G3 Genes | Genomes | Genetics

INVESTIGATION

RNA-Seq-Based Analysis Reveals Heterogeneity in Mature 16S rRNA 3′ Termini and Extended Anti-Shine-Dalgarno Motifs in Bacterial Species

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

*Ottawa Institute of Systems Biology, Ottawa, Ontario, Canada K1H 8M5

ORCID IDs: 0000-0002-1578-1933 (J.R.S.); 0000-0003-1218-2086 (Y.W.); 0000-0002-3092-7566 (X.X.)

ABSTRACT We present an RNA-Seq based approach to map 3′ end sequences of mature 16S rRNA (3′ TAIL) in bacteria with single-base specificity. Our results show that 3′ TAILs are heterogeneous among species; they contain the core CCUCC anti-Shine-Dalgarno motif, but vary in downstream lengths. Importantly, our findings rectify the mis-annotated 16S rRNAs in 11 out of 13 bacterial species studied herein (covering Cyanobacteria, Deinococcus-Thermus, Firmicutes, Proteobacteria, Tenericutes, and Spirochaetes). Furthermore, our results show that species-specific 3′ TAIL boundaries are retained due to their high complementarity with preferred Shine-Dalgarno sequences, suggesting that 3′ TAIL bases downstream of the canonical CCUCC motif play a more important role in translation initiation than previously reported.

KEYWORDS
Gene Expression
RNA-Seq
Translation
Initiation

Understanding bacterial translation is important to pharmaceutical industries seeking to optimize protein biosynthesis (Xia 2018a). In this process, the rate-limiting step is generally considered to be initiation (Kudla *et al.* 2009; Tuller *et al.* 2010; Xia 2015) and the most prominently cited mechanism of initiation in bacteria (Shine and Dalgarno 1974, 1975) involves an interaction between a pyrimidine-rich anti-Shine-Dalgarno (aSD) sequence at the 3′ end of the 16S rRNA (3′ TAIL) and a purine-rich Shine-Dalgarno (SD) sequence in the mRNA translation initiation region (TIR) of protein coding genes. Pairing between these two sequences helps the ribosome dock near the start codon.

Efficient SD-mediated translation initiation requires optimal SD:aSD binding location and pairing potential (Schurr *et al.* 1993; Osterman *et al.* 2013; Prabhakaran *et al.* 2015; Abolbaghaei *et al.* 2017; Hockenberry *et al.* 2017; Wei *et al.* 2017). The canonical core aSD motif, CCUCC, is widely believed to elevate initation efficiency because of its strong complementarity with SD sequences and conservation across phyla (Shine and Dalgarno 1974; Woese *et al.* 1975; Schurr *et al.* 1993; Starmer *et al.* 2006; Vimberg *et al.* 2007; Nakagawa *et al.* 2010). Yet what constitutes ideal SD:aSD complementarity remains a subject of debate. Some researchers contend that there is weak association between SD:aSD binding affinity and initiation efficiency (Li *et al.* 2012), but others suggest that intermediate binding affinities optimize initiation efficiency in *Escherichia coli* and *Bacillus subtilis* when a broader range of SD:aSD interactions is considered (Vimberg *et al.* 2007; Osterman *et al.* 2013; Hockenberry *et al.* 2017). Furthermore, when a SD sequence that binds to the *B. subtilis* 3′ TAIL is substituted with a shorter SD sequence pairing with *E. coli*'s 3′ TAIL, interferon plasmids' expression levels decrease drastically (Band and Henner 1984). These findings emphasize the importance of characterizing the full extent of the 3′ TAIL.

The 3′ TAIL boundary remains ambiguous for most bacterial species because the precise 3′ maturation process of the 16S precursor sequence remains unclear (Sulthana and Deutscher 2013; Deutscher 2015), and only a few mature 16S rRNA sequences have been experimentally verified (Woese *et al.* 1980). Consequently, determination of the 16S rRNA is frequently automated based on sequence similarity (Lin *et al.* 2008; Nakagawa *et al.* 2010). However, this process is often unreliable (Starmer *et al.* 2006; Jones *et al.* 2007; Lagesen *et al.* 2007; Lin *et al.* 2008) and many such 16S ribosomal RNA sequence annotations have been discontinued in NCBI's Gene database. For example, 16S rRNA

doi: https://doi.org/10.1534/g3.118.200729
Manuscript received September 12, 2018; accepted for publication October 21, 2018; published Early Online October 24, 2018.

Supplemental material available at Figshare: https://doi.org/10.25387/g3.7081094.
[1]Equal contribution.
[2]Corresponding author: Department of Biology, University of Ottawa, 30 Marie Curie, Ottawa, Ontario, Canada, K1N 6N5. Tel: (613) 562-5800 ext. 6886, Fax: (613) 562-5486. E-mail: xxia@uottawa.ca

entries for *Streptococcus pyogenes* (NC_002737), *Bacillus anthracis* (NC_005945), and *Legionella pneumophila* (NC_005823) are all truncated such that their annotated 3′ ends do not encompass the canonical CCUCC motif.

To circumvent the aforementioned problem, we devise strategies to map RNA transcripts from high-throughput RNA sequencing (RNA-Seq) data (Lister *et al.* 2008; Wang *et al.* 2009; Anders *et al.* 2013) to the 16S rDNA genomic sequence with single base specificity. The feasibility of this approach was shown recently in a study (Wei *et al.* 2017) where we successfully recovered the *E. coli* and *B. subtilis* 3′ TAILs documented in literature (Shine and Dalgarno 1974; Woese *et al.* 1975). Our present objective is to advance our RNA-Seq framework to characterize the 3′ TAIL in any bacterial species, especially those that have not been experimentally verified.

The challenge associated with our approach is the limited availability of suitable data. There is a complete lack of publicly available RNA-Seq data in GEO DataSets for many species, such as *Acidithiobacillus ferrooxidans*, *Microcystis aeruginosa*, *Shigella flexneri*, and *Yersinia pestis*. Furthermore, many experiments remove rRNAs prior to sequencing (O'Neil *et al.* 2013) in an effort to enrich the target RNA molecules, such as mRNAs (Choi and Hagedorn 2003). Fortunately, our findings suggest that ribo-depletion is often incomplete, and enough 16S rRNA reads will persist to allow for 3′ TAIL characterization. The inclusion of 13 species studied herein (covering Cyanobacteria, Deinococcus-Thermus, Firmicutes, Proteobacteria, Tenericutes, and Spirochaetes) is thus predicated on the availability of usable RNA-Seq datasets in NCBI's GEO (Edgar *et al.* 2002) database (see Materials and Methods for additional details). Additionally, the availability of protein abundance data in PaxDb (Wang *et al.* 2012, 2015) for all species studied allow us to investigate the effect of SD:aSD complementarity on protein production in real genes.

Comprehensive comparative sequence analyses (Nakagawa *et al.* 2010, 2017) claim 5′-CCUCCU-3′ is the functionally constrained 3′ TAIL terminus. In other words, the motif is conserved among bacterial species because it pairs with SD sequences effectively. However, several bases further downstream are conserved in the genomic sequences of closely related species. We suspect that this is the result of functional constraint imposed by the SD:aSD interaction further downstream of 5′-CCUCCU-3′. Accordingly, we hypothesize that downstream bases are retained in 3′ TAILs because they effectively interact with species-specific SD sequences as previously observed for *E. coli* and *B. subtilis* (Band and Henner 1984; Abolbaghaei *et al.* 2017; Wei *et al.* 2017).

Our findings corroborate previous studies suggesting that intermediate binding affinity is preferred (Osterman *et al.* 2013; Hockenberry *et al.* 2017; Wei *et al.* 2017). The 3′ termini downstream of the core CCUCC are heterogeneous among species, but fall within the conserved boundary at the genomic level. Furthermore, terminal bases are preferred in SD:aSD binding in most species, albeit having weaker binding affinity than CCUCC. These findings demonstrate the importance of considering bases downstream of CCUCC in SD:aSD binding.

MATERIALS AND METHODS

Processing genomic and RNA-Seq data

The annotated genomes of 26 species in GenBank formats were retrieved from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov). Next, the NCBI annotated 16S rRNA was retrieved. In the case where multiple 16S rRNA entries exist, the first one listed is selected.

High-throughput RNA-Seq SRX runs of wildtype species were downloaded from GEO DataSets in FASTQ format. The FASTQ files were first converted to FASTQ+ format using ARSDA 1.1 (Xia 2017), grouping identical reads under a single ID while also indicating the number of copies (SeqID_# of copies), in order to reduce the size of the datasets prior to adapter trimming. The FASTQ+ data were then processed using CutAdapt 1.17 (Martin 2011) to trim off the 3′ flanking adapter sequences. In experiments that use the oligo(dT)-adapter primer, RNA fragments are first poly-adenylated at the 3′ end, we thus set CutAdapt to recognize "AAAAA". In others that use specific sets of primers ligated to random hexamers, we set CutAdapt to recognize all possible adapters in the kits' index, with 10% error rate. Regardless of whether poly-As or barcode adapters were trimmed, we only retained reads that were 25 nt or longer after the trimming process to mitigate bias in expression levels (Williams *et al.* 2016). Next, we used Trimmomatic 0.38 (Bolger *et al.* 2014) to remove poor quality sequences with average Phred scores lower than 20 (1% probability of a base calling error) (Ewing and Green 1998). Since adapters were trimmed after reads were grouped in FASTQ+ format, sequences that were previously unique due to the presence of adapter nucleotides may become identical (such as for SeqGr176560_1 and SeqGr558077_1, Figure 1d). The processed FASTQ+ datasets were subsequently converted into FASTA format for multiple sequence alignment.

Aligning RNA-Seq reads to annotated rRNA sequences

We next mapped reads in the FASTA files onto the 16S rDNA genomic sequence. The FASTA+ files were converted into BLAST databases using the "Create BLAST DB" function in ARSDA. The BLAST query sequence was selected using genomic sequences 100 nt upstream and downstream of the core CCUCC motif (205 nt total query length). For each species, the query sequence was searched against BLAST databases using the BLAST function (Altschul *et al.* 1990) implemented in ARSDA. We used an E-value cutoff of 10^{-5} (with the exception of *Bacillus anthracis*, for which we used an E-value of 10^{-3} due to the relatively shorter average read length and smaller database size) paired with a minimum word length of 12 to balance the quantity and quality of hits, as well as search speed, against the >= 25 nt reads in the ribo-depleted datasets. Then, sequence hits were retrieved from the FASTA files using seqtk (Li 2012) and complementary strand sequences were eliminated. Finally, remaining hits were aligned to the query sequence using multiple sequence alignment (Clustal Omega algorithm (Sievers and Higgins 2014) implemented in DAMBE, default parameters).

Determining putative SD sequences based on pairing potential, location, and binding affinity

For each species, our characterized 3′ TAILs (Table 1) were used as the complementary sequence in identifying putative SD sequences. To ensure that determined putative SD sequences are from real genes, we map protein IDs in PaxDb 4.0 (Wang *et al.* 2012, 2015) to Gene IDs in NCBI and only use CDSs that have protein expressions. Using DAMBE7 (Xia 2018b), we followed the methods used in previous studies (Nussinov *et al.* 1978; Waterman and Smith 1978): 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. Site-specific observed and expected aSD usage values were retrieved from the DAMBE when SD sequences are determined.

DATA AVAILABILITY

Supplementary file S1 contains RNA-Seq BLAST hits and file S2 contains the list of genes with protein abundance data that were used to determine putative SD sequences in all species studied; Figure S1 contain the 3′ TAIL map for the remaining 11 species. Supplemental material available at Figshare: https://doi.org/10.25387/g3.7081094.

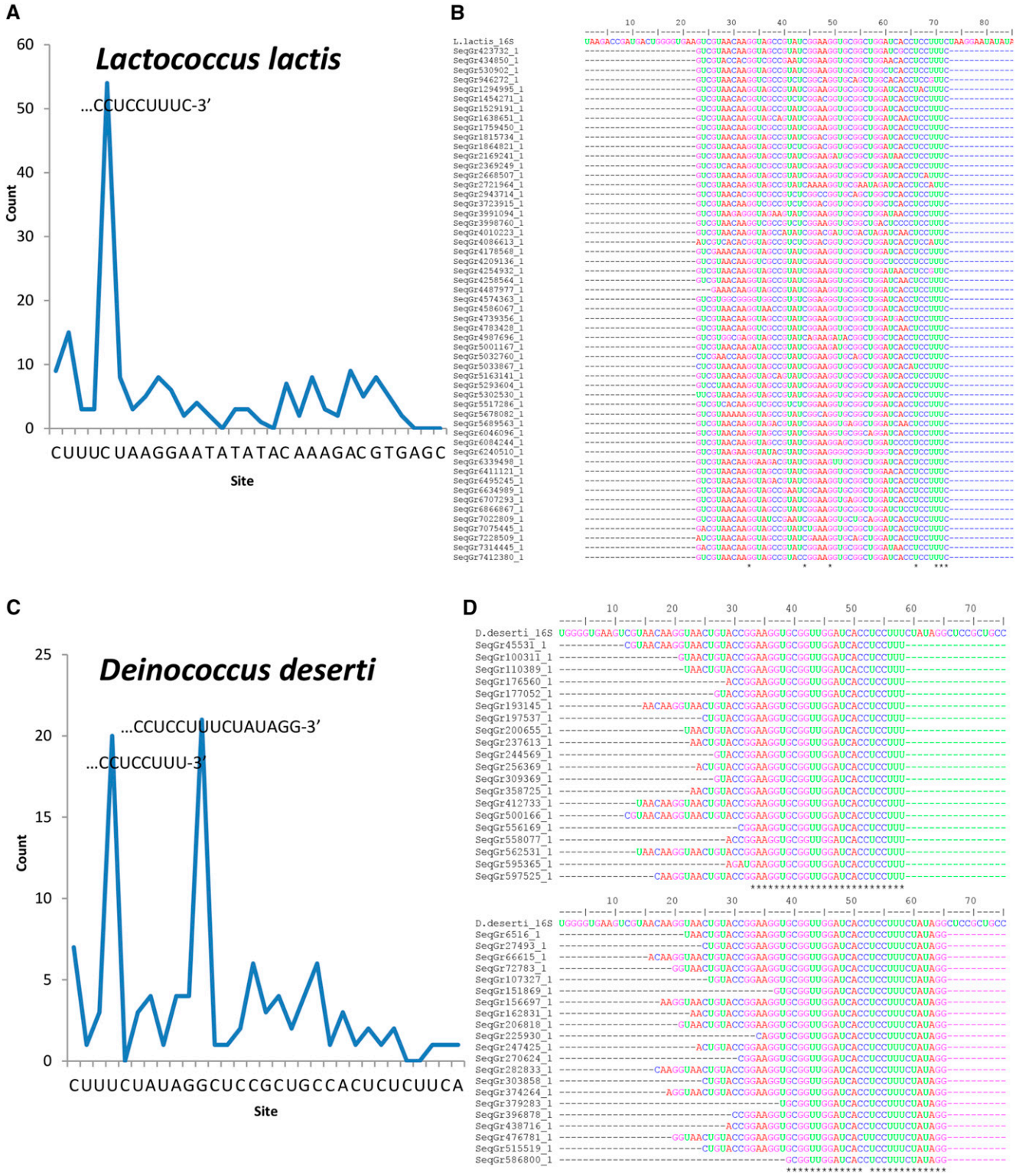

Figure 1 The count of mapped 3′ ends of RNA-Seq reads (A, C) and sequence alignments (B, D) for *Lactococcus lactis* and *Deinococcus deserti*. Mapped regions start with the last C of CCUCC as the first site, extended by 30 nt downstream. The 3′ ends of sequence alignments represent local reads mapped to the single major peak in *L. lactis* and the two major peaks in *D. deserti*. The complete length of the query genomic sequence is 205 nt long.

Table 1 The RNA-Seq corrected 3′ TAIL in 13 bacterial species. RNA-Seq determined 3′ TAILs are shaded gray. The NCBI annotated 3′ TAILs are in black fonts, extensions revealed by RNA-Seq data are underlined and ambiguities in bold

Species	16S 3′ TAIL	Putative pre-16S rRNA	NCBI accession	SRA accession
Listeria monocytogenes	GAUCACCUCCUUUCU		NC_003210	SRX2771238-41
Streptococcus pyogenes*	GAUCACCUCCUUUCU		NC_002737	SRX3036007, 08, 10, 11
Lactococcus lactis*	GAUCACCUCCUUUC		NC_002662	SRX2140913
Bacillus anthracis*	GAUCACCUCC		NC_005945	SRX129739
Neisseria meningitidis*	GAUCACCUCCUUUCU**A**†		NC_003112	SRX2005108, 10
Clampylobacter jejuni*	GAUCACCUCCUUUC		NC_002163	SRX326863
Deinococcus deserti*	GAUCACCUCCUUUCUA	GAUCACCUCCUUUCUAUAGG	NC_012526	SRX497284
Mycoplasma pneumoniae*	GAUCACCUCCUUUCUAAUGGAG	GAUCACCUCCUUUCUAAUGGAG	NC_017504	SRX1122953
Salmonella enterica*	GAUCACCUCCUUA		NC_003198	SRX2409112, 3
Legionella pneumophila*	GAUCACCUCC	GAUCACCUCCUUACAUAGAAAGGCAC	NC_002942	SRX041877
Desulfovibrio vulgaris*	GAUCACCUCCUU		NC_002937	SRX066256
Leptospira interrogans	GAACACCUCCUUUUUAAGGAG	GAACACCUCCUUUUUAAGGAGAAUCAAAGG	NC_005823	SRX2448245-52
Synechocystis sp.*	GAUCACCUCCUUAAGGG		NC_000911	SRX2694285-8

*Species whose characterized 3′ TAIL differ from NCBI annotation.
†The use of poly-adenylated data makes it difficult to determine whether the terminal nucleotide is U or A in this case.

RESULTS AND DISCUSSION

Characterizing the 3′ TAIL in bacteria using an improved RNA-Seq-based approach

We improve upon our method of 3′ TAIL characterization (Wei *et al.* 2017) by processing the RNA-Seq data more rigorously. To ensure quality and single-base specificity for reads mapped to a reference genomic sequence, we used CutAdapt (Martin 2011) to trim adapters flanking raw RNA-Seq reads because these sequences obscure the true end of RNA fragments (see Materials and Methods for more detail). We subsequently filtered out poor quality reads by discarding those with average Phred scores ≤ 20 using Trimmomatic (Bolger *et al.* 2014); in other words, we retained reads with average base-calling error rates of $< 1\%$ (Ewing and Green 1998). A caveat of using poly-adenylated RNA-Seq datasets for *Neisseria meningitidis* is that we cannot distinguish between 5′-CCUCCUUUCU-3′ and 5′-CCUCCUUUCUA-3′ as the 3′ TAIL; it is unclear whether the first adenosine is associated with the 3′ TAIL or the poly-A chain (Table 1).

To map the 16S rRNA, we generated a BLAST library using the quality filtered datasets and performed ungapped local similarity search using BLAST (Altschul *et al.* 1990) between RNA-Seq reads and a 205 nt genomic sequence with the canonical CCUCC motif at the center (100 nt extending from each side). We next aligned the BLAST hits by multiple sequence alignment (Clustal Omega algorithm (Sievers and Higgins 2014) implemented in DAMBE (Xia 2018b), default parameters) against the reference genomic sequence. In all species, we define the terminus of the 3′ TAIL using two criteria: 1) it must contain the canonical CCUCC, and 2) it is the most mapped site at or near CCUCC. The underlying assumption for the second criterion is that the mature 16S rRNA is more abundant than precursor transcripts, as is the case in *E. coli* (Cangelosi and Brabant 1997), because precursors are continuously degraded by exoribonucleases (Sulthana and Deutscher 2013).

The 3′ TAIL termini are heterogeneous but functionally constrained

Following our two criteria, we have characterized the 3′ TAIL in 13 out of 26 species in PaxDb (Table 1). Figure 1 shows the sequence map and alignments for *Lactococcus lactis* and *Deinococcus deserti*. The sequences mapped for the 11 others are present in Supplementary Figure S1. Two others, *E. coli* and *B. subtilis*, were previously determined (Wei *et al.* 2017). The remaining 11 species could not be characterized because of the aforementioned absence of data in four species (*Acidithiobacillus ferrooxidans*, *Microcystis aeruginosa*, *Shigella flexneri*, and *Yersinia pestis*), and because no convincing peaks were observed in the region of interest (up to 30 nt downstream of CCUCC) in the remaining seven species (*Bacterioides thetaiotaomicron*, *Bateonella henselae*, *Helicobacter pylori*, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, and *Shewanella oneidensis*), likely due to effective ribo-depletion in their RNA-Seq datasets. We considered a peak to be convincing when the counts mapping to that site were at least 3 fold higher than background (counts of any four flanking sites on either side). Importantly, in the characterized 13 species, we made corrections to annotations in eight species (NC_002662 *L. lactis*, NC_002163 *Clamylobacter jejuni*, NC_000911 *Synechocystis* sp., NC_003112 *N. meningitidis*, NC_*012526 D. deserti*, NC_017504 *Mycoplasma pneumoniae*, NC_003198 *Salmonella enterica*, NC_002937 *Desulfovibrio vulgaris*), and redefined the 3′ TAIL in three others (NC_002737 *S. pyogenes*, NC_005945 *B. anthracis*, and NC_002942 *L. pneumophila*) that were certainly mis-annotated due to their failure to incorporate the canonical CCUCC. Resultantly, the annotated 3′ TAILs of only two out of 13 species, NC_003210 *Listeria*

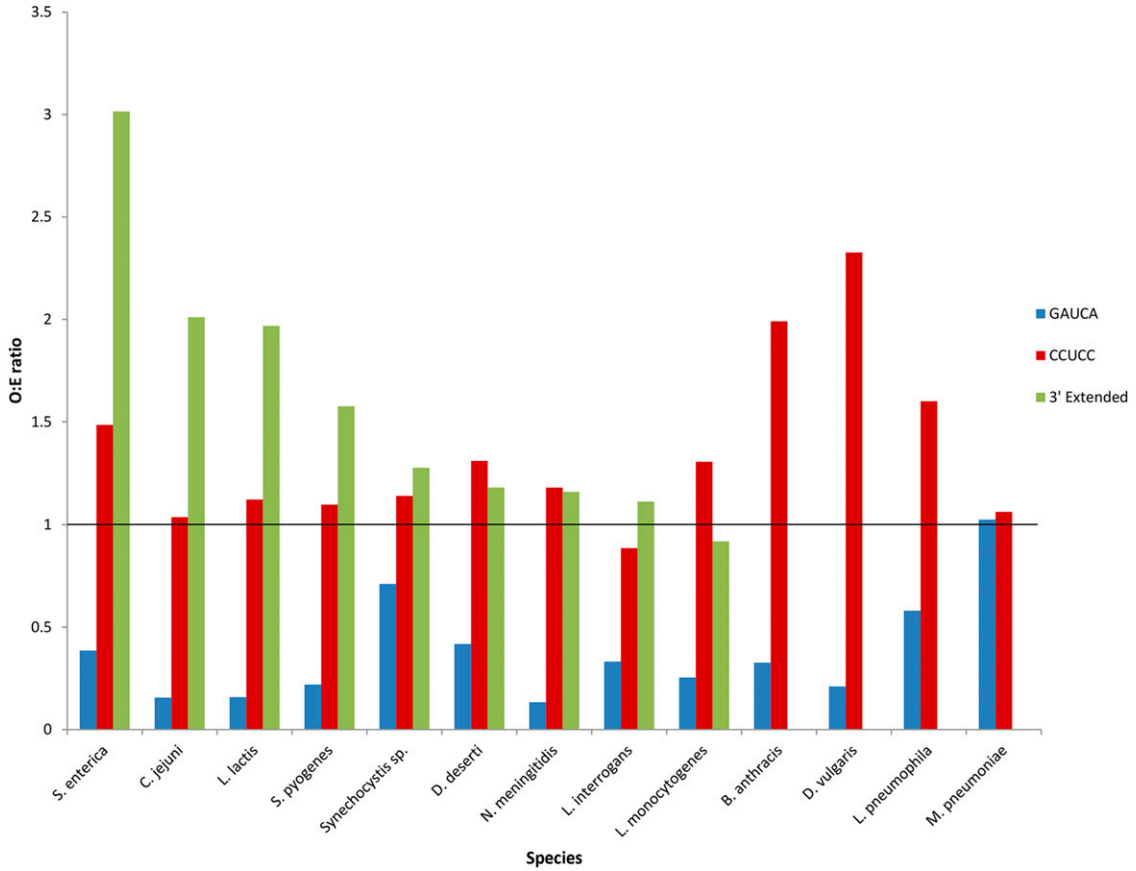

Figure 2 Mean ratio of observed over expected SD:aSD complementarity (O:E ratio) in 13 species at conserved 5′-GAUCA-3′ (blue), and 5′-CCUCC-3′ motifs (red). The average O:E ratio is also shown for the characterized sequences downstream of CCUCC (green) in the nine bacterial species that have extended ends.

monocytogenes and NC_005823 *Leptospira interrogans* were left unchanged. In short, the 3′ TAILs can variably extend up to six bases downstream of CCUCC in the majority of species studied.

The 3′ TAILs vary among species, but bases downstream of the CCUCC motif are conserved among bacteria. The 16S genomic sequences are largely conserved for several bases beyond CCUCC: *e.g.*, 5′-GAUCACCUCCUUUCUA-3′ in Bacilli and 5′-GAUCACCUCCUUA-3′ in Beta- and Gammaproteobacteria. This conservation suggests that the regions downstream of CCUCC may also be important in SD:aSD pairing. Importantly, 3′ TAIL terminal bases downstream of CCUCC are species-specific, but do not extend past the conserved genomic boundaries, in all species studied except in *L. interrogans*. In other words, the 3′ TAIL falls variably short of 5′-GAUCACCUCCUUUCUA-3′. This finding further suggests that both CCUCC and downstream bases are conserved regions that are preferred in SD:aSD pairing in most species.

To offer a plausible reason for the unexpected length of the 3′ TAIL in *L. interrogans*, it is worth noting that the dependence on the SD:aSD interaction for efficient translation is dynamic (Nakagawa *et al.* 2017). In genes that have strong secondary structure within the TIR, ribosome recruitment is facilitated by RPS1 (Nakagawa *et al.* 2010; Osterman *et al.* 2013). This protein binds U-rich regions (Boni *et al.* 1991; Komarova *et al.* 2005) to unfold double-stranded RNA (Qu *et al.* 2012; Duval *et al.* 2013). Furthermore, RPS1's domains appear to be under higher functional constraint in species possessing few SD-containing genes, such as *L. interrogans* (Nakagawa *et al.* 2010, 2017); the reliance on RPS1 reduces the dependence on a SD:aSD interaction and may relax 3′ TAIL boundary constraints.

Notably, four species (*D. deserti*, *M. pneumoniae*, *L. pneumophila*, and *L. interrogans*) have a secondary peak of mapped reads within 20 nt downstream of CCUCC (Figure 1, Table 1, Supplementary Fig. S1). We propose that the secondary peak farther downstream is the pre-16S rRNA; it is too far downstream of CCUCC to be considered as the mature 16S rRNA 3′ end based on sequence conservation (Nakagawa *et al.* 2010). The prominence of this second peak may be due to the accumulation of the endoribonuclease cleaved pre-16S rRNA intermediate, because the localization of exoribonuclease to this precursor sequence is a rate limiting step. However, the intermediate sequence is rapidly continuously degraded once it is targeted by these enzymes. This would explain the lack of sequences mapped between the mature 16S rRNA and the intermediate sequence (the two peaks) (Figure 1c, Supplementary Fig. S1).

The 3′ TAIL terminal bases are preferred in SD:aSD binding

We define an aSD site to be preferred if the observed number of times the base is involved in SD pairing is greater than expected. In the absence of SD usage bias, a putative SD sequence of 4 to 12 nt can be expected to pair anywhere within the boundary of the aSD sequence, as long as complete complementarity is achieved. Here, we designate the aSD sequence to constitute the 3′ TAIL, beginning with the conserved 5′-GAUCA-3′, followed by the core motif CCUCC, and extended by variable lengths of

terminal bases characterized herein (Table 1, *e.g.*, in *L. monocytogenes*, the aSD sequence is 5′GAUCACCUCCUUUCU-3′ and the terminal bases are 5′-UUUCU-3′). Then, taking *L. monocytogenes* as example, the maximum number of possible pairs at the first complementary aSD site (*aSD_1*) by the total pool of 4 nt to 12 nt putative SD sequences is calculated by equation (1), with N_m denoting N number of putative SD sequences of length m:

$$aSD\_1 = \sum_{m=4}^{12} \frac{N_m}{15 - m + 1} \quad (1)$$

However, the number of possible base-pairs resulting in perfect complementarity varies. For example, a 12 nt putative SD sequence may start pairing at the first, but not the sixth, base on a complementary aSD sequence that is 15 nt long, and the maximum usage of the sixth aSD site (*aSD_6*) is calculated instead by equation (2):

$$aSD\_6 = 4\times\frac{N_4}{12} + 5\times\frac{N_5}{11} + 6\times\frac{N_6}{10} + 6\times\frac{N_7}{9} + 6\times\frac{N_8}{8} + 6\times\frac{N_9}{7} + 6\times\frac{N_{10}}{6} + 6\times\frac{N_{11}}{5} + 5\times\frac{N_{12}}{4} \quad (2)$$

The expected usage is then calculated by taking the relative proportions of maximum usage at each site (adding up to 1) multiplied by the total number of observed putative SD sequences of various lengths. These calculations are implemented in DAMBE (Xia 2018b) under the 'Analyze 5UTR' function.

As defined, a preferred aSD site will have an observed to expected SD:aSD usage ratio (O:E) > 1. Since expected SD:aSD count is calculated in absence of any selection bias in SD usage, an O:E > 1 suggests presence of selection bias in observed SD usage. Figure 2 shows an average O:E > 1 at CCUCC for all species, with the exception of *L. interrogans*. Indeed, CCUCC is preferentially used in SD:aSD pairing. Meanwhile, average O:E is <1 for 5′-GAUCA-3′ in all species except *M. pneumoniae*; hence, this conserved region is avoided by SD:aSD pairing in most species. Lastly, in keeping with expectations, conserved regions downstream of CCUCC have an O:E > 1 in all species except *L. monocytogenes*. These observations indicate that downstream bases are retained because they are preferred in SD:aSD binding, despite their weaker binding affinity than CCUCC. This result corroborates recent studies suggesting that intermediate SD:aSD complementarity increase initiation efficiency (Osterman *et al.* 2013; Hockenberry *et al.* 2017; Wei *et al.* 2017).

In this study we present an RNA-Seq based approach to characterize the 3′ end of mature 16S rRNA in bacterial species across different lineages. There is weaker 3′ TAIL conservation at the RNA level than at the DNA level. Nonetheless, the presence of 3′ termini bases downstream of CCUCC falls within the conserved boundary at the genomic level. Furthermore, the usage of terminal bases is favored in SD:aSD binding. Alternatively, RPS1-mediated initiation may relax the functional constraint at the 3′ TAIL of *L. interrogans*, explaining its exceptional length. Our findings complement previous studies investigating the role of CCUCC in translation initiation and suggest that transcribed bases downstream of this canonical motif also play an important role in translation efficiency.

ACKNOWLEDGMENTS

This work was supported by the Discovery Grant of Natural Science and Engineering Research Council of Canada to X.X. (NSERC, RGPIN/2018-03878), and the Ontario Graduate Scholarship 2018-2019 to Y.W. The manuscript was substantially improved by the comments of two anonymous reviewers, and we are grateful for their insight.

LITERATURE CITED

Abolbaghaei A., J. R. Silke, and X. Xia, 2017 How Changes in Anti-SD Sequences Would Affect SD Sequences in *Escherichia coli* and *Bacillus subtilis*. *G3 Genes|Genomes|Genetics (Bethesda)* 7: 1607–1615. https://doi.org/10.1534/g3.117.039305

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Anders, S., D. J. McCarthy, Y. Chen, M. Okoniewski, G. K. Smyth *et al.*, 2013 Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat. Protoc. 8: 1765–1786. https://doi.org/10.1038/nprot.2013.099

Band, L., and D. J. Henner, 1984 *Bacillus subtilis* Requires a "Stringent" Shine-Dalgarno Region for Gene Expression. DNA 3: 17–21. https://doi.org/10.1089/dna.1.1984.3.17

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Boni, I. V., D. M. Isaeva, M. L. Musychenko, and N. V. Tzareva, 1991 Ribosome-messenger recognition: mRNA target sites for ribosomal protein S1. Nucleic Acids Res. 19: 155–162. https://doi.org/10.1093/nar/19.1.155

Cangelosi, G. A., and W. H. Brabant, 1997 Depletion of pre-16S rRNA in starved *Escherichia coli* cells. J. Bacteriol. 179: 4457–4463. https://doi.org/10.1128/jb.179.14.4457-4463.1997

Choi, Y. H., and C. H. Hagedorn, 2003 Purifying mRNAs with a high-affinity eIF4E mutant identifies the short 3′ poly(A) end phenotype. Proc. Natl. Acad. Sci. USA 100: 7033–7038. https://doi.org/10.1073/pnas.1232347100

Deutscher, M. P., 2015 Twenty years of bacterial RNases and RNA processing: how we've matured. RNA 21: 597–600. https://doi.org/10.1261/rna.049692.115

Duval, M., A. Korepanov, O. Fuchsbauer, P. Fechter, A. Haller *et al.*, 2013 *Escherichia coli* Ribosomal Protein S1 Unfolds Structured mRNAs Onto the Ribosome for Active Translation Initiation. PLoS Biol. 11: e1001731. https://doi.org/doi: 10.1371/journal.pbio.1001731

Edgar, R., M. Domrachev, and A. E. Lash, 2002 Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30: 207–210. https://doi.org/10.1093/nar/30.1.207

Ewing, B., and P. Green, 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8: 186–194. https://doi.org/10.1101/gr.8.3.186

Hockenberry, A. J., A. R. Pah, C. J. Jewett, and L. A. N. Amaral, 2017 Leveraging genome-wide datasets to quantify the functional role of the anti-Shine-Dalgarno sequence in regulating translation efficiency. Open Biol. 7: 160239. http://dx.doi.org/10.1098/rsob.160239

Jones, C. E., A. L. Brown, and U. Baumann, 2007 Estimating the annotation error rate of curated GO database sequence annotations. BMC Bioinformatics 8: 170. https://doi.org/10.1186/1471-2105-8-170

Komarova, A. V., L. S. Tchufistova, M. Dreyfus, and I. V. Boni, 2005 AU-Rich Sequences within 5′ Untranslated Leaders Enhance Translation and Stabilize mRNA in *Escherichia coli*. J. Bacteriol. 187: 1344–1349. https://doi.org/10.1128/JB.187.4.1344-1349.2005

Kudla G., A. W. Murray, D. Tollervey, and J. B. Plotkin, 2009 Coding-Sequence Determinants of Gene Expression in *Escherichia coli*. *Science* 80: 255–258. https://doi.org/10.1126/science.1170160

Lagesen, K., P. Hallin, E. A. Rodland, H. H. Staerfeldt, T. Rognes *et al.*, 2007 RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35: 3100–3108. https://doi.org/10.1093/nar/gkm160

Li, G.-W., E. Oh, and J. S. Weissman, 2012 The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484: 538–541. https://doi.org/10.1038/nature10965

Li H., 2012 seqtk Toolkit for processing sequences in FASTA/Q formats.

Lin, Y. H., B. C. Chang, P. W. Chiang, and S. L. Tang, 2008 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

Lister, R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry *et al.*, 2008 Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133: 523–536. https://doi.org/10.1016/j.cell.2008.03.029

Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal; Vol 17, No 1 Next Gener. Seq. Data Anal. https://doi.org/10.14806/ej.17.1.200

Nakagawa, S., Y. Niimura, K. Miura, and T. Gojobori, 2010 Dynamic evolution of translation initiation mechanisms in prokaryotes. Proc. Natl. Acad. Sci. USA 107: 6382–6387. https://doi.org/10.1073/pnas.1002036107

Nakagawa, S., Y. Niimura, and T. Gojobori, 2017 Comparative genomic analysis of translation initiation mechanisms for genes lacking the Shine–Dalgarno sequence in prokaryotes. Nucleic Acids Res. 45: 3922–3931. https://doi.org/10.1093/nar/gkx124

Nussinov, R., G. Pieczenik, J. R. Griggs, and D. J. Kleitman, 1978 Algorithms for Loop Matchings. SIAM J. Appl. Math. 35: 68–82. https://doi.org/10.1137/0135006

O'Neil, D., H. Glowatz, and M. Schlumpberger, 2013 Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Curr. Protoc. Mol. Biol.* Chapter 4: Unit 4.19. https://doi.org/10.1002/0471142727.mb0419s103

Osterman, I. A., S. A. Evfratov, P. V. Sergiev, and O. A. Dontsova, 2013 Comparison of mRNA features affecting translation initiation and reinitiation. Nucleic Acids Res. 41: 474–486. https://doi.org/10.1093/nar/gks989

Prabhakaran, R., S. Chithambaram, and X. Xia, 2015 *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

Qu, X., L. Lancaster, H. F. Noller, C. Bustamante, and I. Tinoco, 2012 Ribosomal protein S1 unwinds double-stranded RNA in multiple steps. Proc. Natl. Acad. Sci. USA 109: 14458–14463. https://doi.org/10.1073/pnas.1208950109

Schurr, T., E. Nadir, and H. Margalit, 1993 Identification and characterization of *E.coli* ribosomal binding sites by free energy computation. Nucleic Acids Res. 21: 4019–4023. https://doi.org/10.1093/nar/21.17.4019

Shine, J., and L. Dalgarno, 1974 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. https://doi.org/10.1073/pnas.71.4.1342

Shine, J., and L. Dalgarno, 1975 Determinant of cistron specificity in bacterial ribosomes. Nature 254: 34–38. https://doi.org/10.1038/254034a0

Sievers, F., and D. G. Higgins, 2014 Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol. Biol. 1079: 105–116. https://doi.org/10.1007/978-1-62703-646-7_6

Starmer, J., A. Stomp, M. Vouk, and D. Bitzer, 2006 Predicting Shine–Dalgarno Sequence Locations Exposes Genome Annotation Errors. PLOS Comput. Biol. 2: e57. https://doi.org/10.1371/journal.pcbi.0020057

Sulthana, S., and M. P. Deutscher, 2013 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

Tuller, T., Y. Y. Waldman, M. Kupiec, and E. Ruppin, 2010 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

Vimberg, V., A. Tats, M. Remm, and T. Tenson, 2007 Translation initiation region sequence preferences in *Escherichia coli*. BMC Mol. Biol. 8: 100. https://doi.org/10.1186/1471-2199-8-100

Wang, Z., M. Gerstein, and M. Snyder, 2009 RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10: 57–63. https://doi.org/10.1038/nrg2484

Wang, M., M. Weiss, M. Simonovic, G. Haertinger, S. P. Schrimpf *et al.*, 2012 PaxDb, a Database of Protein Abundance Averages Across All Three Domains of Life. Mol. Cell. Proteomics 11: 492–500. https://doi.org/10.1074/mcp.O111.014704

Wang, M., C. J. Herrmann, M. Simonovic, D. Szklarczyk, and C. von Mering, 2015 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

Waterman, M. S., and T. F. Smith, 1978 RNA secondary structure: a complete mathematical analysis. Math. Biosci. 42: 257–266. https://doi.org/10.1016/0025-5564(78)90099-8

Wei, Y., J. R. Silke, and X. Xia, 2017 Elucidating the 16S rRNA 3′ boundaries and defining optimal SD/aSD pairing in *Escherichia coli* and *Bacillus subtilis* using RNA-Seq data. Sci. Rep. 7: 17639. https://doi.org/10.1038/s41598-017-17918-6

Williams, C. R., A. Baccarella, J. Z. Parrish, and C. C. Kim, 2016 Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics 17: 103. https://doi.org/10.1186/s12859-016-0956-2

Woese, C. R., G. E. Fox, L. Zablen, T. Uchida, L. Bonen *et al.*, 1975 Conservation of primary structure in 16S ribosomal RNA. Nature 254: 83–86. https://doi.org/10.1038/254083a0

Woese, C. R., L. J. Magrum, R. Gupta, R. B. Siegel, D. A. Stahl *et al.*, 1980 Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. Nucleic Acids Res. 8: 2275–2293. https://doi.org/10.1093/nar/8.10.2275

Xia, X., 2015 A Major Controversy in Codon-Anticodon Adaptation Resolved by a New Codon Usage Index. Genetics 199: 573–579. https://doi.org/10.1534/genetics.114.172106

Xia X., 2017 ARSDA: A New Approach for Storing, Transmitting and Analyzing Transcriptomic Data. *G3 Genes|Genomes|Genetics (Bethesda)* 7: 3839–3848. https://doi.org/10.1534/g3.117.300271

Xia, X., 2018a Bioinformatics and translation initiation, pp. 173–195 in *Bioinformatics and the Cell*: modern computational approaches in genomics, proteomics and transcriptomics. Springer, Cham.

Xia, X., 2018b DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. Mol. Biol. Evol. 35: 1550–1552. https://doi.org/10.1093/molbev/msy073

Communicating editor: D. Baltrus