent in B. subtilis than in E. coli, with 1203 SD_Ec in B. subtilis in contrast to 576 in E. coli (Table 3). The difference is also highly significant ($P < 0.0001$) using the same tests for SD_Ec results in Table 2. However, one interesting deviation from the SD_Ec data is that SD_Bs of length 4 exhibit the opposite pattern, being more frequent in E. coli than in B. subtilis (Table 3), which assumes a 3’UCU-ending in B. subtilis 3’TAIL. The pattern is the same with 3’AUCU-ending of the TAIL (Table S1). This observation can be explained by stronger selection against short SD/aSD in B. subtilis than in E. coli. Translation efficiency increases with longer and more stringent SD/aSD binding in B. subtilis, and such dependence is much stronger in B. subtilis than in E. coli (Band and Henner 1984). The predicted free energy of SD/aSD for an average B. subtilis message is at least 6 kcal/mol more than that of an average SD/aSD in E. coli (Hager and Rabinowitz 1985). Thus, a short SD is expected to be selected against, and, consequently, rare in B. subtilis, consistent...
Highly expressed genes tend to have longer SDs

In addition to the observed difference in SD length between *E. coli* and *B. subtilis* (Figure 2 and Table 3; *B. subtilis* SDs tend to be longer than *E. coli* SDs), there is also clear difference between HEGs and LEGs, or between genes of HTE and of LTE. Although SDs of length four are more frequent in *E. coli*, longer SDs are relatively more represented in *B. subtilis* than in *E. coli* (Figure 2A). This is consistent with previous experimental studies demonstrating an optimal SD length of 3–5 nt. In addition to the observed difference in SD length between genes of HTE and of LTE. Although SDs of length four are the most frequent in *E. coli*, longer SDs are relatively more represented in *B. subtilis* than in *E. coli* (Figure 2 and Table 3; *B. subtilis* (Figure 2A). This is consistent with previous experimental studies demonstrating an optimal SD length of six (Schurr et al. 1993; Komarova et al. 2002; Vimbarg et al. 2007). Optimal SDs in *B. subtilis* are even longer (Band and Henner 1984) than in *E. coli* (Figure 2). We thus expect HEGs or HTE genes to have relatively longer SDs than LEGs or LTE genes, especially in *B. subtilis*. Our empirical results (Figure 2) strongly support this expectation. Short SDs are overrepresented in LEGs and LTE genes, and longer SDs overrepresented in HEGs and HTE genes in both *E. coli* and *B. subtilis*, but not in *B. subtilis* (Figure 2). This pattern, i.e., association of long SDs with HEGs and HTE genes is highly significant for *B. subtilis* (chi-square = 12.0375, df = 1, P-value = 0.0005214) when tested by the Cochran-Armitage test (Agresti 2002, pp. 181–182) for contingency tables with a linear trend as implemented in the coin package in R (Hothorn et al. 2006, 2008). The result for *E. coli*, while consistent with the expectation, is not significant at the 0.05 level (chi-square = 3.3948, df = 1, P-value = 0.0654).

Differential usage of SD_{Ec} and SD_{Bs} in HEGs and LEGs

SD_{Ec} is used more frequently in HEGs than LEGs in *E. coli* (Table 4). In contrast, SD_{Bs} is used mainly in LEGs in *B. subtilis* (Table 5), prompting the question of what SDs are used by *B. subtilis* HEGs, and whether the core aSD region (where most HEGs have SD to pair against) for *B. subtilis* HEGs include the trailing 3’UCU (or 3’AUC). The pattern is similar when contrasting between HTE genes and LTE genes (results not shown). The core aSD region is centered at CCUCC in the overwhelming majority of surveyed prokaryotes (Ma et al. 2002; Nakagawa et al. 2010; Lim et al. 2012). If *B. subtilis* has the same core aSD region, then the trailing 3’UCU (or 3’AUC) will be used rarely, consequently with few SD_{Bs} pairing to it. The distribution of SDs in *E. coli* and *B. subtilis* is consistent with this interpretation (Figure 3). SDs overrepresented in HEGs relative to LEGs use exclusively 3’AUUCCUCCA as the core aSD region in *E. coli*, and 3’UUCCUCCA as the core aSD region in *B. subtilis* (Figure 3). The trailing 3’UCU (or 3’AUC) is used as part of aSD mainly by LEGs in *B. subtilis*. The mature ssu rRNA pool may be heterogeneous in *B. subtilis*. A number of 3’→5’ exoribonucleases, such as RNases II, R, and PH, as well as PNPase, participate in maturation of the 3’TAIL of ssu rRNA (Suthana and Deutscher 2013), and nuclease YbY has also been shown recently to participate in the 3’ end maturation of ssu rRNA (Davies et al. 2010; Jacob et al. 2013). The continuous 3’→5’ digestion implies that the 3’AUC will become 3’UCU, 3’CU, and so on. It would make sense for HEGs to use SD pairs with the less volatile part of the 3’TAIL of ssu rRNA (Table 5).

Figure 3, Table 4, and Table 5 suggest that many HEGs in *E. coli* use the species-specific SD_{Ec} and will experience translation initiation problems when translated by the *B. subtilis* translation machinery. In contrast, most HEGs in *B. subtilis* do not use the species-specific SD_{Bs} and will have no translation initiation problems when translated by the *E. coli* translation machinery. Early studies have suggested a more permissible translation machinery in *E. coli* than in *B. subtilis*, i.e., most *E. coli* mRNAs cannot be efficiently translated in *B. subtilis* (McLaughlin et al. 1981;a,b) but most *B. subtilis* mRNAs can be efficiently translated in *E. coli* (Stallcup et al. 1976). The discrepancy in this translation permissibility is often attributed to the presence of the six-domain highly conserved RPSI in gram-negative bacteria (Subramanian 1983) but absent in gram-positive bacteria with translation specificity (Roberts and Rabinowitz 1989). Our results (Figure 3, Table 4, and Table 5) suggest an alternative explanation for the discrepancy. Because these early studies often involve HEGs,
and because E. coli HEGs often use species-specific SDs (Table 4) whereas B. subtilis HEGs rarely use species-specific DBs, it is not surprising that E. coli HEG messages tend to fail in translation initiation in B. subtilis, but B. subtilis HEG messages tend to have no problem in translation initiation in E. coli.

Species-specific SD and host specificity

One rare exception to the general observation that E. coli possesses a more permissible translation machinery than B. subtilis is gene 6 (gp6) of the B. subtilis phage φ29, which can be translated efficiently in B. subtilis but not in E. coli (Vellanoweth and Rabinowitz 1992). Among the 16 nonhypothetical genes in phage φ29, gp6 is the only one that uses a species-specific SDb (UAGAAAG) exclusively (Table 6). This SD used all four nucleotides at 3′TAIL of B. subtilis, and consequently cannot form SD/aSD in E. coli (Table 6). Because gp6 is an essential gene, its use of a SDb may explain its host-specificity. That is, even if it gains entry into an E. coli-like host, it will not be able to survive and reproduce successfully.

Another case of host-specificity that may be explained by SD/aSD binding is E. coli phage PRD1, which has codon usage deviating greatly from that of its host, in contrast to the overwhelming majority of E. coli phages, whose codon usage exhibits high concordance with that of the host (Chithambaram et al. 2014). Phage PRD1 belongs to the peculiar Tectiviridae family whose other members, i.e., phages PR3, PR4, PR5, L17, and PR772, parasitize gram-positive bacteria. Phage PRD1 is the only species in the family known to parasitize a variety of gram-negative bacteria, including Salmonella, Pseudomonas, Escherichia, Proteus, Vibrio, Acinetobacter, and Serratia species (Bamford et al. 1995; Grahn et al. 2006). Phage PRD1 is extremely similar to its sister lineages, parasitizing gram-positive bacteria; there is only one amino acid difference in the coat protein between PRD1 and PR4 (Bamford et al. 1995). It is thus quite likely that the ancestor of phage PRD1 parasitizes gram-positive bacteria. The lineage leading to Phage PRD1 may have switched to gram-negative bacterial hosts only recently, and thus still has codon usage similar to its ancestral gram-positive bacterial host, which is indeed the case (Chithambaram et al. 2014). However, one nonhypothetical gene in phage PRD1 (PRD1_09)
has evolved an \textit{E. coli}-specific SD (UAAG), and does not have alternative SD that can form a well-positioned SD/ASD with \textit{B. subtilis} 3’T\textsc{tail}. This may have contributed to the host limitation of phage PRD1 within \textit{E. coli}-like species.

The study of coevolution between SD and ASD sequences would be facilitated if 3’T\textsc{tails} of many bacterial species were characterized experimentally, and if these 3’T\textsc{tails} differ substantially from each other in different lineages. At present, strong experimental evidence is available for 3’T\textsc{tail} of \textit{E. coli} and \textit{B. subtilis} (except for the uncertainty on whether the 3’T\textsc{tail} ends with 3’UUC or 3’AUCA). However, RNA-Seq data may become available for many bacterial species in the near future, and should pave the way for rapid characterization of 3’T\textsc{tail} of different species by simply mapping the sequence reads to ssu rRNA genes on the genome. One problem to be aware of is that most transcriptomic studies will use an rRNA removal kit to remove the large rRNAs, i.e., 16S and 23S rRNA, in bacteria, because otherwise sequence reads from these large rRNAs will dominate the RNA-seq data. There are two main types of rRNA Remove Kits in the markets: (1) RiboMinus Kit from Invitrogen or MICROBExpress Bacterial mRNA Enrichment Kit (formerly Ambion, now Invitrogen), which have two probes located within the conserved sequence region at each ends of 16S and 23S rRNAs. Full-length rRNA or partial rRNA that pairs with these probes are removed. This implies that such RNA-seq data will lack reads mapped to the 5’ or 3’ ends of ssu rRNAs. The other type of rRNA removal kit is represented by the Ribo-Zero Kit from Epicentre (an Illumina company). This kit removes rRNA across the entire length and does not specifically targets the 5’ and 3’ ends. We used ARSDA (Xia 2017) to confirm that transcriptomic studies using this RNA removal kit have reads that map to the 3’ end of ssu rRNA.

\section*{ACKNOWLEDGMENTS}
We thank J. Wang, M.-A. Akimenko, A. Golshani, and T. Xing for discussion and comments. This study was funded by the Discovery Grant from Natural Science and Engineering Research Council of Canada to X.X. (RGPIN/261252-2013).

\section*{LITERATURE CITED}

Agresti, A., 2002 \textit{Categorical Data Analysis}. Wiley, New Jersey.

Bamford, D. H., J. Caldentey, and J. K. Bamford, 1995 Bacteriophage PRD1: a broad host range DSDNA tectivirus with an internal membrane. Adv. Virus Res. 45: 281–319.

Band, L., and D. J. Henner, 1984 \textit{Bacillus subtilis} requires a ‘stringent’ Shine-Dalgarno region for gene expression. DNA 3: 17–21.

Britton, R. A., T. Wen, L. Schaefer, O. Pellegrini, W. C. Uicker et al., 2007 Maturation of the 5’ end of \textit{Bacillus subtilis} 16S rRNA by the essential ribonuclease YqgC/RNase J1. Mol. Microbiol. 63: 127–138.

Brosius, J., M. L. Palmer, P. J. Kennedy, and H. F. Noller, 1978 Complete nucleotide sequence of a 16S ribosomal RNA gene from \textit{Escherichia coli}. Proc. Natl. Acad. Sci. USA 75: 4801–4805.

Bulmer, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics 129: 897–907.

Chakravorty, S., D. Helb, M. Burday, N. Connell, and D. Alland, 2007 A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J. Microbiol. Methods 69: 330–339.

Chiu, B. K., G. Gronau, U. Mader, B. Hessling, D. Becher et al., 2011 S-bacillithiolation protects against hypochlorite stress in \textit{Bacillus subtilis} as revealed by transcriptomics and redox proteomics. Mol. Cell. Proteomics 10: M111.009506.

Chithambaram, S., R. Prabhakaran, and X. Xia, 2014 The effect of mutation and selection on codon adaptation in \textit{Escherichia coli} bacteriophage. Genetics 197: 301–315.

Claridge, J. E., III, 2004 Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin. Microbiol. Rev. 17: 840–862.

Davies, B. W., C. Kohrer, A. I. Jacob, L. A. Simmons, J. Zhu et al., 2010 Role of \textit{Escherichia coli} YbeY, a highly conserved protein, in rRNA processing. Mol. Microbiol. 78: 506–518.

Dunn, J. E., E. Buzash-Pollert, and F. W. Studier, 1978 Mutations of bacteriophage T7 that affect initiation of synthesis of the gene 0.3 protein. Proc. Natl. Acad. Sci. USA 75: 2741–2745.

Duval, M., A. Korepanov, O. Fuchsauer, P. Fechter, A. Haller et al., 2013 \textit{Escherichia coli} ribosomal protein S1 unloads structured mRNAs onto the ribosome for active translation initiation. PLoS Biol. 11: e1001731.

Eckhardt, H., and R. Luhrmann, 1979 Blocking of the initiation of protein biosynthesis by a pentanucleotide complementary to the 3’ end of \textit{Escherichia coli} 16 S RNA. J. Biol. Chem. 254: 11185–11188.

Farwell, M. A., M. W. Roberts, and J. C. Rabinowitz, 1992 The effect of ribosomal protein S1 from \textit{Escherichia coli} and \textit{Micrococcus luteus} on protein synthesis in vitro by \textit{E. coli} and \textit{Bacillus subtilis}. Mol. Microbiol. 6: 3375–3383.

Gilbert, J., S. O’Donnell, W. J. Etten, and G. R. Janssen, 2012 A 5’-terminal phosphate is required for stable ternary complex formation and translation of leaderless mRNA in \textit{Escherichia coli}. RNA 18: 508–518.

Gold, L., D. Prinbnow, T. Schneider, S. Shinedling, B. S. Singer et al., 1981 Translational initiation in prokaryotes. Annu. Rev. Microbiol. 35: 365–403.

Grahn, A. M., S. J. Butcher, J. K. H. Bamford, and D. H. Bamford, 2006 PRD1: dissecting the genome, structure and entry, pp. 176–185 in \textit{The Bacteriophages}, edited by Calendar, R. Oxford University Press, Oxford.

Hager, P. W., and J. C. Rabinowitz, 1985 Translational specificity in \textit{Bacillus subtilis}, pp. 1–29 in \textit{The Molecular Biology of the Bacilli}, edited by Dubnau. Academic Press, New York.

Hothorn, T., K. Hornik, M. A. van de Wiel, and A. Zeileis, 2006 A Lego system for conditional inference. Am. Stat. 60: 257–263.

Hothorn, T., K. Hornik, M. A. van de Wiel, and A. Zeileis, 2008 Implementing a class of permutation tests: the coin package. J. Stat. Softw. 28: 1–23.

Hui, A., and H. A. de Boer, 1987 Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in \textit{Escherichia coli}. Proc. Natl. Acad. Sci. USA 84: 4762–4766.

Jacob, A. I., C. Kohrer, B. W. Davies, U. L. RajBhandary, and G. C. Walker, 2013 Conserved bacterial RNase YbeY plays key roles in 70S ribosome quality control and 16S rRNA maturation. Mol. Cell 49: 427–438.

Kitahara, K., Y. Yasutake, and K. Miyazaki, 2012 Mutational robustness of 16S rRNA, shown by experimental horizontal gene transfer in \textit{Escherichia coli}. Proc. Natl. Acad. Sci. USA 109: 19220–19225.

Komarova, A. V., L. S. Tchufistova, E. V. Supina, and I. V. Boni, 2002 Protein S1 counteracts the inhibitory effect of the extended Shine-Dalgarno sequence on translation. RNA 8: 1137–1147.

Krishnan, K. M., W. J. Van Etten, III, and G. R. Janssen, 2010 Proximity of the start codon to a leaderless mRNA’s 5’ terminus is a strong positive determinant of ribosome binding and expression in \textit{Escherichia coli}. J. Bacteriol. 192: 6482–6485.

Kudla, G., A. W. Murray, D. Tollervey, and J. B. Plotkin, 2009 Coding-sequence determinants of gene expression in \textit{Escherichia coli}. Science 324: 255–258.

Kurata, T., S. Nakanishi, M. Hashimoto, M. Taoka, Y. Yamazaki et al., 2015 Novel essential gene involved in 16S rRNA processing in \textit{Escherichia coli}. J. Mol. Biol. 427: 955–965.

Liljenström, H., and G. von Heijne, 1987 Translation rate modification by preferential codon usage: intragenic position effects. J. Theor. Biol. 124: 43–55.

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

Lu, P., C. Vogel, R. Wang, X. Yao, and E. M. Marcotte, 2007 Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotechnol. 25: 117–124.
