Characterization of 16S rRNA 3' Termini Using RNA-Seq Data

Jordan Silke

Supervisor: Dr. Xuhua Xia

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
In partial fulfillment of the requirements for a
M.Sc. degree from the
Ottawa-Carleton Institute of Biology

Thèse soumise à la
Faculté des Etudes Supérieures et Postdoctorales
Université d'Ottawa
En vue de l'obtention de la maîtrise ès sciences
L'Institut de Biologie d'Ottawa-Carleton

Abstract

Optimizing the production of useful macromolecules from transgenic microorganisms is crucial to biopharmaceutical companies. Improving bacterial growth and replication depends largely on the efficiency of translation, which is rate-limited by initiation. Among the most important interactions between the mRNA translation initiation region (TIR) and the translation machinery is the association between the Shine-Dalgarno (SD) sequence in the TIR and the complementary anti-SD (aSD) sequence which is located within a short unstructured segment that includes the 3' terminus (3' TAIL) of the mature 16S rRNA. However, the mature 3' TAIL has been poorly characterized in the majority of bacteria, rendering optimal SD/aSD pairing unclear in these species.

In light of this, we established a novel strategy to characterize the mature 3' TAILs of bacterial species that leverages the availability of publically stored RNA sequencing (RNA-Seq) data. In chapter 2, we devised a RNA-Seq-based approach to successfully recover the experimentally verified 3' TAIL in *E. coli* (5'-CCUCCUUA-3') and resolve inconsistencies surrounding the identity of the 3' TAIL in *Bacillus subtilis*. In chapter 3 we improve the method introduced in chapter 2 to clearly and more reliably define the 3' TAIL termini for 13 bacterial species with available protein abundance data.

Our results reveal considerable heterogeneity in the termini of 3' TAILs among closely related species and that sites downstream of the canonical CCUCC aSD motif are more important to initiation than previously believed. My research contributes to advance our understanding in microbial translation efficiency in two significant ways: 1) providing an RNA-Seq-based approach to characterize rRNA transcripts, and 2) elucidating optimal recognition between protein-coding genes and the rRNA translation machinery.

Résumé

L'optimisation de la production de macromolécules à partir de microorganismes transgéniques est important pour les sociétés biopharmaceutiques. L'amélioration de la croissance et réplication bactérienne dépend, en grande partie de l'efficacité de la traduction, qu'est limitée par l'initiation. L'association entre la séquence Shine-Dalgarno (SD) de la région TIR et la séquence complémentaire anti-SD (aSD), située dans une courte séquence non structurée avec une terminaison 3' (3' TAIL) de l'ARNr 16S mature, est l'une des interactions les plus importantes pour la région d'initiation de la traduction de l'ARNm (TIR) et la machinerie de traduction. La plupart de queue 3 ' mature des bactéries sont mal caractérisées, cela rend difficile l'appariement optimal de SD / aSD chez ces espèces.

À la lumière de cette information, nous avons établi une nouvelle stratégie pour caractériser les matures queues 3' d'espèces bactériennes en utilisant les séquençages d'ARN (RNA-Seq) disponible au public. Au chapitre 2, nous avons développé avec succès une approche basée sur l'ARN-Seq pour récupérer la queue 3' expérimentalement vérifiée d'E.coli (5'-CCUCCUUA-3') et nous avons résolu les incohérences d'identité de la queue 3' du Bacillus subtilis. Au chapitre 3, nous avons amélioré la méthode introduite en chapitre 2. Nous avons défini clairement et précisément la queue 3' des 13 espèces bactériennes avec une abondance d'information disponible sur leurs protéines.

Nos résultats révèlent une hétérogénéité considérable dans l'extrémité des queues 3' d'espèces proches. Nous avons aussi remarqué que les sites en aval du motif canonique de la CCUCC sont plus importantes pour l'initiation que nous l'espérions. Ma recherche aide à améliorer notre compréhension de l'efficacité de la traduction microbienne par deux manières significatives: 1) fournir une approche basée sur l'ARN-Seq pour caractériser les transcrits d'ARNr, et 2) élucider la reconnaissance optimale entre les gènes codant pour les protéines et la machinerie de traduction de l'ARNr.

Acknowledgements

It is with the utmost respect that I acknowledge the support and guidance of my supervisor, Dr. Xuhua Xia. The completion of this thesis would not have been possible if not for his research acumen and unwavering belief in me. I am truly grateful to have had the opportunity to work in the Xia lab in pursuit of my Master's degree. I would like to thank all of my lab mates, past and present, for your constructive feedback on my research.

I would be remiss if I failed to recognize the influence that other professors have had on my work. I thank my committee members Dr. Linda Bonen and Dr. Douglas Johnson for their advice and understanding. I would also like to thank Dr. Stéphane Aris-Brosou and Dr. Nicolas Rodrigue for their informative and interesting bioinformatics courses.

I would not be where I am today without the love and support of my family and friends. I am especially grateful to my grandparents (Ray and Irene), my aunt and uncle (Cathy and Syl), my mother (Joanne), and my younger brother (Liam). It is with special consideration that I mention the rest of my brothers, not in blood, but in bond. Tyler and Trevor Schofield, David Beauchesne, and Yulong Wei have stood by me through everything and taught me that true strength lies in unity. They have been my most indispensable allies through times of prosperity and adversity alike, and I lack the words to properly describe my gratitude towards them.

In loving memory of Irene Silke
December 12, 1941 – October 31, 2018

INVICTUS

OUT OF THE NIGHT THAT COVERS ME,
BLACK AS THE PIT FROM POLE TO POLE,
I THANK WHATEVER GODS MAY BE
FOR MY UNCONQUERABLE SOUL.

IN THE FELL CLUTCH OF CIRCUMSTANCE
I HAVE NOT WINCED NOR CRIED ALOUD.
UNDER THE BLUDGEONINGS OF CHANCE
MY HEAD IS BLOODY, BUT UNBOWED.

BEYOND THIS PLACE OF WRATH AND TEARS
LOOMS BUT THE HORROR OF THE SHADE,
AND YET THE MENACE OF THE YEARS
FINDS AND SHALL FIND ME UNAFRAID.

IT MATTERS NOT HOW STRAIT THE GATE,
HOW CHARGED WITH PUNISHMENTS THE SCROLL,
I AM THE MASTER OF MY FATE,
I AM THE CAPTAIN OF MY SOUL.

-William Ernest Henley

List of Publications

A. Publications related to my thesis:

1) Akram Abolbaghaei, Jordan R. Silke, and Xuhua Xia (2017). How Changes in Anti-SD Sequences Would Affect SD Sequences in *Escherichia coli* and *Bacillus subtilis*. *G3: Genes|Genomes|Genetics* 7(5).
2) Yulong Wei, Jordan R. Silke, and Xuhua Xia (2017). Elucidating the 16S rRNA 3' boundaries and defining optimal SD/aSD pairing in Escherichia coli and Bacillus subtilis using RNA-Seq data. *Scientific Reports* 7(1).
3) Jordan R. Silke, Yulong Wei, and Xuhua Xia (2018). RNA-Seq-Based Analysis Reveals Heterogeneity in Mature 16S rRNA 3' Termini and Extended Anti-Shine-Dalgarno Motifs in Bacterial Species. *G3: Genes|Genomes|Genetics* 8(12).

B. Other publications:

1) Karen Massel, Jordan R. Silke, and Linda Bonen (2016). Multiple splicing pathways of group II trans-splicing introns in wheat mitochondria. *Mitochondrion* 28 p. 23-32.
2) Yulong Wei, Jordan R. Silke, and Xuhua Xia (2019). An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. *Scientific Reports* 9(1):3184.

C. Poster presentations:

1) Jordan Silke and Xuhua Xia: "Characterization of novel hypothetical proteins in bacteriophage λ". Ottawa-Carleton Institute of Biology Symposium. April 27-28, 2017. Ottawa, ON, Canada.
2) Jordan Silke and Xuhua Xia: "Characterization of novel hypothetical proteins in bacteriophage λ". Ontario Ecology, Ethology and Evolution Colloquium. May 18-20, 2017. Kingston, ON, Canada. Honourable mention.
3) Jordan Silke, Yulong Wei, and Xuhua Xia: "Determination of bacterial 16S rRNA 3' termini using RNA-Seq". Ontario Ecology, Ethology and Evolution Colloquium. May 10-12, 2018. London, ON, Canada.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

LSU – large ribosomal subunit
SSU – small ribosomal subunit
RPS1 – ribosomal protein S1
TIR – translation initiation region
3' TAIL – the free nucleotides at the 3' end of the 16S rRNA component of SSU
UTR – untranslated region
SD – Shine-Dalgarno
aSD – anti-Shine-Dalgarno
RNA-Seq – RNA sequencing
HEG – highly expressed gene
LEG – lowly expressed gene
nt – nucleotides
DAMBE – Data Analysis in Molecular Biology and Evolution
ARSDA – Analyzing RNA-Seq Data
NCBI – National Center for Biotechnology Information
$D_{toStart}$ – distance (in nucleotides) to the start codon
PaxDb – Protein Abundance Across Organisms
Database GEO – Gene Expression Omnibus RBS –
ribosome binding site
ΔG – the change in Gibbs free energy
indel – insertion or deletion

Chapter One

Introduction

1.0 Bacterial translation initiation

The ribosome is a complex of RNA and protein components made up of two subunits: a 50S large subunit (LSU) and a 30S small subunit (SSU). The translation initiation process begins when the 30S SSU attaches to the messenger RNA (mRNA) at its ribosomal binding site (RBS) located in a translation initiation region (TIR) that includes the start codon and generally has weak secondary structure (Iserentant and Fiers 1980; Laursen *et al.* 2005). This unstructured TIR is generally flanked by more highly structured regions that may facilitate ribosome localization (Iserentant and Fiers 1980; Osterman *et al.* 2013). This initiation process is generally regarded to be the rate-limiting step of translation, which means that the interaction between the RBS and SSU will bottleneck protein production unless it is highly efficient, in which case elongation (driven by tRNA-mediated selection) exerts the most influence on polypeptide biosynthesis (Liljenstrom and von Heijne 1987; Bulmer 1991; Kudla *et al.* 2009).

1.0.1 Role of the SSU in Initiation

When discussing how translation initiation takes place in bacterial species, arguably the most important aspect to consider is the various interactions between the SSU and the TIR that act to streamline the process. The nature of these interactions has been best described in *Escherichia coli* (Czernilofsky *et al.* 1975), and a number of experiments in this model organism suggest that three components of the SSU are particularly important to initiation: including ribosomal proteins S21 (Held *et al.* 1974; Van Duin and Wijnands 1981; Stern *et al.* 1988) and S1

(de Smit and van Duin 1994; Komarova *et al.* 2002; Vimberg *et al.* 2007; Qu *et al.* 2012; Duval *et al.* 2013) as well as the 16S rRNA (Shine and Dalgarno 1974, 1975; Schurr *et al.* 1993; Chen *et al.* 1994). Although the S1 protein and the 16S rRNA have roles that are detailed in literature, S21 remains more enigmatic, but there is evidence that demonstrates the importance of S21 in start codon recognition. The role of S21 in initiation is facilitated by its proximity to the single stranded 3' terminal nucleotides, immediately downstream of the highly conserved helix 45 (Shine and Dalgarno 1974; Cannone *et al.* 2002; Baumgardt *et al.* 2018), of the 16S rRNA (3' TAIL) (Czernilofsky *et al.* 1975).

Early experiments that focused on S21 inactivation (Van Duin and Wijnands 1981) and ribosomal reconstitution in the absence of S21 (Held *et al.* 1974; Van Duin and Wijnands 1981) established that S21 is required to maximize ribosomal activity. Both studies were able to demonstrate that the presence of S21 improved the association between the SSU and mRNA. Furthermore, they demonstrated that SSUs lacking S21 exhibited a ~40% reduction in activity relative to those with S21. It is important to note that this reduction in efficiency was observed only with endogenously produced transcripts (Van Duin and Wijnands 1981). Despite its association with elevated ribosomal activity, the underlying mechanism is not fully understood.

Conversely, the S1 protein is known to facilitate initiation through helicase activity which unwinds double stranded RNA segments within structured TIRs by binding to U-rich regions near the RBS. This S1-mediated initiation mechanism (Boni *et al.* 1991; de Smit and van Duin 1994; Laursen *et al.* 2005; Vimberg *et al.* 2007; Qu *et al.* 2012; Duval *et al.* 2013) can

directly improve start codon accessibility by the ribosomal P-site by reducing intramolecular interactions that would otherwise obscure the RBS.

Likewise, there is a well-documented interaction between a (typically purine-rich) section of TIR upstream of the start codon known as the Shine-Dalgarno (SD) sequence and a (correspondingly pyrimidine-rich) segment of the 3' TAIL termed the anti-SD (aSD) sequence. The effectiveness of this SD/aSD interaction is considered to be one of the most important factors in efficient initiation (Shine and Dalgarno 1974, 1975; Steitz and Jakes 1975; Taniguchi and Weissmann 1978; Schurr *et al.* 1993; Vimberg *et al.* 2007). Due to its unique ability to position S1 and S21 to act on the TIR combined with its intrinsic capacity to form meaningful SD/aSD bonds that enable initiation, the 16S rRNA will be a central focus moving forward.

1.0.2 The *rrn* operon

This operon encodes the 5S, 16S, and 23S rRNAs all separated by spacer regions that often include elements such as tRNAs (Green *et al.* 1985). Given the importance of an operon encoding components that are critical to cellular metabolism, it should come as no surprise that the copy number of this operon is often variable between species or even different isolates of a given species (Amin *et al.* 2018). This copy number variability has been shown to affect the fitness of organisms (Stevenson and Schmidt 2004); generally, high *rrn* operon copy number (more than five copies) is associated with improved fitness relative to bacteria with fewer copies (Klappenbach *et al.* 2000). This may be, in part, due to the abundance of polymorphisms that exist between these copies (Hakovirta *et al.* 2016). Notably, these polymorphisms do not impact the 3' TAILs of bacteria in the vast majority of cases (Nakagawa *et al.* 2010, 2017; Amin

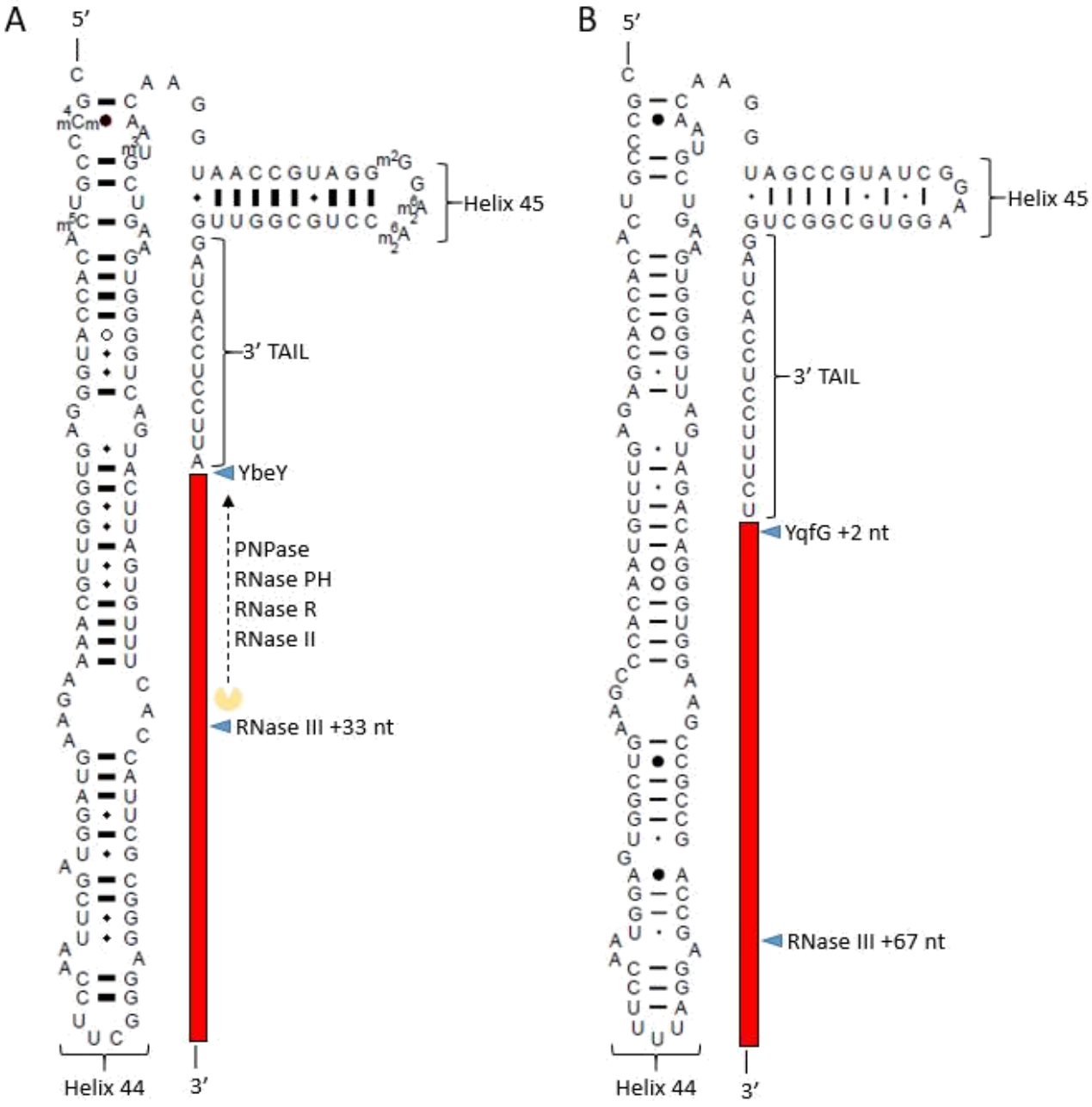
et al. 2018). In contrast, although the *rrn* operon is almost always genomically encoded due to its importance, *Aureimonas* sp. and a subset of its close relatives are known to only possess a single plasmid-bound copy (Anda *et al.* 2015).

Even in extreme cases such as *Aureimonas* sp., the *rrn* operon maintains synteny between its three rDNA components. Due to this co-transcription, the *rrn* transcript is subject to multiple processing steps by a number of RNases in order to generate mature rRNA and tRNA molecules (Sulthana and Deutscher 2013; Jacob *et al.* 2013; Baumgardt *et al.* 2018) that can participate in the translation process.

1.0.3 16S rRNA maturation

Many of the key players in processing the 16S rRNA are documented in *E. coli* and *Bacillus subtilis*. Strikingly, the RNases involved in the process differ widely between these species despite generating mature 3' termini that are quite similar (Sulthana and Deutscher 2013; Baumgardt *et al.* 2018). Figure 1.1 summarizes the inferred structure (Cannone *et al.* 2002) at the 3' end of the 16S rRNA as well as the RNases that participate in its maturation (Baumgardt *et al.* 2018) for both *E. coli* and *B. subtilis*. It is important to note that our structural understanding of the 3' TAIL and its participation in SD/aSD binding is based on what has been previously established in *E. coli* (Shine and Dalgarno 1974, 1975). Indeed, the 3' TAIL lacks experimental validation in the vast majority of bacteria, including all of those not covered herein or previously characterized by Woese and colleagues (Uchida *et al.* 1974; Woese *et al.* 1980). This means that the specific structure in towards the 3' terminus is more difficult to

Figure 1.1. Structural model of the 3' termini of 16S rRNA in A) *E. coli* and B) *B. subtilis*. The schematic begins with helix 44 and proceeds through helix 45, the 3' TAIL, and into the downstream spacer region. Base pairing is denoted by black bars (Watson-Crick), dots (neutral G-U), open circles (A-G), and filled circles (pairings between modified bases). Known methylation sites in *E. coli* are designated. Endonuclease sites (blue triangles) and exonucleases (gold partial circle) that participate in 3' end maturation are labelled alongside the nucleotide position they are known to act on relative to the 3' TAIL terminal nucleotide. Red bars represent the spacer regions after the 3' termini. Adapted from the structural models curated by the Gutell lab (Cannone *et al.* 2002) on the Comparative RNA Web Site and Project (http://www.rna.icmb.utexas.edu/).


A
5′
Helix 45
3′ TAIL
YbeY
PNPase
RNase PH
RNase R
RNase II
RNase III +33 nt
Helix 44
3′
B
5′
Helix 45
3′ TAIL
YqfG +2 nt
RNase III +67 nt
Helix 44
3′

predict, particularly with respect to indels or substitutions that are observable in a subset of classified bacteria (Amin *et al.* 2018).

1.0.4 Shine-Dalgarno sequence-mediated initiation

The highly structured 16S rRNA in the bacterial SSU terminates at the 3' end with a short (~13 nt) 3' TAIL that is free to base pair with the TIR. In the SD-mediated initiation, the aSD sequence in the 3' TAIL base pairs with the SD sequence which is often found in the 5' UTR; this effectively docks the ribosome near the start codon to facilitate translation initiation. Early experimental studies have shown that the 3' TAIL differs among bacterial species (Shine and Dalgarno 1975; Woese *et al.* 1975). In a previous collaboration (Abolbaghaei *et al.* 2017), I have shown that such differences in the 3' TAIL lead to species-specific SD usage due to SD/aSD location constraints. This stresses that SD sequences that pair well to the 3' TAIL with one species will not necessarily have the same binding affinity to a different 3' TAIL in another species. Moreover, SD/aSD pair preference differs between highly and lowly expressed genes (HEGs and LEGs respectively) within a species. These observations lead to my recent collaboration performing RNA-Seq-based analyses to examine SD and aSD sequence usage in order to better explain the varying degree of influence SD-mediated initiation exerts on gene expression across bacterial lineages.

1.0.5 Limitations associated with SD-mediated initiation

The efficiency of SD sequence-mediated initiation is driven up by two factors: 1) SD/aSD sequence binding location and 2) SD/aSD binding affinity (Shine and Dalgarno 1974; Chen *et al.* 1994; Osterman *et al.* 2013; Prabhakaran *et al.* 2015). However, these two components are

poorly understood because what constitutes the aSD sequence is often mis-characterized, even in well studied species such as *B. subtilis* (Abolbaghaei *et al.* 2017; Wei *et al.* 2017). The influence of SD/aSD sequence binding location has been previously established (Chen *et al.* 1994; Osterman *et al.* 2013; Prabhakaran *et al.* 2015). To more accurately measure how well this interaction localizes the ribosome to the start codon, our lab has previously established the measure $D_{toStart}$ (Prabhakaran *et al.* 2015; Abolbaghaei *et al.* 2017), which specifies the distance, in nucleotides, between the 16S rRNA 3' terminus and the start codon. This measurement improves upon the use of aligned spacing distances described previously (Chen *et al.* 1994) because $D_{toStart}$ is strongly constrained within a narrower range and considers the relative positioning of the mRNA and the ribosome. To accurately measure optimal local SD/aSD pairing location in $D_{toStart}$, it is necessary to characterize the precise 3' terminus of the 16S rRNA.

Considering SD/aSD binding affinity, CCUCC is considered to be the core aSD sequence motif due to two characteristics: 1) it has the highest binding affinity with SD sequences, and 2) it is conserved across bacteria (Shine and Dalgarno 1975; Woese *et al.* 1975; Nakagawa *et al.* 2010). However, how SD/aSD binding affinity influence translation initiation has been a subject of debate. Some researchers propose that SD sequences binding to the core CCUCC increase gene expression (Jacob *et al.* 1987; Nakagawa *et al.* 2010), yet others find no significant relationship between binding affinity and gene expression (Li *et al.* 2012). Our findings provide support for the notion that it is not strong, but intermediate SD/aSD binding affinity which best increases gene expression (Vimberg *et al.* 2007; Hockenberry *et al.* 2017). It stands to reason that bases downstream of CCUCC that are retained in the 3' TAIL must be characterized before

one can compare the gene expression levels associated with strong and weak SD/aSD binding affinities.

1.1 Characterizing RNA transcripts using RNA-Seq data

Since its emergence (Mortazavi *et al.* 2008; Lister *et al.* 2008; Wang *et al.* 2009), RNA-Seq has rose to prominence as one of the most widely used techniques to profile RNA transcripts. The approach uses isolated RNA subjected to fragmentation to generate more stable complementary DNA (cDNA) counterparts. Known adapters are then ligated to the short cDNA copies and they are amplified in preparation for sequencing. Although there are a number of proprietary RNA-Seq approaches, perhaps the most commonly used approach to sequencing by synthesis is Illumina's method (Goodwin *et al.* 2016) which generally generates reads of about 100 bases in length (Liu *et al.*, 2012). These reads can then be mapped to the transcriptome of the corresponding organism to provide a measure of transcript expression. When the number of reads associated with a given transcript is properly normalized, RNA-Seq can be used to accurately quantify gene expression between experiments with single base specificity (Wang *et al.* 2009). Given the ubiquity of RNA-Seq experiments in recent years combined with efforts to archive these valuable data, many such datasets find their way into public databases like the National Center for Biotechnology Information's (NCBI) gene expression omnibus (GEO) DataSets (Edgar *et al.* 2002).

1.1.1 A proposal to characterize rRNA sequences using RNA-Seq data

Due to that lack of structure in the free 3' TAIL, the nucleotides of interest should be relatively easy to sequence. Hence, it stands to reason that this wealth of RNA-Seq data could

be exploited in order to gain further insight into the identity of mature 3' TAIL termini in bacterial species. A caveat is that researchers often attempt to remove ribosomal RNAs in a majority of RNA-Seq experiments because rRNA is abundant, encompassing more than 80-90% (O'Neil et al. 2013) of cellular RNA content, and often will often skew the sequencing data to prevent the examination of the target molecules. In addition to the potential decrease in data availability, RNA-Seq reads must be processed for primers used in cDNA synthesis, but the 3' TAIL is relatively short (13 nt in *E. coli*) and may be lost during data processing. Hence, due to accessibility and abundance of the template, the ease with which RNA-Seq reads can recovered the 3' TAIL is specific to the RNA-Seq data. In chapter two and three, we investigated both ribo-depleted and untreated datasets and devised an effective strategy to operationally characterize the 3' TAIL with single-base specificity.

We investigated both ribo-depleted and untreated datasets and devised an effective strategy to operationally characterize the 3' TAIL with single-base specificity. We are the first to show that RNA-Seq provides a promising means to investigate mature 3' TAIL termini for two major reasons. First, primers used in cDNA synthesis can easily anneal to the unstructured 3' TAIL. This means that stringent sequencing protocols that are normally employed for highly structured transcripts such as tRNAs (Zheng et al. 2015) are not strictly required. Second, RNA is abundant, and the ribo-depletion treatments are never 100% effective. There is evidence which indicates that mature rRNAs are particularly abundant in *E. coli* (Cangelosi and Brabant 1997) due to the high demand for ribosomal components to drive protein synthesis. Our results in chapters two and three show that even such data can be meaningfully interpreted.

1.2 Significance

Our approach to define the 3' TAIL in bacteria, especially in species that are less frequently studied, is important for a number of reasons. From a methodological standpoint, our solution takes advantage of publically available data, so it can be replicated with no further wet lab work. Moreover, our implementation scales proportionally with the improvement of available RNA-Seq technologies and mapping software: as these become cheaper, faster, and more reliable, so does our approach. Another advantage of our process is that it can be applied to study other molecules, such as tRNAs (Zheng *et al.* 2015; Cozen *et al.* 2015), given the right datasets.

Mapping the precise 3' TAIL in bacterial species is important for two major related reasons. First, the overwhelming majority of 3' TAILs have not been experimentally corroborated. As such, almost all of the current annotations are based on DNA-level sequence similarity (Jones *et al.* 2007; Lin *et al.* 2008) combined with the known structure of the 3' terminus in *E. coli* (Shine and Dalgarno 1974). Additionally, it is impossible to characterize effective SD/aSD binding without first knowing the extent of the 3' TAIL, otherwise too few or too many nucleotides may be considered which can obscure biologically relevant trends. This means that a researcher will be unable to adequately determine methods to optimize the expression of particular genes unless the complete 3' TAIL, i.e. the full range of nucleotides capable of participating in SD/aSD binding, is revealed.

Importantly, when our method is extended to pathogenic species (especially those that are known to rely heavily on SD-dependent translation), characterizing the 3' TAIL can be used to better understand how their translation can be optimized. This paves the way to genetically

engineer more effective bacteriophages by optimizing the SD sequences and codon usage of phage genes to outcompete highly expressed genes of the host species by better tailoring them to interact with the host translation machinery. Such phages could be an attractive solution to dealing with bacteria that are highly resistant to multiple antibiotics when locally applied to areas of infection (Cabot *et al.* 2016).

Chapter Two

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

2.0 Comments

This chapter has been published in Scientific Reports (2017) as “Elucidating the 16S rRNA 3' boundaries and defining optimal SD/aSD pairing in *Escherichia coli* and *Bacillus subtilis* using RNA-Seq data.” YW and JS contributed equally to the analysis and interpretation of results as well as figure preparation and writing the manuscript. This research has been presented as a poster at OE3C 2018. This chapter is formatted to reflect the submission guidelines of Scientific Reports with the exception of the in text citations: these were altered to retain consistency with the rest of this document.

2.1 Abstract

Bacterial translation initiation is influenced by base pairing between the Shine-Dalgarno (SD) sequence in the 5’ UTR of mRNA and the anti-Shine-Dalgarno (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.

2.2 Introduction

Protein production is a highly controlled and optimized process in bacterial species (Li *et al.* 2014), and translation initiation is often recognized as the rate-limiting step of the translation process (Kudla *et al.* 2009; Tuller *et al.* 2010; Xia 2015). 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 (Walsh 2005). 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) (Shine and Dalgarno 1974; Hui and de Boer 1987). Furthermore, the location of the SD/aSD interaction relative to the start codon must also be considered to ensure the pairing positions of ribosomal P-site at the start codon (Shine and Dalgarno 1974; Hui and de Boer 1987; Osterman *et al.* 2013; Hockenberry *et al.* 2017).

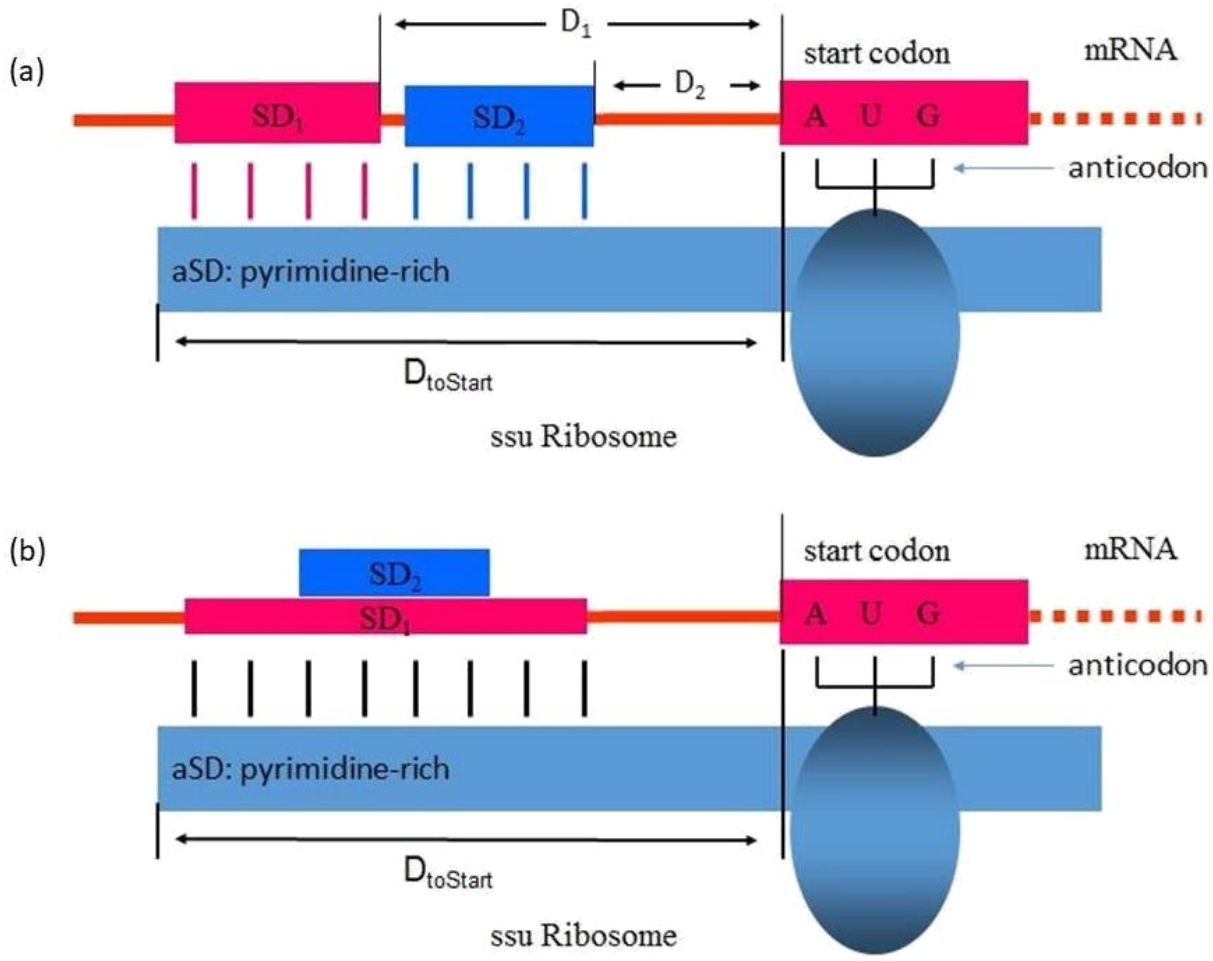
A recent model of SD/aSD interaction (Prabhakaran *et al.* 2015; Abolbaghaei *et al.* 2017) (Fig. 2.1) suggests that optimal SD/aSD pairing may depend on three factors: 1) $D_{toStart}$ (Fig. 2.1) which specifies the distance, in nucleotides, between the 16S rRNA 3' terminus and the start codon, 2) SD/aSD binding affinity (Fig. 2.1b), and 3) "leash" distance measured by D_1 and D_2 (Fig. 2.1). $D_{toStart}$ 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 (de Smit and van Duin 1994; Prabhakaran *et al.* 2015). Characterizing these features demands the precise terminus of the 16S rRNA which is often unclear, as is the case for *Bacillus subtilis*.

2.2.1 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' (Murray and Rabinowitz 1982) 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 (Murray and Rabinowitz 1982; Green *et al.* 1985) arose because Woese and colleagues published the details of their RNA sequencing method (Uchida *et al.* 1974) as well as the 3' TAILs in a number of bacterial species (Woese *et al.* 1975). Since then, alternative rDNA annotations of the *B. subtilis* 3' TAIL have emerged, including 5'-CCUCCUUUCUA-3' (NC_000964) (Barbe *et al.* 2009) and 5'-CCUCCUUUCUAA-3' (NZ_CP010052) which have been used in recent studies on *B. subtilis* 16S rRNA (Sohmen *et al.* 2015; Hockenberry *et al.* 2017). Discrepancies in these reported 3' TAILs likely arose due to the fact that multiple exoribonucleases participate in the maturation process of the 3' TAIL (Deutscher 2015). These include *RNase II*, *RNase R, PNPase* and *RNase PH* (Sulthana and Deutscher 2013), as well as *YbeY* (Jacob *et al.* 2013); hence, the 3' TAIL is continuously degraded.

Figure 2.1. Schematic model of SD/aSD interaction, illustrating (a) $D_{toStart}$ (distance, in nucleotides, between the start codon and the 3' terminus of the 16S rRNA) and the difference in the two "leash" distances D_1 and D_2 (number of nt between the 3' end of the SD sequence and the start codon) and in (b) binding affinity between two SD/aSD interactions involving SD_1 and SD_2.


(a)
D1
D2
start codon
mRNA
SD1
SD2
A U G
anticodon
aSD: pyrimidine-rich
DtoStart
ssu Ribosome
(b)
SD2
SD1
start codon
mRNA
A U G
anticodon
aSD: pyrimidine-rich
DtoStart
ssu Ribosome

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 (Mortazavi *et al.* 2008; Wang *et al.* 2009; Liu *et al.* 2012; Li *et al.* 2013) 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 (Murray and Rabinowitz 1982). 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 (Shine and Dalgarno 1974).

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.

2.2.2 Determining the optimal SD/aSD interaction that maximizes initiation efficiency

It was generally believed that high SD/aSD binding affinity facilitated translation initiation (Schurr *et al.* 1993; Ma *et al.* 2002; Starmer *et al.* 2006; Lim *et al.* 2012); accordingly, the core aSD motif (CCUCC) was characterized based on its high binding affinity (most negative

change in Gibbs free energy [G]). Furthermore, CCUCC is conserved in 277 prokaryotic species using multiple sequence alignment in MAFFT (Nakagawa *et al.* 2010). In practice, putative SD sequences are determined based on their complementarity with an extended sequence at the 3' TAIL (Li *et al.* 2012; Osterman *et al.* 2013; Li 2015; Prabhakaran *et al.* 2015; Abolbaghaei *et al.* 2017; Hockenberry *et al.* 2017): the inclusion of the core motif CCUCC is canonical, but what constitutes the full extent of the core aSD sequence remains unclear (Hockenberry *et al.* 2017).

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 (Li *et al.* 2014) 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 (Li 2015). A similar study (Hockenberry *et al.* 2017) 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 (Vimberg *et al.* 2007; Osterman *et al.* 2013) showing that SD sequences with intermediate levels of aSD (5'-ACCUCCUU-3') binding occur most frequently in *E. coli* genes (Osterman *et al.* 2013) and that six SD/aSD base pairs lead to more efficient translation and growth than shorter or longer SD/aSD pairs (Vimberg *et al.* 2007). 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.

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 (Schurr *et al.* 1993; Ma *et al.* 2002; Starmer *et al.* 2006; Lim *et al.* 2012), the corroborative reasoning that CCUCC is conserved among bacterial species is a weak one, as 5'-GAUCACCUCCU-3' is highly conserved among 277 bacterial species (Nakagawa *et al.* 2010), not just CCUCC.

2.3 Results and Discussion

2.3.1 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. 2.2a, 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. 2.2b, 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. 2.3) in both species.

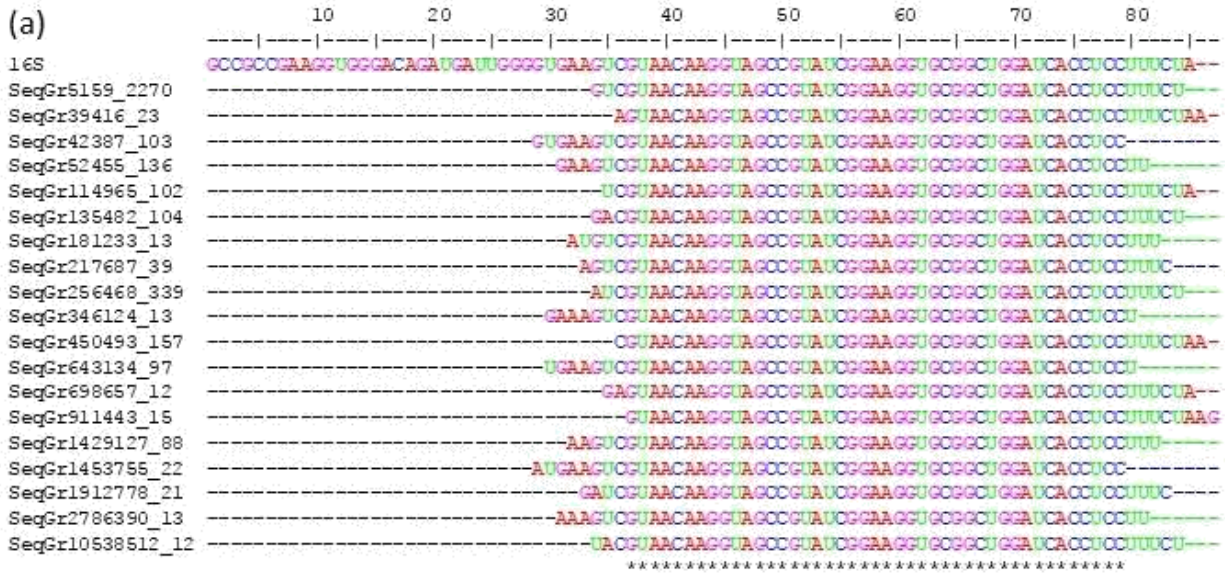
We expect to recover the universally accepted 3' terminus reported for *E. coli* (Shine and Dalgarno 1974) and, at minimum, the 5'-CCUCCUUUCU-3' end reported for *B. subtilis* (Murray and Rabinowitz 1982). In keeping with expectations, the data shows dominant usage of the

originally reported 5'-CCUCCUUUCU-3' end in *B. subtilis,* and provides no basis for the inclusion of downstream nucleotides such as A(Sohmen *et al.* 2015) (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 (Shine and Dalgarno 1974). Figure 2.3b 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 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. S2.1). 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.

Figure 2.2. 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).

(a)


10 20 30 40 50 60 70 80
16S
SeqGr5159_2270
SeqGr39416_23
SeqGr42387_103
SeqGr52455_136
SeqGr114965_102
SeqGr135482_104
SeqGr181233_13
SeqGr217687_39
SeqGr256468_339
SeqGr346124_13
SeqGr450493_157
SeqGr643134_97
SeqGr698657_12
SeqGr911443_15
SeqGr1429127_88
SeqGr1453755_22
SeqGr1912778_21
SeqGr2786390_13
SeqGr10538512_12

(b)

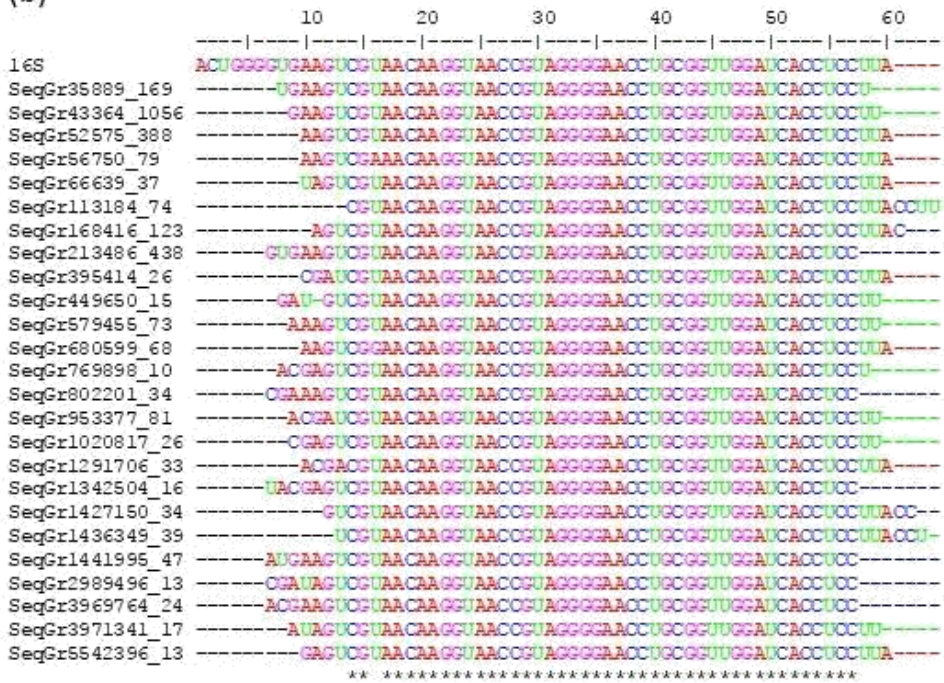
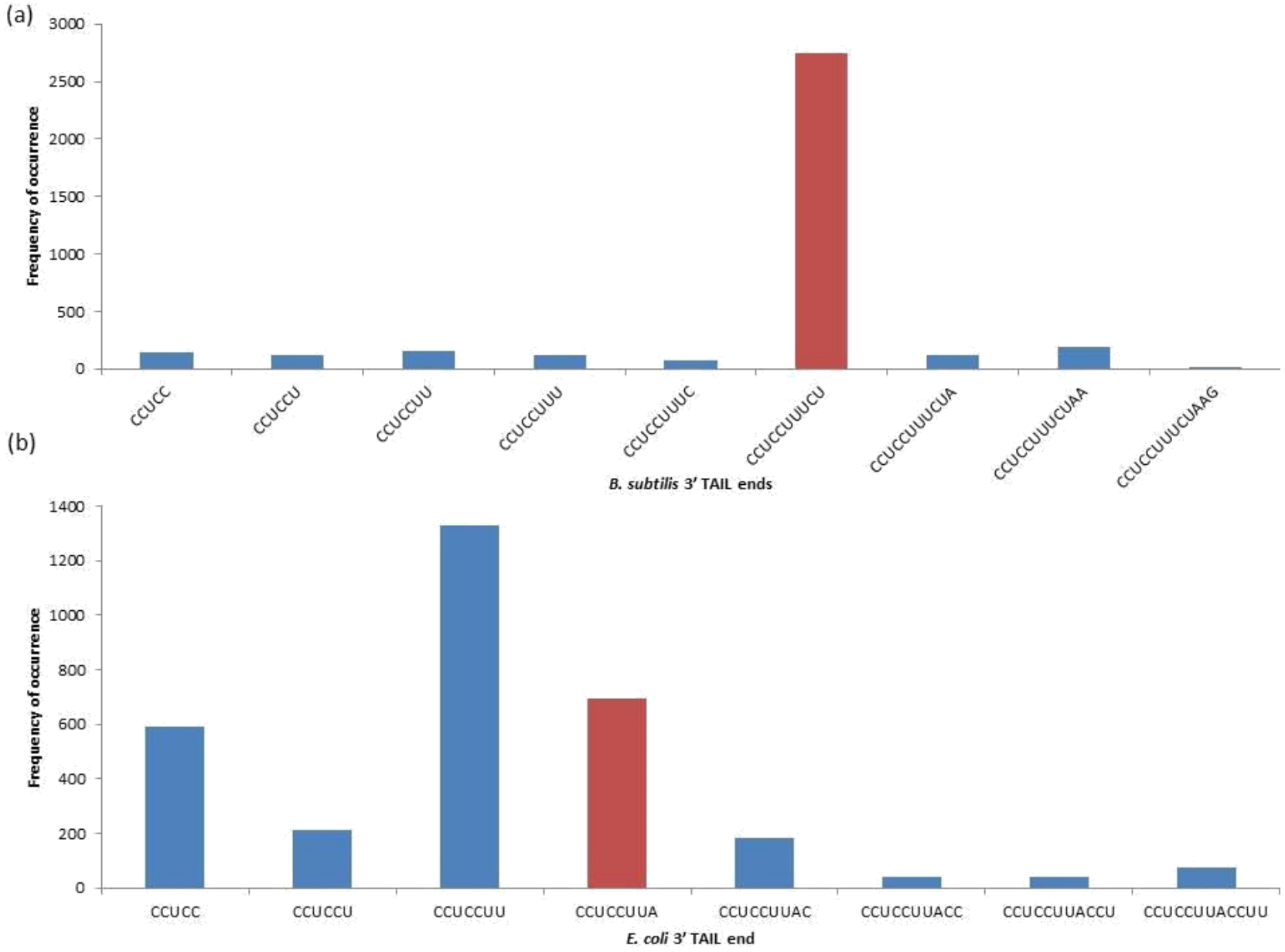

10 20 30 40 50 60
16S
SeqGr35889_169
SeqGr43364_1056
SeqGr52575_388
SeqGr56750_79
SeqGr66639_37
SeqGr113184_74
SeqGr168416_123
SeqGr213486_438
SeqGr395414_26
SeqGr449650_15
SeqGr579455_73
SeqGr680599_68
SeqGr769898_10
SeqGr802201_34
SeqGr953377_81
SeqGr1020817_26
SeqGr1291706_33
SeqGr1342504_16
SeqGr1427150_34
SeqGr1436349_39
SeqGr1441995_47
SeqGr2989496_13
SeqGr3969764_24
SeqGr3971341_17
SeqGr5542396_13

Figure 2.3. 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 with each hit count for a given terminus represented a single time. Red bars represent the frequencies associated with the first reported 3' ends in literature (Shine and Dalgarno 1974, 1975) for each species.


(a)
Frequency of occurrence
3000
2500
2000
1500
1000
500
0
CCUCC
CCUCCU
CCUCCUU
CCUCCUUU
CCUCCUUUC
CCUCCUUUCU
CCUCCUUUCUA
CCUCCUUUCUAA
CCUCCUUUCUAAG
B. subtilis 3' TAIL ends
(b)
Frequency of occurrence
1400
1200
1000
800
600
400
200
0
CCUCC
CCUCCU
CCUCCUU
CCUCCUUA
CCUCCUUAC
CCUCCUUACC
CCUCCUUACCU
CCUCCUUACCUU
E. coli 3' TAIL end

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 (Lin *et al.* 2008; Nakagawa *et al.* 2010) is inadequate. The underlying issue surrounding these instances of poor annotation is the ease with which they are propagated in automated annotation (Jones *et al.* 2007; Lagesen *et al.* 2007). Using such annotations may potentially skew conclusions in studies on translation initiation. For example, investigations considering the *B. subtilis* 3' TAIL 5'-GAUCACCUCCUUUCUA-3' (Osterman *et al.* 2013; Li *et al.* 2014; Abolbaghaei *et al.* 2017; Hockenberry *et al.* 2017) 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.

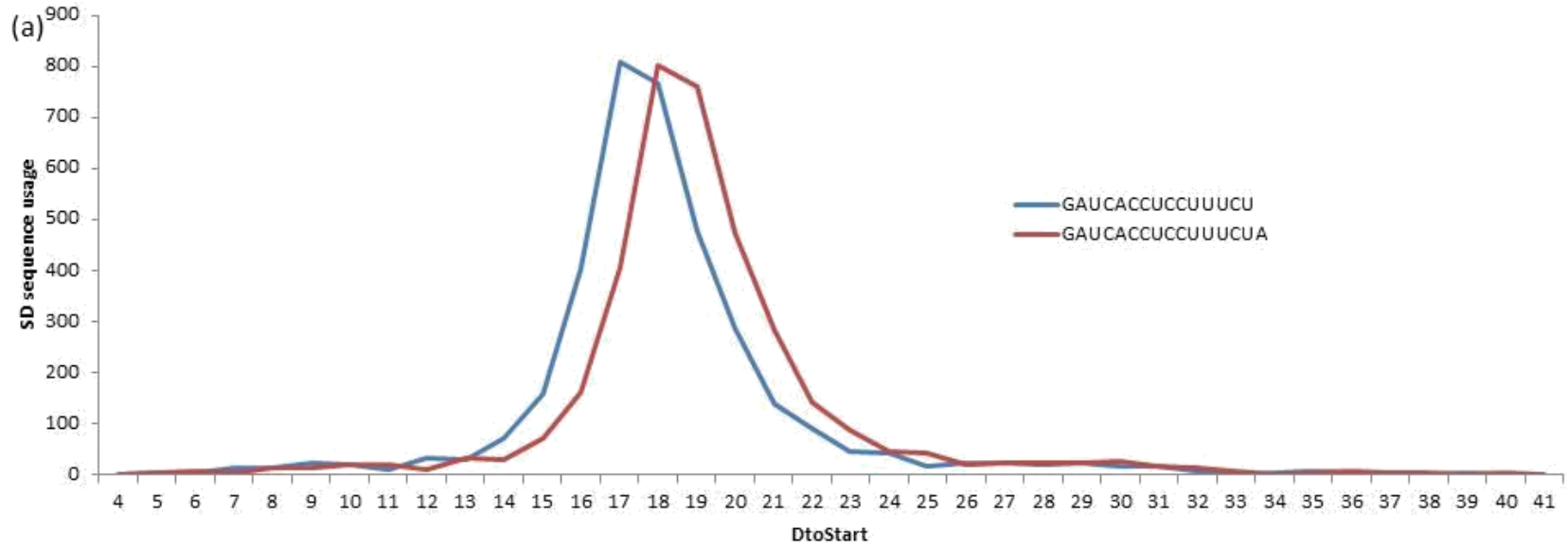
2.3.2 The effect of SD/aSD pairing location on initiation efficiency

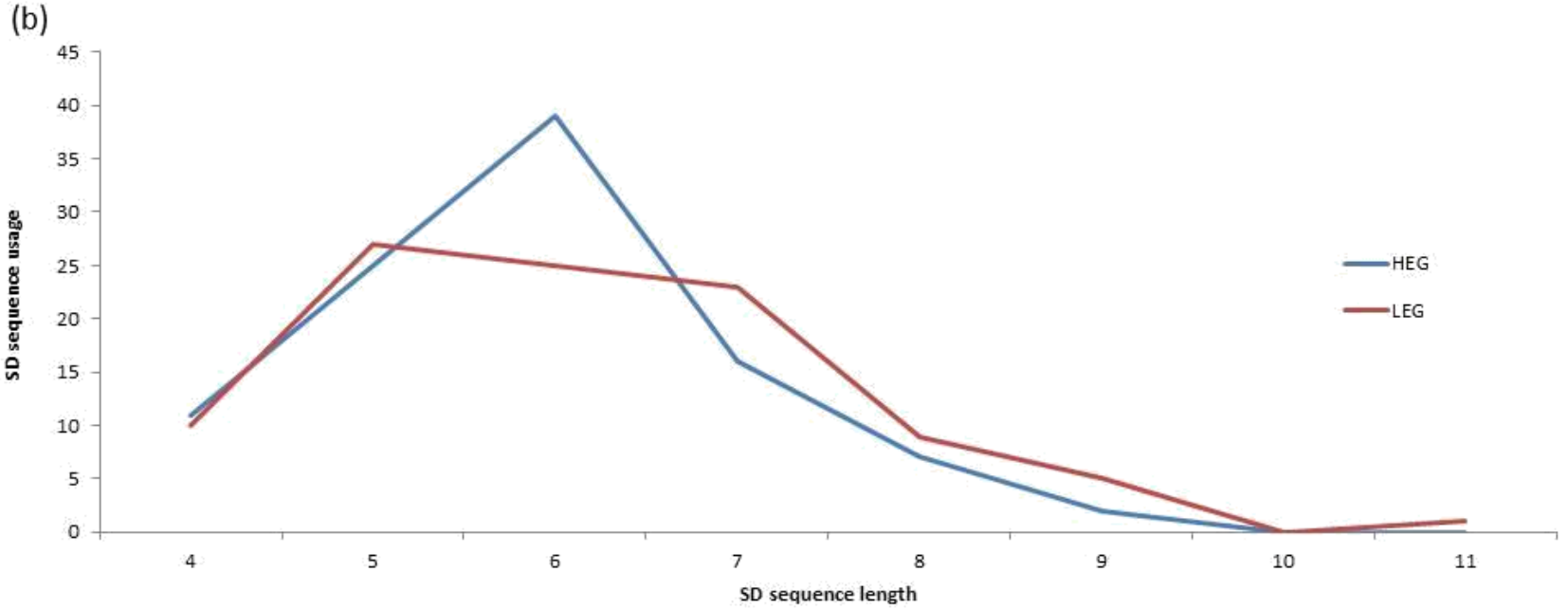
The mature 3' TAIL in *B. subtilis* identified here (5'-GAUCACCUCCUUUCU-3') requires that the optimal range for $D_{toStart}$ positions, described in a previous study as 15-25 using 5'-GAUCACCUCCUUUCUA-3' (Abolbaghaei *et al.* 2017), to be redefined. In order to accomplish this, we determined all putative 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. 2.4a. As for *E.*

coli, the previously reported $D_{toStart}$ range of 10 to 21 nt (Abolbaghaei *et al.* 2017) 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. 2.4a; $D_{toStart}$ = 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. 2.4b). 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. 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 (Lorenz *et al.* 2011)). 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.

Figure 2.4. (a) The distance, in nucleotides, between the 3’ terminus of the 16S rRNA and the start codon ($D_{toStart}$) is constrained to a narrow range for all *B. subtilis* putative SD sequences, but the distances vary depending on the terminal position of the 3’ TAIL. Altering the length of the 3’ TAIL causes a directly proportional shift in $D_{toStart}$. (b) Difference in motif length preference of SD sequences with a fixed $D_{toStart}$ = 17 in highly and lowly expressed genes.


(a)
SD sequence usage
900
800
700
600
500
400
300
200
100
0
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
DtoStart
GAUCACCUCCUUUCU
GAUCACCUCCUUUCUA


(b)
SD sequence usage
45
40
35
30
25
20
15
10
5
0
4 5 6 7 8 9 10 11
SD sequence length
HEG
LEG

2.3.3 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. 2.5; 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. 2.5. Since CCUCC constitutes the middle segment of the 3' TAIL: 5'-GAUCACCUCCUUA-3' (Shine and Dalgarno 1974) and 5'-GAUCACCUCCUUUCU-3' (Murray and Rabinowitz 1982) in *E. coli* and *B. subtilis*, respectively, it is unsurprising that the expected usage of this motif is the highest, as illustrated in Fig. 2.6. 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. 2.6ac), or has an observed/expected usage ratio > 1 (Fig. 2.6bd). In *B. subtilis*, aSD sites 5'-CUCCUUU-3' are favorably selected (Fig. 2.6ab), and in

Figure 2.5. 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.

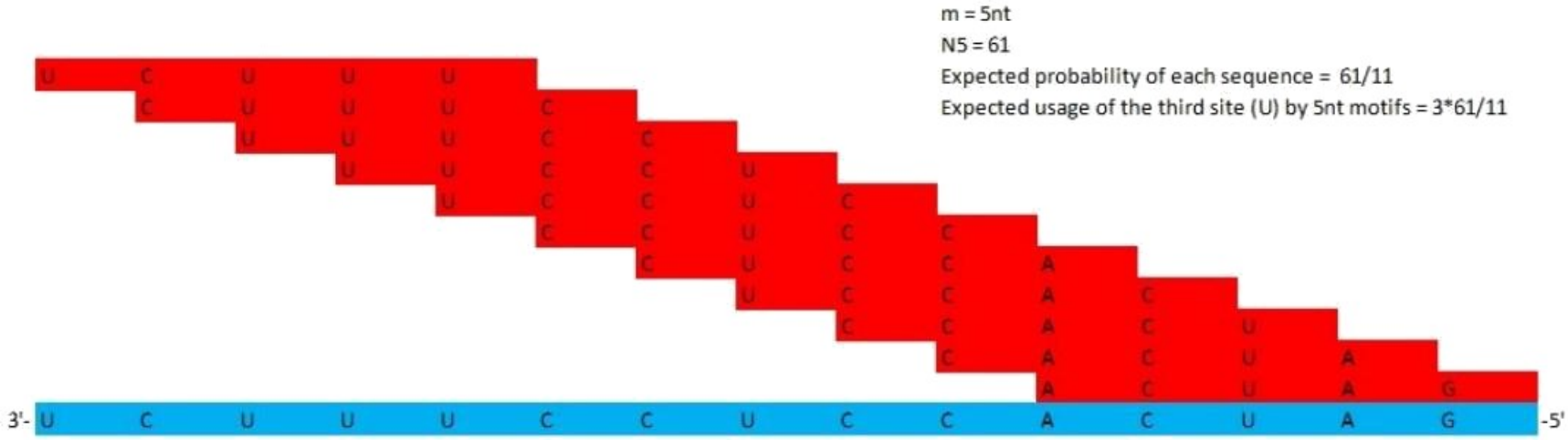
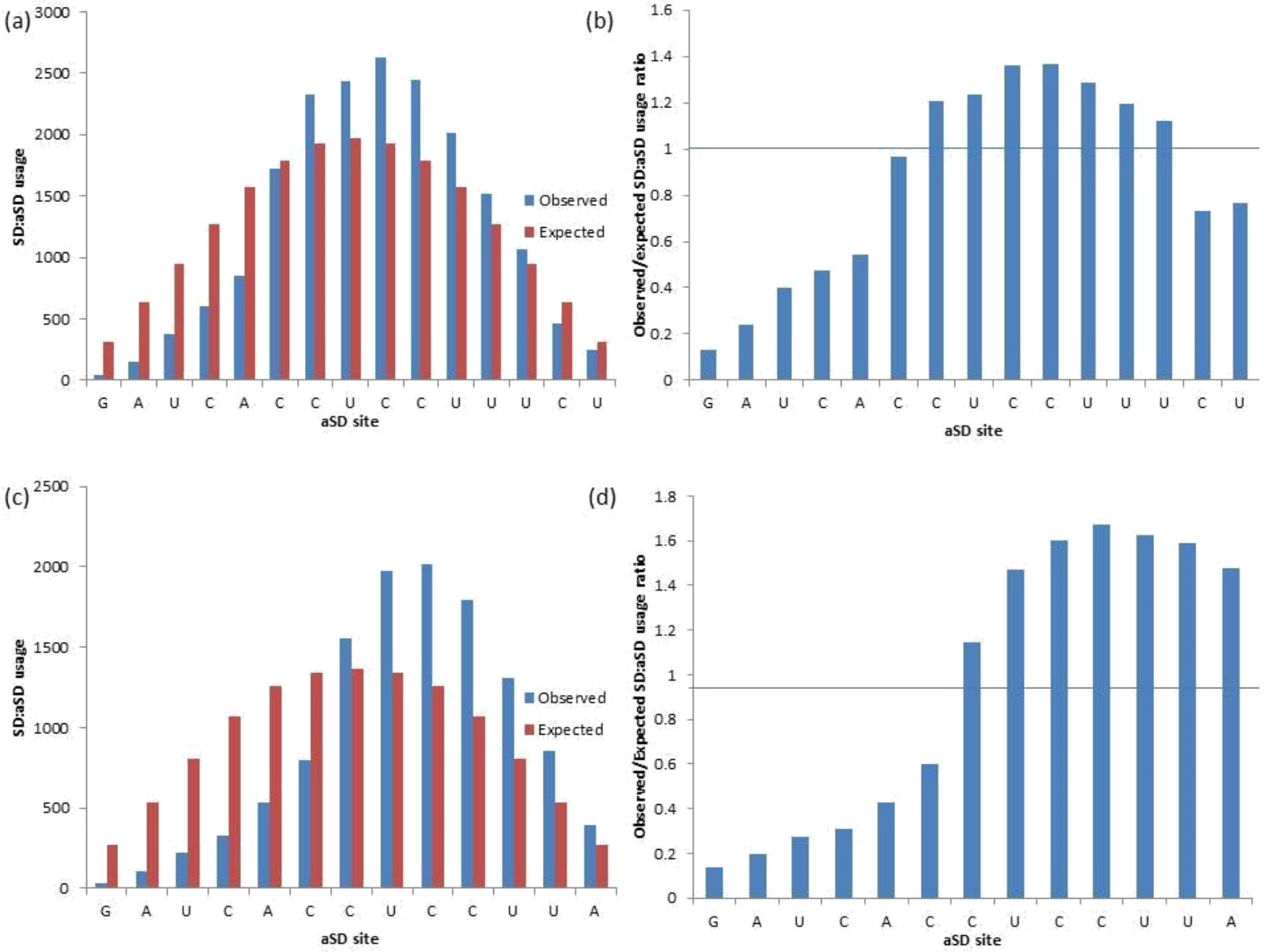

m = 5nt
N5 = 61
Expected probability of each sequence = 61/11
Expected usage of the third site (U) by 5nt motifs = 3*61/11
3'- U C U U U C C U C C A C U A G -5'

Figure 2.6. 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 between 4 and 12 nt are determined at optimal $D_{toStart}$ locations (15 to 21 in *B. subtilis*, 10 to 21 in *E. coli*).


(a)
SD:aSD usage
3000
2500
2000
1500
1000
500
0
Observed
Expected
G A U C A C C U C C U U U C U
aSD site
(b)
Observed/expected SD:aSD usage ratio
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
G A U C A C C U C C U U U C U
aSD site
(c)
SD:aSD usage
2500
2000
1500
1000
500
0
Observed
Expected
G A U C A C C U C C U U A
aSD site
(d)
Observed/Expected SD:aSD usage ratio
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
G A U C A C C U C C U U A
aSD site

E. coli, aSD sites 5'-CUCCUUA-3' were found to be favorably selected (Fig. 2.6cd). 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 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* (Fig. 2.7a), and 5'-CCUCCUUA-3' in *E. coli* (Fig. 2.7b). 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. 2.8a: approximately -9kcal/mol to -7kcal/mol, $P = 0.001915$, $R^2 = 0.7282$) and in *E. coli* (Fig. 2.8b: approximately -6kcal/mol to -4kcal/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

Figure 2.7. 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. Binding strength values were determined through adjusted ΔG heterodimer values (Mathews *et al.* 2004) implemented in RNAcofold (Lorenz *et al.* 2011).

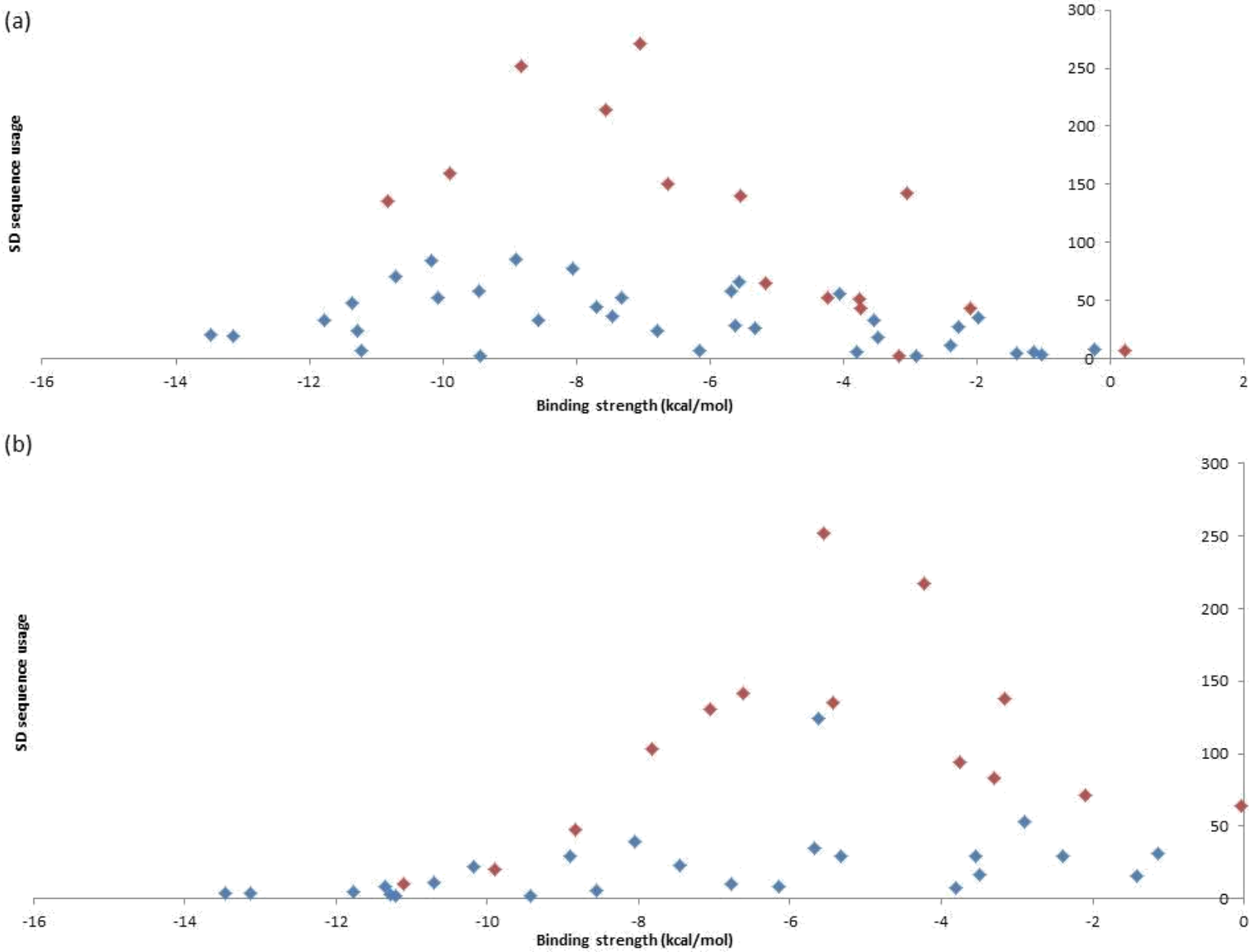
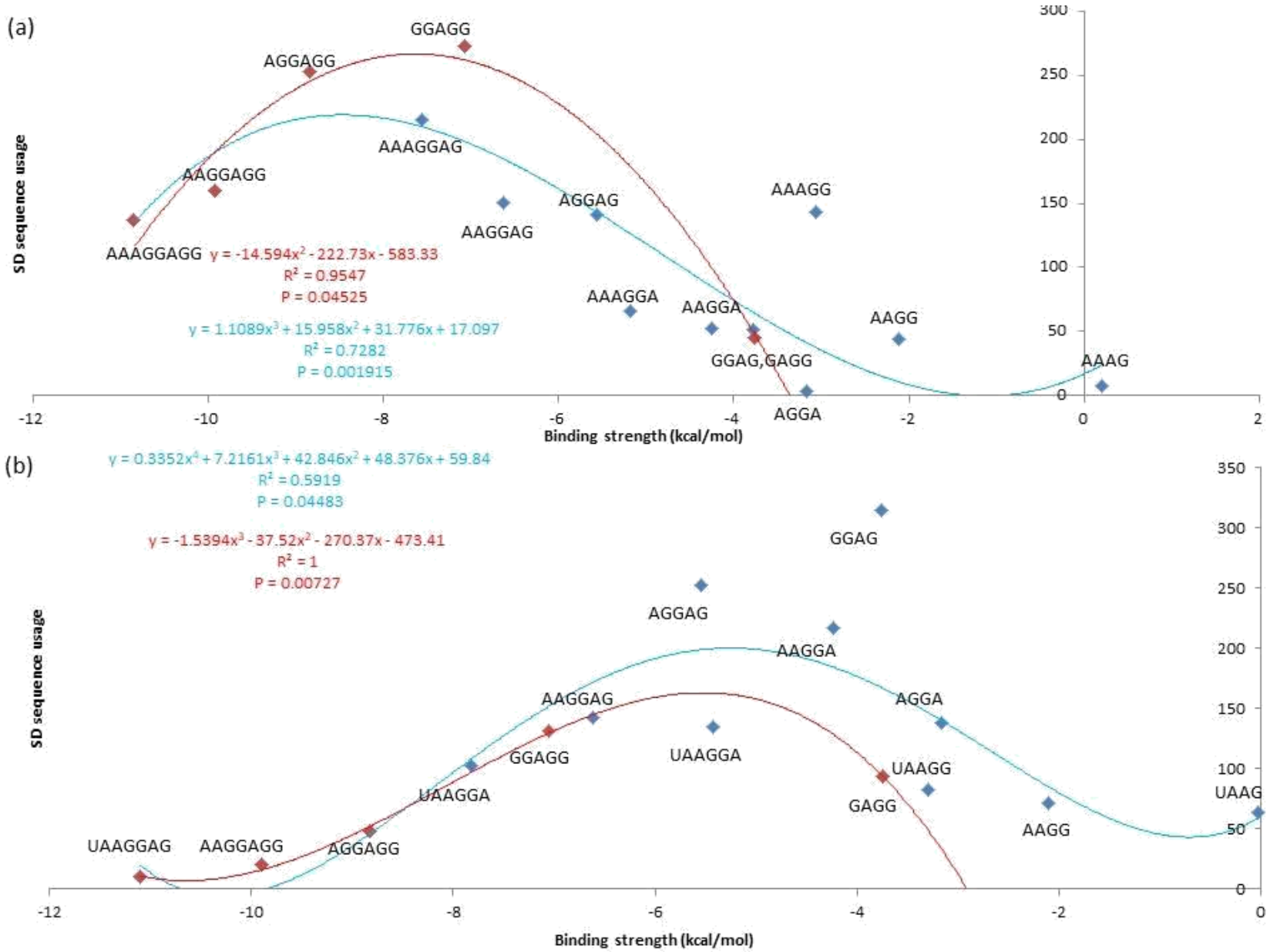

(a)
SD sequence usage
300
250
200
150
100
50
0
-16
-14
-12
-10
-8
-6
-4
-2
0
2
Binding strength (kcal/mol)
(b)
SD sequence usage
300
250
200
150
100
50
0
-16
-14
-12
-10
-8
-6
-4
-2
0
Binding strength (kcal/mol)

Figure 2.8. 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′ (*E. coli*). Highlighted in red are SD sequences that were identified after the core aSD sequences were extended to encompass CCUCC. Binding strength values were determined through adjusted ΔG heterodimer values (Mathews *et al.* 2004) implemented in RNAcofold (Lorenz *et al.* 2011).

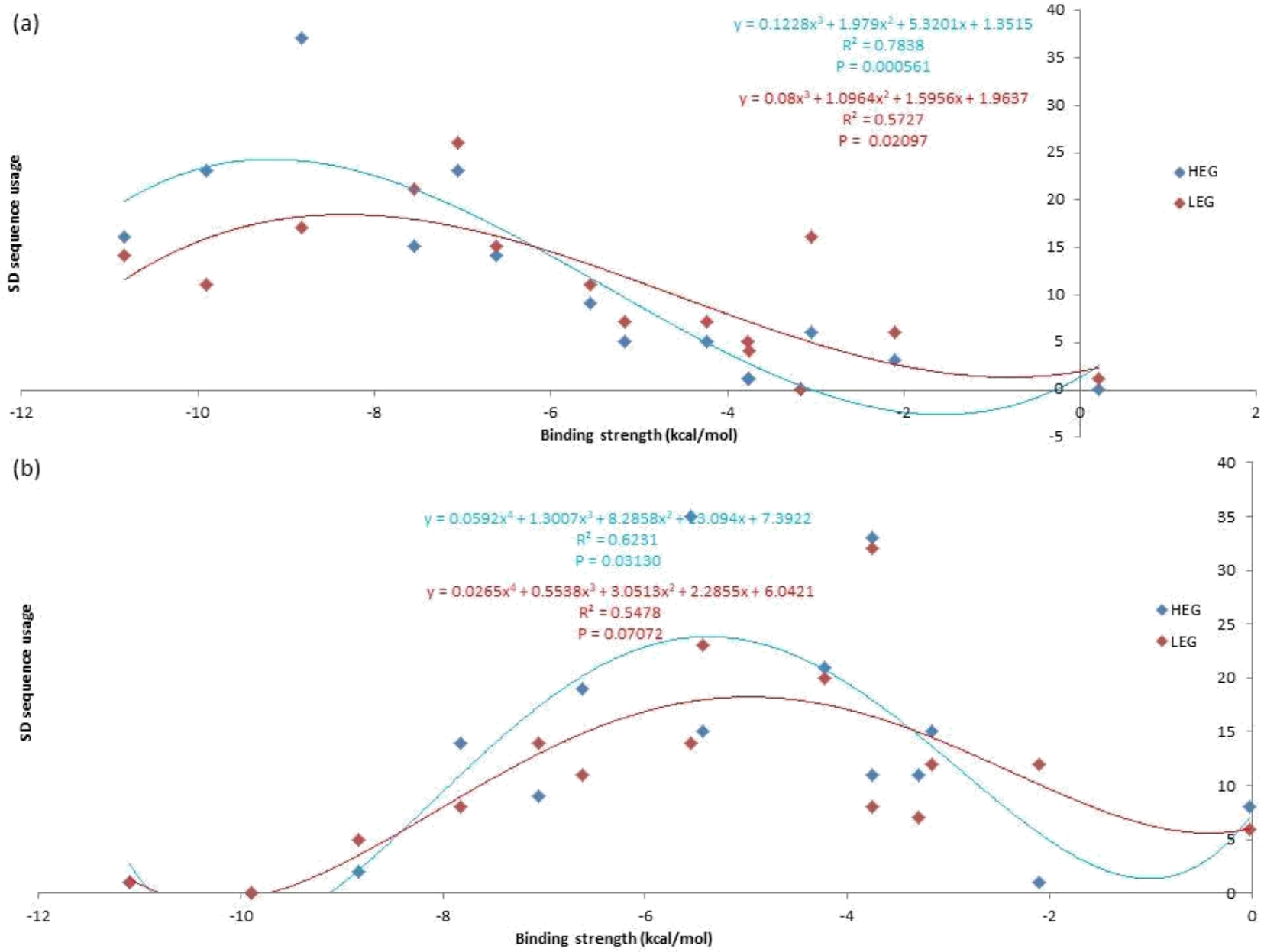

(a)
SD sequence usage
Binding strength (kcal/mol)
GGAGG
AGGAGG
AAAGGAG
AAGGAGG
AGGAG
AAGGAG
AAAGG
AAAGGAGG
y = -14.594x² - 222.73x - 583.33
R² = 0.9547
P = 0.04525
y = 1.1089x³ + 15.958x² + 31.776x + 17.097
R² = 0.7282
P = 0.001915
AAAGGA
AAGGA
AAGG
GGAG,GAGG
AGGA
AAAG
(b)
y = 0.3352x⁴ + 7.2161x³ + 42.846x² + 48.376x + 59.84
R² = 0.5919
P = 0.04483
y = -1.5394x³ - 37.52x² - 270.37x - 473.41
R² = 1
P = 0.00727
GGAG
AGGAG
AAGGA
AAGGAG
AGGA
GGAGG
UAAGGA
UAAGGA
UAAGG
GAGG
AAGG
UAAG
UAAGGAG
AAGGAGG
AGGAGG

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 (-7.05kcal/mol) are less selected for in the former than the latter due to the high binding affinity of the motif (Fig. 2.8b). 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. 2.6b).

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. 2.9). 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*. (2017) 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. 2.9), 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

Figure 2.9. The association between SD sequence usage and binding strength 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*). Binding strength values were determined through adjusted ΔG heterodimer values (Mathews *et al.* 2004) implemented in RNAcofold (Lorenz *et al.* 2011).


(a)
y = 0.1228x³ + 1.979x² + 5.3201x + 1.3515
R² = 0.7838
P = 0.000561
y = 0.08x³ + 1.0964x² + 1.5956x + 1.9637
R² = 0.5727
P = 0.02097
HEG
LEG
SD sequence usage
Binding strength (kcal/mol)
(b)
y = 0.0592x⁴ + 1.3007x³ + 8.2858x² + 23.094x + 7.3922
R² = 0.6231
P = 0.03130
y = 0.0265x⁴ + 0.5538x³ + 3.0513x² + 2.2855x + 6.0421
R² = 0.5478
P = 0.07072
HEG
LEG
SD sequence usage
Binding strength (kcal/mol)

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 (Li *et al.* 2014). 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 induced by the fact that intermediate levels of binding affinity are preferable.

2.4 Materials and Methods

2.4.1 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.

2.4.2 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 $1*10^{-17}$ 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. 2.3.

2.4.3 Classifying genes according to gene expression

We used protein abundance (ppm) data as proxies of gene expression. The integrated datasets were downloaded from PaxDB (Wang *et al.* 2015) 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.

2.4.4 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 (Xia 2017a), following the methods used in two previous studies (Prabhakaran *et al.* 2015; Abolbaghaei *et al.* 2017): 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 (Lorenz *et al.* 2011). This approach to calculating ΔG differs from standard minimum free energy calculations by attempting to account for chemical modifications that occur on bases *in vivo*. As such, scoring penalties are applied to terminal G-C and A-U parings, and a separate set of scoring penalties are applied to G-U pairs and G-U pairs with adjacent G-U pairs as previously described (Mathews *et al.* 2004).

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* (Abolbaghaei *et al.* 2017) 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.

2.4.5 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:

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

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

$$4 \times \frac{4}{12} + 5 \times \frac{5}{11} + 6 \times \frac{6}{10} + 6 \times \frac{7}{9} + 6 \times \frac{8}{8} + 6 \times \frac{9}{7} + 6 \times \frac{10}{6} + 6 \times \frac{11}{5} + 5 \times \frac{12}{4} \quad (2)$$

The same methodology is applied to measure usage of *E. coli* SD sequences. These computations are implemented in DAMBE (Xia 2013, 2017a) 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/).

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).

Additional Information

Supplementary Figure S2.1 is available in Supplementary Information; all data used to make the figures in the manuscript are available in Supplementary Dataset 1.

Chapter Three

RNA-Seq-based analysis reveals heterogeneity in mature 16S rRNA 3' termini and extended anti-Shine-Dalgarno motifs in bacterial species

3.0 Comments

This chapter is published in the December edition of G3: Genes|Genomes|Genetics under the title "RNA-Seq-based analysis reveals heterogeneity in mature 16S rRNA 3' termini and extended anti-Shine-Dalgarno motifs in bacterial species." JS and YW contributed equally to the analysis and interpretation of results as well as figure preparation and writing the manuscript. This chapter is formatted to reflect the submission guidelines of G3. Instances of SD:aSD (in reference to the sequence pairings) from the published article were changed to SD/aSD here to maintain consistency.

3.1 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.

3.2 Introduction

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 constitues ideal SD/aSD complementarity remains a subject of debate. Some researchers contend that there is no 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 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.

3.3 Materials and Methods

3.3.1 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.

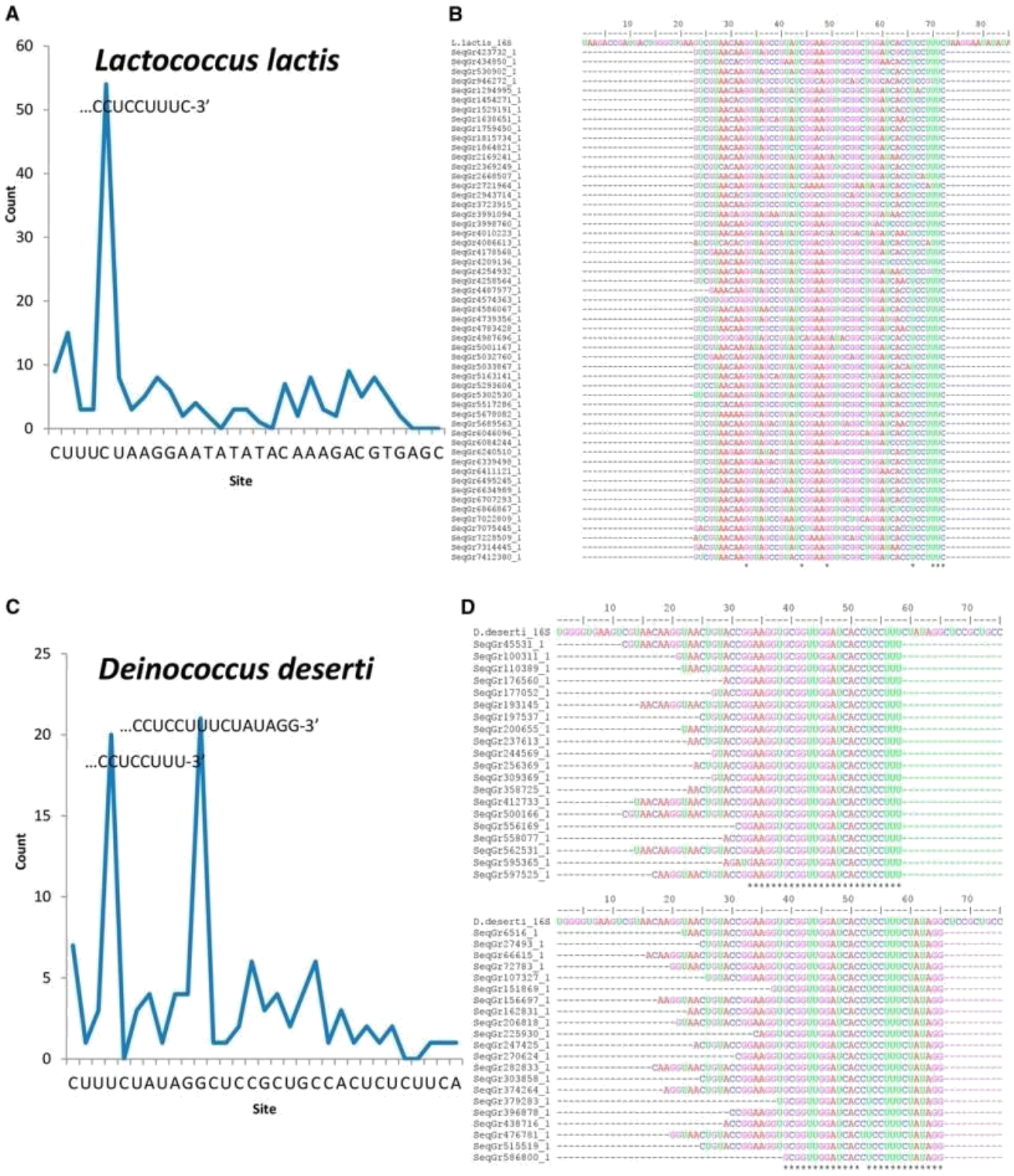
High-throughput RNA-Seq SRX runs of wildtype species were downloaded from GEO DataSets in FASTQ format. The FASTQ files were first converted to FASTQ+ format using ARSDA 1.1 (Xia 2017b), grouping identical reads under a single ID while also indicating the number of copies (SeqID_# of copies), in order to reduce the size of the datasets prior to adapter trimming. The FASTQ+ data was 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, Fig. 3.1d). The processed FASTQ+ datasets were subsequently converted into FASTA format for multiple sequence alignment.

3.3.2 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

Figure 3.1. 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.


A
Lactococcus lactis
...CCUCCUUUC-3'
Count
Site
CUUUC UAAGGAATATATAC AAAGAC GTGAGC
B
C
Deinococcus deserti
...CCUCCUUUCUAUAGG-3'
...CCUCCUUU-3'
Count
Site
CUUUCUAUAGGCUCCGCUGCCACUCUCUUCA
D

and quality of hits, as well as search speed, against the >= 25 nt reads in the ribo-depleted datasets. Then, the hits were retrieved from the FASTA files using seqtk (Li *et al.* 2012) and complementary strand sequences were eliminated. Finally, the hits were aligned to the query sequence using multiple sequence alignment (Clustal Omega algorithm (Sievers and Higgins 2014) implemented in DAMBE, default parameters).

3.3.3 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.* 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 S3.1 contains the 3' TAIL map for the remaining 11 species. They are available at FigShare: https://figshare.com/s/766dea5a60a413f3f147.

3.4 Results and Discussion

3.4.1 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).

3.4.2 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 3.1 shows the sequence map and alignments for *Lactococcus lactis* and *Deinococcus deserti*. The sequences mapped for the 11 others are present in Supplementary Figure S3.1. 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

Table 3.1. 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*	GAUCACCUCCUUUC**UA**†		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.	GAUCACCUCCUUUAAGGG		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.

due to their failure to incorporate the canonical CCUCC. Resultantly, the annotated 3' TAILs of only two out of 13 species, NC_003210 *Listeria 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 Gamma-proteobacteria. 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 (Fig. 3.1, Table 1, Supplementary Fig. S3.1). 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) (Fig. 3.1c, Supplementary Fig. S3.1).

3.4.3 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 U 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:

$$\frac{12}{\sum_{x=4} \quad _{15-x+1}} \qquad (1)$$

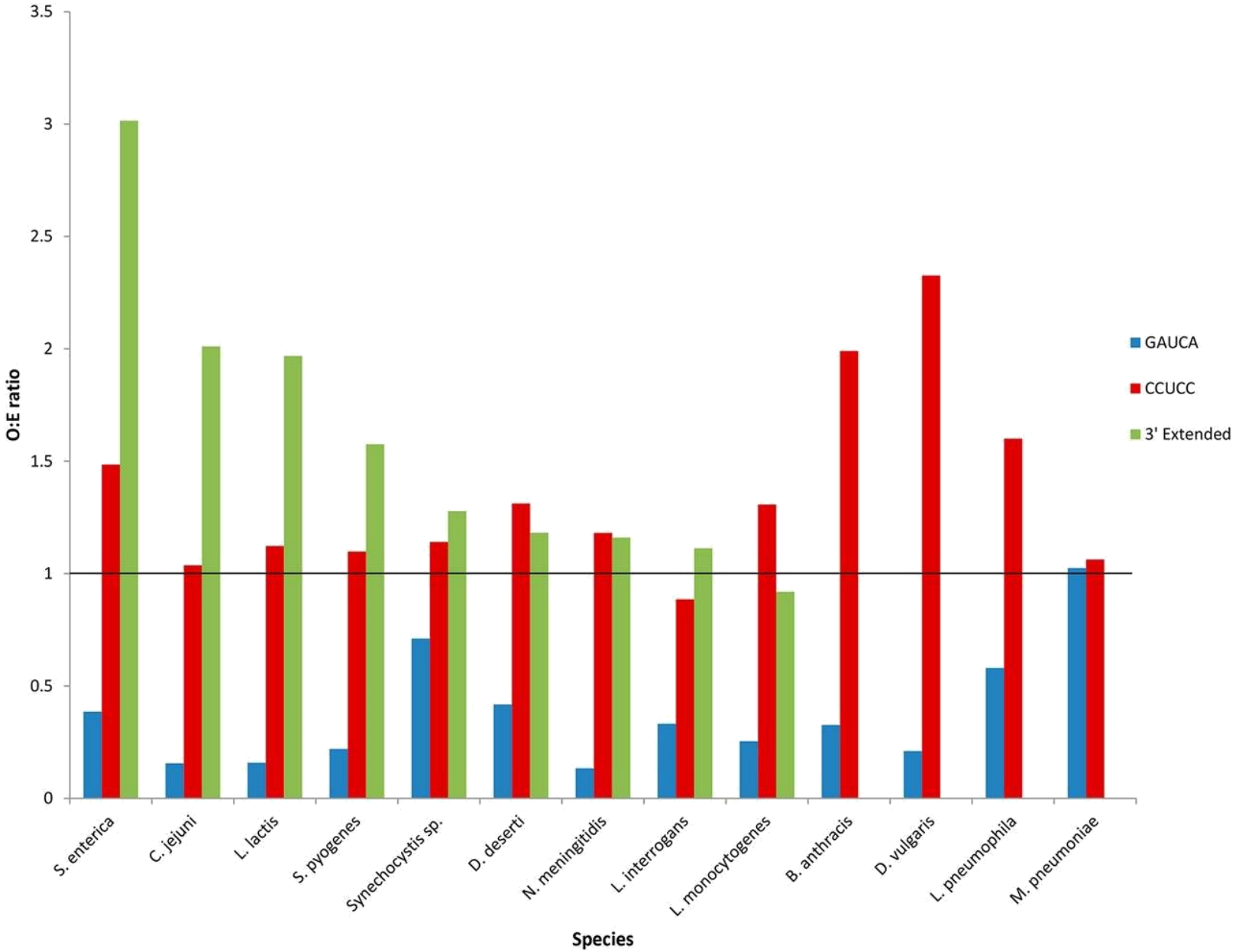
However, the number of possible base-pairs resulting in perfect complementarity varies. For example, a 12 nt putative SD sequence may start pairing at the first, but not the sixth, base on a complementary aSD sequence that is 15 nt long, and the maximum usage of the sixth aSD site is calculated instead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

$$4\times\frac{4}{12} \quad +5\times\frac{5}{11} \quad +6\times\frac{6}{10} \quad +6\times\frac{7}{9} \quad +6\times\frac{8}{8} \quad +6\times\frac{9}{7} \quad +6\times\frac{10}{6} \quad +6\times\frac{11}{5} \quad +5\times\frac{12}{4} \qquad (2)$$

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 3.2 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

Figure 3.2. Mean ratio of observed over expected SD/aSD complementarity (O:E ratio) in 13 species at conserved 5′-GAUCA-3′ (blue), and CCUCC 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.


3.5
3
2.5
2
1.5
1
0.5
0
O:E ratio
GAUCA
CCUCC
3' Extended
S. enterica
C. jejuni
L. lactis
S. pyogenes
Synechocystis sp.
D. deserti
N. meningitidis
L. interrogans
L. monocytogenes
B. anthracis
D. vulgaris
L. pneumophila
M. pneumoniae
Species

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.

Acknowledgements

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.

Supplementary information is made available in the Supplementary files S1, S2, and Fig. S3.1 at FigShare: https://figshare.com/s/766dea5a60a413f3f147.

Chapter Four

Discussion

4.0 Advancements in the 3' TAIL mapping approach

Finding discrepant reports on the 3' TAIL terminus of *B. subtilis* motivated us to characterize the 3' TAIL using an RNA-Seq approach in chapter two. We have also included *E. coli* as a proof of concept because the reported 3' TAIL remains undisputed(Shine and Dalgarno 1974; Woese *et al.* 1975). By recovering the originally reported 3' TAILs from both species, we established that our method worked as intended. The results from chapter two prompted us to further improve our RNA-Seq based approach in chapter three so that it could be widely applicable to bacterial species.

Chapter three significantly improves upon the RNA-Seq-based strategy introduced in chapter two by enabling it to characterize the 3' TAILs in species where small datasets must be used due limited data availability. For instance, our quality control was initially based on a site-specific assessment of the average sequence quality across all sequences. This is problematic in small datasets because site-specific assessments do not provide enough information about the reliability of specific sequences. In cases where there are only a few reads that map to a given molecule, we would not reliably determine the precise 3' TAIL due to noise. Fortuitously, the site-specific average quality scores were high for both *E. coli* and *B. subtilis* datasets considered in chapter two. Without controlling for sequence-specific quality we would have to manually select high quality mapped sequences for all matches, which we did for the two species studied in chapter two.

Another advancement to our initial approach is to pre-process the data for adapter sequences. This step adds a number of adapter-containing reads to the alignment that would otherwise be eliminated during mapping. This increase in read depth is especially important for small datasets. Applying these two processing steps to the first iteration of our approach increases the abundance of reads mapped to the true terminus of the 3' TAIL. These improvements were necessary because the higher volume of data considered were generated in different labs; thus these data are heterogeneous in terms of quality and size. For example, using sequence-specific quality control, we increase our confidence that mappings are not spurious in species such as *C. jejuni*, which has very few overall mappings. Additionally, trimming adapter sequences afforded us more mappings than we would have otherwise obtained.

Arguably the greatest challenge across both chapters two and three was in obtaining usable data. As previously stated, the ubiquity of ribo-depletion treatments used in publicly archived RNA-Seq experiments rendered about 90% of the data we processed unusable for the purposes of characterizing 16S rRNAs. This resulted in very little choice as to the species we were able to represent in our analyses; however, despite this limitation, we were able to observe differences in the 3' termini of even closely related species such as *E. coli* and *S. enterica*. This implies that the extent of the mature 16S rRNA is likely far more diverse among bacterial species than rDNA sequence similarity would lead us to believe.

4.1 Heterogeneity in 3’ TAIL termini influences SD/aSD paring

Our results show that the terminus of the mature 3’ TAIL is unexpectedly heterogeneous among species, even for closely related species such as *B. anthracis* (5’- GAUCACCUCC-3’) and *B. subtilis* (5’-GAUCACCUCCUUUCU-3’), or *E. coli* (5’-GAUCACCUCCUUA-3’) and *S. enterica* (5’-GAUCACCUCCU-3’). Importantly, we find that despite this heterogeneity, 3’ TAIL termini do not extend past genomically conserved sequences across bacteria, except in *L. interrogans*. Moreover, species-specific 3’ TAILs can be explained when the functional constraint imposed by SD/aSD pairing is taken into consideration.

In addition to heterogeneity between species, chapter three also revealed substantial heterogeneity within a subset of the studied species. We previously postulated that, in some cases, the observed 3’ TAIL heterogeneity within a given species could represent pre-16S RNA processing intermediates. However, an alternative hypothesis that was not previously discussed is that the 16S molecules with extended 3’ termini may represent alternate mature 3’ TAILs that are preferred to facilitate the initiation of genes that are not constitutively expressed, or are inducible under specific conditions. These hypotheses are compatible with the findings of a recent comparative analysis between the 3’ TAILs of about 35,000 prokaryotes (Amin *et al.* 2018). Although the DNA-level nature of their analysis was unable to provide evidence for or against the extension of 3’ TAILs beyond a consensus motif, they observed that the canonical CCUCC aSD motif was absent in a small subset of 3’ TAILs that otherwise retained an identifiable helix 45.

Most notably, Amin and colleagues (2018) concluded that species which lacked a CCUCC aSD motif exhibited no signs of compatible SD/aSD interactions; however, this conclusion is problematic for two reasons. First, they did not account for 3' TAILs beyond 13 nt in length, even though our results in chapters two and three demonstrate that there are species exhibiting extended 3' TAILs. Additionally, species they identified as having no CCUCC also had the most substitutions in helix 45. Taken together, it is not excluded that the substitutions in helix 45 could potentially alter the structure in this region in such a way as to shift the 3' TAIL further downstream than the sites that were considered by Amin and colleagues (2018). Although this possibility was not considered, the hypothesis they put forward is a valid one, namely that species which lack identifiable SD/aSD interactions rely more on non-SD-mediated initiation mechanisms.

Furthermore, our results reveal that many species use bases downstream of CCUCC in SD/aSD pairs more frequently than expected. There are two key factors that play a role in increased usage of bases downstream of the energetically more stable pairing between GGAGG (SD) and CCUCC (aSD). One is that increased selection for intermediate SD/aSD binding affinities may reduce the transition time between initiation and elongation whereas strong SD/aSD pairing lead to translational pausing (Li *et al.* 2012). The second invokes a reduced dependence on SD-mediated initiation, as was observed for *L. interrogans*, in favor of other mechanisms such as S1-mediated or leaderless initiation.

4.2 Future directions

We have recently made further improvements to the previous mapping strategy shown in chapters two and three. By using kallisto (Bray *et al.* 2016) to pseudoalign RNA-Seq reads to a given target or set of target sequences in addition to ARSDA (Xia 2017b), we achieve faster results that allow for more precise and accurate transcript quantification than our BLAST mapping approach. This has allowed us to quantify bacterial tRNA expression in a manner that yields results consistent with previously reported tRNA abundances (Wei *et al.* 2019) from RNA fingerprinting experiments (Ikemura 1981, 1985; Dong *et al.* 1996; Kanaya *et al.* 1999). This demonstrates that our method can be successfully applied to other transcripts, expanding its use to tasks such as operon mapping.

One specific application of our improved method is to elucidate the nature of the 3' TAIL heterogeneity reported in chapter three, especially in cases such as *D. deserti*, wherein two major termini are observed. In these instances, both observed 3' TAILs can be subjected to SD/aSD pairings analogously to what we have previously done in chapters two (Fig. 2.6) and three (Fig. 3.2) by using different genes sets organized by expression (HEGs, LEGs, all genes). There are two main hypotheses to explain the observed heterogeneity in the extended 3' TAILs:

1) they represent pre-16S molecules that are processing intermediates, and 2) they are alternative mature products.

Our major prediction in support of the first hypothesis is that the extended 3' TAIL region should not participate in SD/aSD binding more frequently than expected at any sites for any particular gene set. Further evidence to support this hypothesis could be obtained through

structural analysis of the 3' TAIL region using metrics such as minimum free energy to assess energetically favored conformations. In cases where secondary structure is high within the 3' TAIL region, nucleotides are more prone to intramolecular than intermolecular binding, making it unlikely for them to participate in SD/aSD binding. Similarly, demonstrating higher activity of alternative initiation mechanisms, such as S1-mediated initiation, would suggest that constraint on SD-mediated initiation is more relaxed. In turn, this could lead to mutations that alter the expressed structure of the 3' TAIL at the RNA-level.

The second hypothesis has, arguably, more interesting implications and requires more rigorous testing than the first, in part because evidence disfavoring it provides indirect support for the first hypothesis. Our major prediction in support of an alternative 3' TAIL is that it would be functionally relevant in a subset of genes. Such genes are more likely to be inducible rather than constitutively expressed, in part because producing longer products is more resource-intensive than shorter alternatives, even if difference is marginal. Given that the experimental data we examined was taken from wild-type stains at culture conditions as close to standard for the particular organism as possible, most inducible genes are likely to be best represented by the lowest expression category. Promising gene targets for further experimentation can be identified by determining the candidates that best participate in SD/aSD interactions with downstream positions exclusive to the extended 3' TAILs.

Many such inducible genes are likely to share categories of exploitable commonalities, such as being present in the same operon and/or otherwise having similar regulatory elements that facilitate tandem increases in their expression. Wet lab experiments can be performed to

upregulate such genes and determine whether or not an accompanying increase in expression is observed for the associated extended 3' TAIL. This investigation could optionally broaden the focus to assess translation elongation in these genes by also examining actively translated mRNAs using ribosome profiling (Ribo-Seq) (Ingolia *et al.* 2009) and steady-state protein levels using a technique such as APEX (Arike *et al.* 2012). These changes could be compared to potential expression-level differences in the tRNA pool brought about by the induction stimuli (van Weringh *et al.* 2011; Wei *et al.* 2019).

In the long term, the approaches outlined above can open the door to a concerted approach whereby pathogenic bacterial species can be studied in more detail in the hopes of developing effective solutions to their increasing antibiotic resistance, especially for species that heavily rely on SD-mediated initiation and tRNA-mediated codon usage. A basic outline of the approach entails characterizing the 3' TAIL of the pathogen in question as well as quantifying its transcriptome and tRNA pool under desired experimental conditions (such as during the course of phage infection) to understand shifts in tRNA expression caused by the stimuli of interest. By combining these data with codon usage information, Ribo-Seq data, and steady-state protein abundances, it is possible to revisit phage therapies that focus on genetically engineering phages to replicate rapidly or effectively compete for translation machinery in the host pathogen. This can be achieved by engineering the phages to best take advantage the host translation machinery by using emerging genome editing technologies with high precision, such as CRISPR/Cas (Cong *et al.* 2013), to add in optimized regulatory elements. This serves the dual purpose of mitigating the amount of harmful toxins produced by the pathogen due to increased

competition for gene expression while also increasing the efficiency and rate of the cell lysis by bacteriophage.

References

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 7: 1607–1615.

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.

Amin M. R., A. Yurovsky, Y. Chen, S. Skiena, and B. Futcher, 2018 Re-annotation of 12,495 prokaryotic 16S rRNA 3' ends and analysis of Shine-Dalgarno and anti-Shine-Dalgarno sequences. PLoS One 13: e0202767.

Anda M., Y. Ohtsubo, T. Okubo, M. Sugawara, Y. Nagata, *et al.*, 2015 Bacterial clade with the ribosomal RNA operon on a small plasmid rather than the chromosome. Proc. Natl. Acad. Sci. 112: 14343 LP-14347.

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.

Arike L., K. Valgepea, L. Peil, R. Nahku, K. Adamberg, *et al.*, 2012 Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli. J. Proteomics 75: 5437–5448.

Band L., and D. J. Henner, 1984 Bacillus subtilis Requires a "Stringent" Shine-Dalgarno Region for Gene Expression. DNA 3: 17–21.

Barbe V., S. Cruveiller, F. Kunst, P. Lenoble, G. Meurice, *et al.*, 2009 From a consortium sequence to a unified sequence: the Bacillus subtilis 168 reference genome a decade later. Microbiology 155: 1758–1775.

Baumgardt K., L. Gilet, S. Figaro, and C. Condon, 2018 The essential nature of YqfG, a YbeY homologue required for 3' maturation of Bacillus subtilis 16S ribosomal RNA is suppressed by deletion of RNase R. Nucleic Acids Res. 46: 8605–8615.

Bolger A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

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.

Bray N. L., H. Pimentel, P. Melsted, and L. Pachter, 2016 Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34: 525.

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

Cabot G., L. Zamorano, B. Moya, C. Juan, A. Navas, *et al.*, 2016 Evolution of Pseudomonas aeruginosa Antimicrobial Resistance and Fitness under Low and High Mutation Rates. Antimicrob. Agents Chemother. 60: 1767–1778.

Cangelosi G. A., and W. H. Brabant, 1997 Depletion of pre-16S rRNA in starved Escherichia coli cells. J. Bacteriol. 179: 4457–4463.

Cannone J. J., S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, *et al.*, 2002 The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 3: 2.

Chen H. Y., M. Bjerknes, R. Kumar, and E. Jay, 1994 Determination of the Optimal Aligned Spacing between the Shine-Dalgarno Sequence and the Translation Initiation Codon of Escherichia-Coli Messenger-Rnas. Nucleic Acids Res. 22: 4953–4957.

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. U. S. A. 100: 7033–8.

Cong L., F. A. Ran, D. Cox, S. L. Lin, R. Barretto, *et al.*, 2013 Multiplex Genome Engineering Using CRISPR/Cas Systems. Science (80-.). 339: 819–823.

Cozen A. E., E. Quartley, A. D. Holmes, E. Hrabeta-Robinson, E. M. Phizicky, *et al.*, 2015 ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. Nat. Methods 12: 879–884.

Czernilofsky A. P., C. G. Kurland, and G. Stoffler, 1975 30S ribosomal proteins associated with the 3'-terminus of 16S RNA. FEBS Lett. 58: 281–284.

Deutscher M. P., 2015 Twenty years of bacterial RNases and RNA processing: how we've matured. RNA 21: 597–600.

Dong H., L. Nilsson, and C. G. Kurland, 1996 Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J. Mol. Biol. 260: 649–663.

Duin J. Van, and R. Wijnands, 1981 The function of ribosomal protein S21 in protein synthesis. Eur. J. Biochem. 118: 615–619.

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.

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.

Ewing B., and P. Green, 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8.

Goodwin S., J. D. McPherson, and W. R. McCombie, 2016 Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17: 333.

Green C. J., G. C. Stewart, M. A. Hollis, B. S. Vold, and K. F. Bott, 1985 Nucleotide sequence of the Bacillus subtilis ribosomal RNA operon, rrnB. Gene 37: 261–266.

Hakovirta J. R., S. Prezioso, D. Hodge, S. P. Pillai, and L. M. Weigel, 2016 Identification and Analysis of Informative Single Nucleotide Polymorphisms in 16S rRNA Gene Sequences of the Bacillus cereus Group. J. Clin. Microbiol. 54: 2749–2756.

Held W. A., M. Nomura, and J. W. Hershey, 1974 Ribosomal protein S21 is required for full activity in the initiation of protein synthesis. Mol. Gen. Genet. 128: 11–22.

Hockenberry A. J., A. J. Stern, L. A. N. Amaral, and M. C. Jewett, 2017 Diversity of translation initiation mechanisms across bacterial species is driven by environmental conditions and growth demands. bioRxiv.

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 Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. 84: 4762–4766.

Ikemura T., 1981 Correlation between the Abundance of Escherichia-Coli Transfer-Rnas and the Occurrence of the Respective Codons in its Protein Genes - a Proposal for a Synonymous Codon Choice that is Optimal for the Escherichia-Coli Translational System. J. Mol. Biol. 151: 389–409.

Ikemura T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2: 13–34.

Ingolia N. T., S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman, 2009 Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science (80-.). 324: 218–223.

Iserentant D., and W. Fiers, 1980 Secondary structure of mRNA and efficiency of translation initiation. Gene 9: 1–12.

Jacob W. F., M. Santer, and A. E. Dahlberg, 1987 A single base change in the Shine-Dalgarno region of 16S rRNA of Escherichia coli affects translation of many proteins. Proc. Natl. Acad. Sci. U. S. A. 84: 4757–4761.

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.

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.

Kanaya S., Y. Yamada, Y. Kudo, and T. Ikemura, 1999 Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238: 143–155.

Klappenbach J. A., J. M. Dunbar, and T. M. Schmidt, 2000 rRNA Operon Copy Number Reflects Ecological Strategies of Bacteria. Appl. Environ. Microbiol. 66: 1328 LP-1333.

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-a Publ. Rna Soc. 8: 1137–1147.

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.

Kudla G., A. W. Murray, D. Tollervey, and J. B. Plotkin, 2009 Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science (80-.). 324: 255–258.

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.

Laursen B. S., H. P. Sorensen, K. K. Mortensen, and H. U. Sperling-Petersen, 2005 Initiation of protein synthesis in bacteria. Microbiol. Mol. Biol. Rev. 69: 101–+.

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.

Li S., X. Dong, and Z. Su, 2013 Directional RNA-seq reveals highly complex condition-dependent transcriptomes in E. coli K12 through accurate full-length transcripts assembling. BMC Genomics 14: 520.

Li G.-W., D. Burkhardt, C. Gross, and J. S. Weissman, 2014 Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell 157: 624–635.

Li G.-W., 2015 How do bacteria tune translation efficiency? Curr. Opin. Microbiol. 24: 66–71.

Liljenstrom 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.

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.

Lin Liu; Yinhu Li; Siliang Li; Ni Hu; Yimin He; Ray Pong; Danni Lin; Lihua Lu; Maggie Law, 2012 Comparison of Next-Generation Sequencing Systems. J. Biomed. Biotechnol. 2012: 11.

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.

Liu Q., Y. Guo, J. Li, J. Long, B. Zhang, *et al.*, 2012 Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genomics 13.

Lorenz R., S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, *et al.*, 2011 ViennaRNA Package 2.0. Algorithms Mol Biol 6: 26.

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

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

Mathews D. H., M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, *et al.*, 2004 Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc. Natl. Acad. Sci. U. S. A. 101: 7287–7292.

Mortazavi A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, 2008 Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5: 621–628.

Murray C. L., and J. C. Rabinowitz, 1982 Nucleotide sequences of transcription and translation initiation regions in Bacillus phage phi 29 early genes. J. Biol. Chem. 257: 1053–1062.

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

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.

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

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.

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.

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.

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 U S A 109: 14458–14463.

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.

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. U. S. A. 71: 1342–1346.

Shine J., and L. Dalgarno, 1975 Determinant of cistron specificity in bacterial ribosomes. Nature 254: 34–38.

Sievers F., and D. G. Higgins, 2014 Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol. Biol. 1079: 105–116.

Smit M. H. de, and J. van Duin, 1994 Translational initiation on structured messengers. Another role for the Shine-Dalgarno interaction. J. Mol. Biol. 235: 173–184.

Sohmen D., S. Chiba, N. Shimokawa-Chiba, C. A. Innis, O. Berninghausen, *et al.*, 2015 Structure of the Bacillus subtilis 70S ribosome reveals the basis for species-specific stalling. Nat. Commun. 6: 6941.

Starmer J., A. Stomp, M. Vouk, and D. Bitzer, 2006 Predicting Shine–Dalgarno Sequence Locations Exposes Genome Annotation Errors. PLoS Comput Biol 2.

Steitz J. A., and K. Jakes, 1975 How ribosomes select initiator regions in mRNA: base pair formation between the 3’ terminus of 16S rRNA and the mRNA during initiation of protein synthesis in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. 72: 4734–4738.

Stern S., T. Powers, L. M. Changchien, and H. F. Noller, 1988 Interaction of ribosomal proteins S5, S6, S11, S12, S18 and S21 with 16 S rRNA. J. Mol. Biol. 201: 683–695.

Stevenson B. S., and T. M. Schmidt, 2004 Life history implications of rRNA gene copy number in Escherichia coli. Appl. Environ. Microbiol. 70: 6670–6677.

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.

TANIGUCHI T., and C. WEISSMANN, 1978 Inhibition of Qβ RNA 70S ribosome initiation complex formation by an oligonucleotide complementary to the 3’ terminal region of E. coli 16S ribosomal RNA. Nature 275: 770.

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. U. S. A. 107: 3645–3650.

Uchida T., L. Bonen, H. W. Schaup, B. J. Lewis, L. Zablen, *et al.*, 1974 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.

Vimberg V., A. Tats, M. Remm, and T. Tenson, 2007 Translation initiation region sequence preferences in Escherichia coli. Bmc Mol. Biol. 8: 100.

Walsh G., 2005 Therapeutic insulins and their large-scale manufacture. Appl. Microbiol. Biotechnol. 67: 151–159.

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

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.

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.

Waterman M. S., and T. F. Smith, 1978 RNA secondary structure: a complete mathematical

analysis. Math. Biosci. 42: 257–266.

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.

Wei Y., J. R. Silke, and X. Xia, 2019 An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. Sci. Rep. 9: 3184.

Weringh A. van, M. Ragonnet-Cronin, E. Pranckeviciene, M. Pavon-Eternod, L. Kleiman, *et al.*, 2011 HIV-1 Modulates the tRNA Pool to Improve Translation Efficiency. Mol. Biol. Evol. 28: 1827–1834.

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.

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.

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.

Xia X., 2013 DAMBE5: A Comprehensive Software Package for Data Analysis in Molecular Biology and Evolution. Mol. Biol. Evol. 30: 1720–1728.

Xia X., 2015 A Major Controversy in Codon-Anticodon Adaptation Resolved by a New Codon Usage Index. Genetics 199: 573–579.

Xia X., 2017a DAMBE6: New tools for microbial genomics, phylogenetics and molecular evolution. J Hered.

Xia X., 2017b ARSDA: A New Approach for Storing, Transmitting and Analyzing Transcriptomic Data. G3 Genes|Genomes|Genetics 7: 3839–3848.

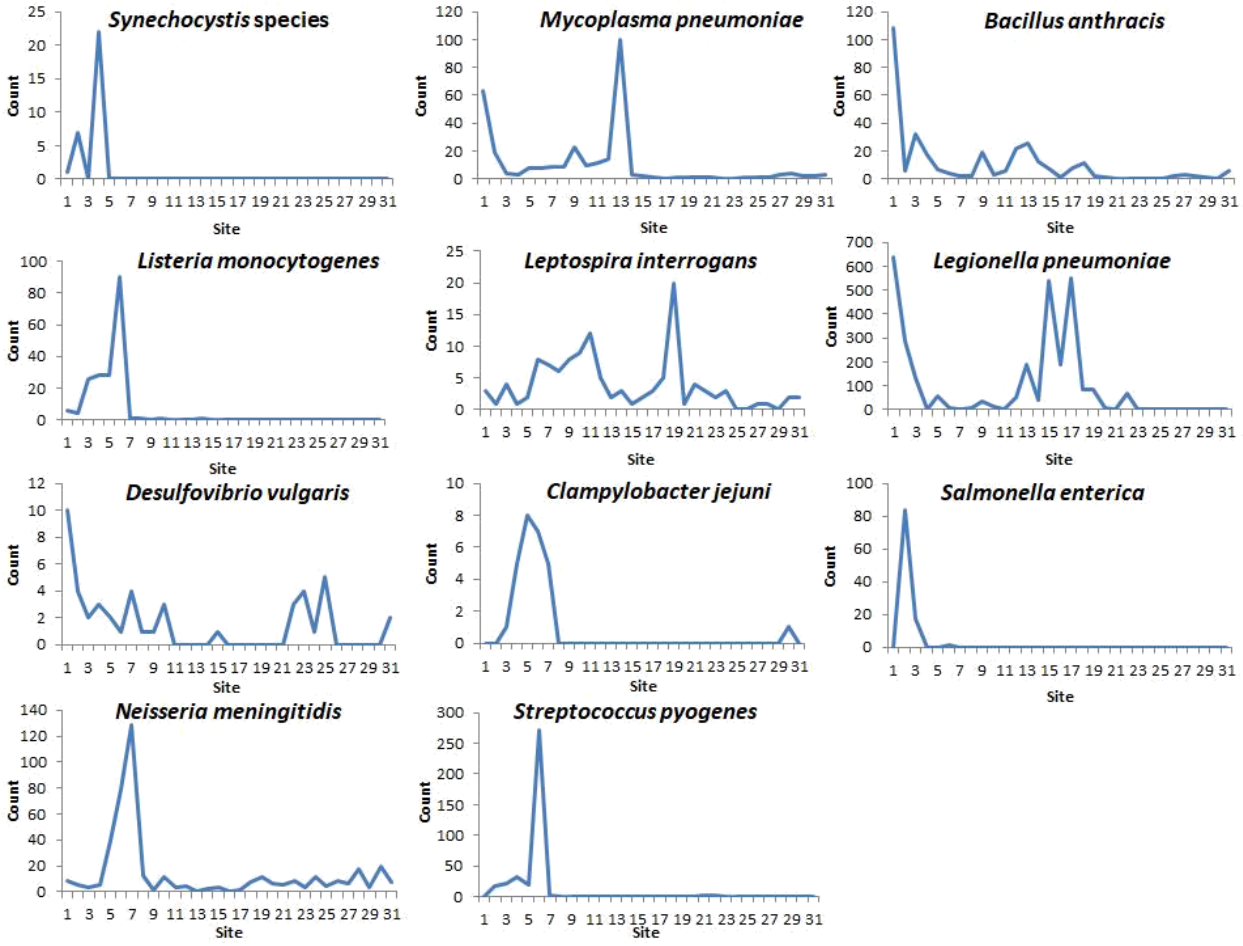
Xia X., 2018a Bioinformatics and the Cell, pp. 173–195 in *Bioinformatics and the Cell*, Copyright Holder Springer Science+Business Media LLC eBook ISBN 978-3-319-90684-3 DOI 10.1007/978-3-319-90684-3 Hardcover ISBN 978-3-319-90682-9 Edition Number 2 Number of Pages XIII, 489 Number of Illustrations 63 b/w illustrations, 59 illustrations in c.

Xia X., 2018b DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. Mol. Biol. Evol. 35: 1550–1552.

Zheng G., Y. Qin, W. C. Clark, Q. Dai, C. Yi, *et al.*, 2015 Efficient and quantitative high-throughput tRNA sequencing. Nat. Methods 12: 835.

Supplementary Information

Supplementary Fig. S2.1. Sequence alignment between the annotated 16S rRNA (rrnA16S, NC_000964) and 75 nt RNASeq reads (SRR3466199). RNA-Seq segment IDs are in FASTA+ format (e.g. SeqGr134077_56 represents the 134077th unique sequence, and it has 56 identical copies).



Supplementary Fig. S3.1. Counts of 3' ends from RNA-Seq reads mapped to the genomic 16S rRNA. Site 1 corresponds with the 3' terminal C in the canonical CCUCC aSD motif, and site 31 represents the 30th base downstream of CCUCC.