The Impact of Ribosome Binding Site Sequences on Translation in Mycobacteria

A Major Qualifying Project

Submitted to the Faculty of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Bachelor of Science

In

Biology & Biotechnology

By:

Catherine Masiello

April 28th, 2022

Approved by:

Scarlet S. Shell, PhD

Abstract

Tuberculosis (TB), caused by the bacteria *Mycobacterium tuberculosis* (Mtb), is one of the world's deadliest infectious diseases. One mechanism by which Mtb can develop resistance to current TB treatments and adapt to stress conditions in a host is through regulation of gene expression. Translation efficiency can impact gene expression and is itself influenced by characteristics of a gene's ribosomal binding site (RBS) within the 5' UTR. Here we used both computational and experimental approaches to investigate RBS characteristics to determine their impact on translation efficiency in the model organism *Mycolicibacterium smegmatis*. We focused on the presence of Shine-Dalgarno (SD) sequence motifs and secondary structure. Sequences complementary to the loops of the B11 sRNA, which is thought to play a role in the regulation of translation efficiency in mycobacteria, were inserted in the RBS of fluorescent reporters. The insertion of these sequences decreased fluorescence in both wild-type and ΔB11 strains, but for one of them the decrease was greater in the wild-type strain, consistent with the idea that B11 negatively regulates translation. We then analyzed the relationship between SD motifs and ribosome occupancy. When controlling for secondary structure we found that the presence of any SD motif with 4 nucleotides or more of complementarity to the anti-SD was associated with higher ribosome occupancy. However, there were not differences in ribosome occupancy among genes with SD motifs of different lengths or with complementarity to different parts of the anti-SD. We also assessed the impact of SD motif location on ribosome occupancy. We found that there seems to be an optimal position for the SD motif relative to the start codon that is not greatly affected by which part of the anti-SD the SD binds to. Our results may be used in further investigation of the role of B11 and in future plasmid designs to maximize gene expression.

Introduction

Tuberculosis (TB), caused by the bacteria *Mycobacterium tuberculosis* (Mtb), is the world's deadliest infectious disease aside from COVID-19, infecting 10 million people and causing 1.5 million deaths per year (World Health Organization, 2021). Though it can be curable, the antibiotic treatment process for TB takes a minimum of six months, involves the use of multiple drugs, and can lead to unwanted side effects in patients (World Health Organization, 2021; Center for Disease Control and Prevention, 2016). Additionally, the rise of antibiotic resistance in Mtb renders current treatment methods less effective in patients infected with resistant strains (Streicher et al, 2012). In order to develop new treatments, it is important to understand basic mechanisms in bacterial species that may be targeted through therapeutics (Casali et al, 2014; Olson et al, 2011). One such mechanism of interest is the regulation of gene expression, which is key in allowing bacterial pathogens, such as Mtb, to adapt to stress conditions within a host during infection (Mvubu et al, 2016). Translation efficiency is one factor that influences gene expression, both through impacting the amount of protein made in a cell and the transcription rate of genes through the coupling of translation and transcription in bacteria, making it a feature of interest for such research (Proshkin et al, 2010; Iyer et al, 2018).

In leadered genes, the process of translation requires the binding of a ribosome to the ribosome binding site (RBS) in the 5' untranslated region (UTR). The RBS contains the Shine-Dalgarno sequence (SD), located about eight or nine nucleotides upstream of the start codon, which the ribosome recognizes and binds to. The ribosome's small subunit contains the anti-Shine-Dalgarno sequence (anti-SD) in its 16S rRNA, which is complementary to the SD and allows the ribosome to bind to the mRNA in the correct location and initiate translation. While the anti-SD is highly conserved among different bacterial species, the SD varies among genes (Nakagawa et al, 2010). When a particular variation of the SD is found in multiple genes, located

at an appropriate distance from the gene's start codon, and is substantially complementary to the anti-SD of the ribosome, it is called an SD motif. Previous research by Saito *et al* has shown a linear relationship in *E. coli* genes between the presence of strong SD motifs and ribosome occupancy, a measure of translation efficiency that allows researchers to determine how many ribosomes are bound to mRNA (Saito et al, 2020; Ingolia et al, 2009). While particular SD motifs are not required for translation initiation, the study suggested they may impact translation efficiency (Saito et al, 2020; Weyens et al, 1988; Ma et al, 2020). A study conducted by Hockenberry *et al* determined that *E. coli, B. subtilis,* and *C. crescentus* genes with intermediate to strong binding strength between their SD and the anti-SD had a higher translation efficiency than genes with weaker binding strength (Hockenberry et al, 2017). In this study, the researchers took into consideration other SD factors besides the sequence and its complementarity to the anti-SD when analyzing the data that are thought to influence translation, such as the distance between the SD sequence and the start codon impacting the optimal location for anti-SD binding of specific sequences and secondary structure around the start codon influencing SD accessibility (Hockenberry et al, 2017; Komarova et al, 2020; Mao et al, 2014).

The presence of secondary structure around the RBS is a characteristic that is known to impact translation efficiency and gene expression (Nguyen et al, 2020; Mao et al, 2014). It can prevent translation by decreasing the accessibility of the region to the ribosome (de Smit & van Duin, 1990; Tholstrup et al, 2012). Research by Studer & Joseph provided evidence of an inverse correlation between secondary structure levels and the association of the ribosome to mRNA, lending to the theory that more stable secondary structure can decrease translation efficiency (Studer & Joseph, 2006; Kudla et al, 2009) Another study by Del Campo *et al* found that the area around the suspected SD upstream of the start codon was the most unstructured region of *E. coli* genes (Del Campo et al, 2015). The study further presented ribosome profiling data to support

the idea that this area lacks structure so that the ribosome may bind for translation (Del Campo et al, 2015). In some cases, however, secondary structure can also increase translation. In *E. coli* for instance, stem loops found in the coding sequence promote translation of the *bamA* gene (Jagodnik et al, 2017).

Alongside those features of the genes themselves, small non-coding RNAs (sRNAs) play a role in influencing translation efficiency. sRNAs can bind to the 5' UTR of mRNA and influence the stability of the mRNA or block binding of the ribosome to the RBS, which may prevent or lower translation (Baker et al, 2002; Baker et al, 2020; Thomason et al, 2019; Liu et al, 1995; Wang et al, 2005; Heidrich et al, 2007; Jagodnik et al, 2017). Given their effect on translation, sRNAs have been identified in bacteria as gene expression regulators. For instance, the MicA sRNA in *E. coli* binds to the *ompA* RBS and prevents ribosome binding for translation of *ompA* into the OmpA outer membrane protein (Udekwu et al, 2005), while the Mcr7 sRNA in Mtb controls the Tat secretion system, which impacts the function of other secretion systems associated with virulence, by preventing ribosome binding and translation of *tatC* (Solans et al, 2014).

One mechanism that sRNAs use to regulate gene expression is through binding to sequences within a gene's RBS that are complementary to the sRNA's own sequence. Research by Bar-Oz *et al* suggests that the B11 sRNA uses this mechanism of gene regulation in *Mycobacteroides abscessus*, a bacterial pathogen that can cause infections and severe decline in the lung function of cystic fibrosis patients (Bar-Oz et al, SUBMITTED; Center for Disease Control and Prevention, 2010; Qvist et al, 2016). Some clinical isolates of *M. abscessus* have been shown to have low levels of B11 expression or mutations in B11, and the sRNA affects multiple pathogenic characteristics of the bacteria, such as cell morphology and drug resistance (Bar-Oz et al, SUBMITTED). B11 may play a role in regulating the ESX-4 secretion system as

well, which is a virulence factor similar to the ESX-1 secretion system of Mtb (Bar-Oz et al, SUBMITTED; Laencina et al, 2018). B11 has two single-stranded loops of nine and twelve nucleotides in length, that are rich in C nucleotides and can bind to complementary sequences found in target mRNAs (Bar-Oz et al, SUBMITTED; Mai et al, 2019). The RBS of *M. abscessus* genes associated with the ESX-4 secretion system contain these B11-complementary sequences, suggesting that B11 may be able to bind to the RBS and block ribosomal binding to said genes (Bar-Oz et al, SUBMITTED).

The mechanism of gene regulation theorized in *M. abscessus* may be applicable to other mycobacteria, as B11 is conserved in many mycobacteria species, including both Mtb and *M. smegmatis* (Arnvig & Young, 2009). A study by Mai *et al* provides evidence for this method of gene regulation in *M. smegmatis*, where they showed that the binding of the B11 loops to stretches of complementary G nucleotides in mRNA 5' UTRs allowed the sRNA to regulate expression of the *panD* and *dnaB* genes by decreasing translation efficiency (Mai et al, 2019). The influence of the complementary sequence's location within the RBS, its length, and the presence of complementary nucleotides besides the stretch of G nucleotides remains to be determined however. Additionally, it is unknown if the presence of other RBS characteristics, such as SD motifs or secondary structure, influence the level of impact on translation efficiency.

Here we study the impact of RBS characteristics on mycobacterial translation efficiency using *M. smegmatis* as a model. To investigate the influence of different SD sequences, we created plasmid constructs with a YFP fluorescent reporter, a fixed Pmyc1 promoter, and an altered SD sequence to insert into our *M. smegmatis* strains. After insertion of the plasmids, we quantified fluorescence as a metric of translation efficiency. Through this technique, we additionally examined the direct relationship between the binding of B11 to *M. smegmatis* RBS and changes in translation efficiency. We found that the presence of B11-complementary

sequences in the Pmyc1 RBS decreased translation efficiency, as demonstrated by a lower level of fluorescence in *M. smegmatis* strains with the altered Pmyc1 RBS compared to strains with unaltered Pmyc1 RBS, though we were unable to conclusively determine that the direct binding of B11 to these complementary sequences caused the change in YFP expression. Additionally, we analyzed the relationships between *M. smegmatis* 5' UTR characteristics, such as SD motif presence, motif location, and secondary structure levels, and measures ribosome occupancy and half-life as measures of translation and mRNA stability respectively. We found that the presence of any SD motif may increase ribosome occupancy but no one motif was better than another. We also determined that our initial metric of determining motif location based on an alignment with a theoretically perfect SD sequence may not be relevant for optimal anti-SD binding.

Materials and Methods

Shine-Dalgarno Motif Identification

20 nucleotides upstream of the start codon of all leadered *M. smegmatis* genes, found in the genome annotation in Supplemental Table S1, were run through MEME on the MEME Suite website. The software identified motifs found within the uploaded sequences. Motifs with significant p-values ($\alpha = 0.05$) were further examined. Using Python, the position of the starting nucleotide of the motif and the position of the nucleotide in the motif that is aligned with a theoretically perfect SD sequence were identified (Nguyen et al, 2020). Supplemental Figure S1 displays the alignment of motifs with the theoretically perfect SD sequence, with the aligned nucleotide position for each motif bolded (Appendix A). GraphPad Prism was then used to make frequency distributions of the aligned nucleotide position for each of the significant motifs. Based on the peak in the frequency distributions, motifs that were over-represented within the suspected SD region of -17 to -4 upstream of the start codon were selected for further analysis.

Supplemental Table S2 lists these chosen motifs, the genes containing the motif, the position of the motif's starting nucleotide, and the position of the motif's aligned nucleotide.

G and A Nucleotide Content Analysis

Microsoft Excel was used to identify the occurrence of G and A nucleotides within the region of -17 to -4 upstream of the start codon in all leadered *M. smegmatis* genes (Supplemental Table S1). The occurrence of individual G and A nucleotides were used to calculate the percentage of G's and A's making up the region. Microsoft Excel was also used to identify occurrences of GA di-nucleotide motifs (GA, AG, GG, or AA) and return a count of all di-nucleotide repeats found within overlapping 2-nucleotide windows in the -17 to -4 region for each leadered gene. Supplemental Table S3 contains the G and A percentage and GA di-nucleotide motif counts for the *M. smegmatis* leadered genes.

Feature Table Analysis

A feature table of *M. smegmatis* characteristics, compiled by Shell lab member Huaming Sun and found in Supplemental Table S4, was analyzed to determine the relationship between different characteristics of leadered *M. smegmatis* genes. All features pertaining to secondary structure were divided into four groups based on the quartiles of genes with particular values for the feature. Quartiles for each feature were determined using GraphPad Prism. A Python script was used to categorize genes and their associated feature table data into these groups for analysis.

Using GraphPad Prism, Spearman Correlation, Kruskal-Wallis, Dunn's Multiple Comparison, Kolmogorov-Smirnov, and Mann-Whitney tests were run to determine the statistical significance of the relationship between features based on their groupings.

Additionally, GraphPad Prism was used to make histograms and scatter plots using the data from significant relationships. Supplemental Table S5 lists all features that were tested, the groupings of the feature values, and the results of each test.

Shine-Dalgarno Motif Location Analysis

When analyzing the influence of SD motif location on translation efficiency, two metrics were used to determine motif location. The first metric involved the use of aligning all motifs of interest in accordance with a theoretically perfect SD sequence (see Supplemental Figure S2a) (Appendix B). The number of nucleotides between the aligned nucleotide position and the start codon was used as the motif location for any analysis with this metric (see Supplemental Figure S2b) (Appendix B). The second distance metric used the number of nucleotides between the 3' end nt of the motif and the start codon as the motif location. For both metrics, motifs and their locations were searched for within a range of 20 nucleotides upstream of the start codon.

For all analyses pertaining to motif location, a Python script and Microsoft Excel were used to remove any genes that contained multiple occurrences of the same motif or different motifs in different locations within the 20 nucleotide range.

Bacterial Strains and Growth Conditions

In liquid culture, *M. smegmatis* strains were grown at 37℃ in Difco Middlebrook 7H9 media containing 0.2% glycerol, 0.05% Tween 80, and Albumin Dextrose Catalase (ADC). *E. coli* strains were grown at 37℃ in LB broth. On solid media, *M. smegmatis* was plated on 7H10 containing 0.5% glycerol and ADC and incubated at 37℃, while *E. coli* was plated on LB agar and incubated at 37℃. For both liquid cultures and solid media, the appropriate antibiotics were added based on the bacteria strain and species. *M. smegmatis* antibiotic concentrations were 250

μg/mL hygromycin B (Hyg) and 25 μg/mL kanamycin (Kan). *E. coli* concentrations were 200 μg/mL Hyg and 50 μg/mL Kan.

The wild-type strain was mc^2155, and the ΔB11 strain was a gift from Daniel Barkan and derived from mc^2155.

Identification of B11 Loop Complementary Sequences

The nine nucleotide B11 loop 1 sequence is GACCCCCCG and the 12 nucleotide B11 loop 2 sequence is UCCUCCCCCCCU. The loop sequences are conserved between *M. tuberculosis*, *M. smegmatis*, and *M. abscessus*. Sequences complementary to these two loops were identified in Microsoft Excel using an annotation of the *M. abscessus* genome, found in Supplemental Table S6 (Bar-Oz et al, SUBMITTED). The 50 nucleotides upstream of the start codon of each leadered gene was split into two 25 nucleotide regions. Subsections of each loop complementary sequence, listed below in Table 1, were searched for and counted in each region using Microsoft Excel. Supplemental Table S7 lists all *M. abscessus* genes and any complementary sequences found in either of the two 25 nucleotide regions.

Table 1: *M. abscessus* B11 Loop Complementary Sequences

Loop	**Sequence Length (nt)**	**Complementary Sequence**
Loop 1	9	CGGGGGGTC
	8	GGGGGGTC
		CGGGGGGT
	7	GGGGGGT
		GGGGGTC

		CGGGGGG
	6	GGGGTC
		GGGGGT
		GGGGGG
		CGGGGG
	5	GGGTC
		GGGGT
		GGGGG
		CGGGG
Loop 2	12	AGGGGGGGAGGA
	11	AGGGGGGGAGG
		GGGGGGGAGGA
	10	AGGGGGGGAG
		GGGGGGGAGG
		GGGGGGAGGA
	9	AGGGGGGGA
		GGGGGGGAG
		GGGGGGAGG
		GGGGGAGGA
	8	AGGGGGGG

		GGGGGGGA
		GGGGGGAG
		GGGGGAGG
		GGGGAGGA
	7	AGGGGGG
		GGGGGGG
		GGGGGGA
		GGGGGAG
		GGGGAGG
		GGGAGGA
	6	AGGGGG
		GGGGGG
		GGGGGA
		GGGGAG
		GGGAGG
		GGAGGA
	5	AGGGG
		GGGGG
		GGGGA
		GGGAG

		GGAGG
		GAGGA

Table 1 lists all of the complementary sequences for B11 loop 1 and loop 2 searched for in the *M. abscessus* genome annotation.

Plasmid Construction

HiFi Assembly (NEB) was used to construct plasmids to alter *M. smegmatis* strains. Two types of plasmids were constructed: plasmids integrating at the Giles site and plasmids integrating at the L5 site, which are listed below in Table 2. For the Giles site plasmids, which were used to introduce B11 loop complementary sequences into the strains, a pSS303 backbone (Nguyen et al, 2020) containing Hyg resistance, a fixed Pmyc1 promoter, and a YFP fluorescence marker was used. The pSS303 was isolated and purified from a frozen glycerol *E. coli* stock (SS-E_0195) using a ZR Plasmid Miniprep Kit. The insert sequence for the plasmids were the B11 loop complementary sequences, an eight nucleotide sequence of GGGGGGTC for loop 1 or a nine nucleotide sequence of AGGGGGGGA for loop 2. These sequences were the longest B11-complementary sequences identified in *M. abscessus* genes (Supplemental Table S6). Since these inserts were short, primers containing the entirety of the insert with overhangs for the backbone were designed in Benchling, found below in Table 3. These primers were used in Q5 polymerase chain reaction (PCR) to amplify the backbone and integrate the complementary sequences into the plasmid directly upstream or two nucleotides upstream of the YFP start codon in two separate fragments (see Table 4) for B11 loop 1 or loop 2 respectively. The total volume of Q5 PCR reaction was 50 µL, containing 22.5 µL of H_2O, 10 µL of Q5 enhancer, 10 µL of Q5 buffer, 2.5 µL of the respective forward primer at 10 mM, 2.5 µL of the respective reverse primer at 10 mM, 1 µL of 10 mM each dNTPs, 0.5 µL of Q5 polymerase, and

1 µL of template pSS303 plasmid at 1 to 5 ng/µL. In the thermocycler, the template was denatured at 98℃ for 2 minutes, followed by 35 cycles of 98℃ for 20 seconds, 68℃ for 30 seconds, and 72℃ for 3 minutes. After completion of the 35 cycles, the reaction elongated at 72℃ for 5 minutes and was then stored at 4℃. 1 µL of Dpn1 was added to the Q5 PCR products and incubated at 37℃ for 30 minutes and 80℃ for 20 minutes to cleave any methylated template in the product. To check the success of the plasmid amplification, the PCR products were run through a 1% TAE gel. Bands of the expected size were cut from the gel and the DNA was extracted and purified using a Zymoclean Gel DNA Recovery Kit. To perform the HiFi Assembly with a total volume of 6 µL, 3 µL of 2x HiFi DNA Assembly Master Mix was added to 2.5 µL of the long plasmid fragment and 0.5 µL of the short plasmid fragment (see Table 4). The sample was incubated at 50℃ for 2 hours, then transformed into *E. coli*. A ZR Plasmid Miniprep Kit was used to purify the plasmid from the *E. coli* culture, which was then sent for Sanger sequencing to confirm proper integration of the loop complementary sequences and a lack of mutations in the YFP gene.

For the L5 site plasmids, which were used for complementation of *M. smegmatis* B11 into ΔB11 strains, a pJEB402 backbone (Lee, Pascopella, Jacobs, & Hatfull, 1991) with Kan resistance and the MOP promoter removed was used. pJEB402 was isolated and purified from a frozen glycerol *E. coli* stock (SS-E_0062) using a ZR Plasmid Miniprep Kit. The insert sequence, which was integrated where the MOP promoter had originally been located in pJEB402, was the predicted *M. smegmatis* B11 sequence, 237 base pairs downstream of the sequence, and 200 base pairs upstream of the sequence. This sequence was selected by inspection of RNAseq coverage using the Artemis software. 20 nucleotide primers on either side of the MOP promoter to amplify the pJEB402 backbone and 20 nucleotide primers for the insert sequence with additional 20 nucleotide overhangs complementary to the backbone were

designed in Benchling (see Table 3). The general process of running Q5 PCR, Dpn1, a 1% TAE gel for DNA recovery, HiFi Assembly, and sequencing was implemented as described above with a few changes in the procedure. For Q5 PCR thermocycler settings, amplification of the backbone had an annealing step at 61℃ and an elongation step of 2.5 minutes while amplification of the insert had an annealing step of 64℃ and an elongation step of 10 seconds. For the HiFi Assembly, the reaction was done with a total of 6 µL that contained 3 µL of 2x HiFi DNA Assembly, 1.5 µL of the backbone at 25 ng/µL, and 1.5 µL of the insert at 25 ng/µL.

Table 2: Plasmids

Plasmid	**Integration Site**	**Description**
pSS303	Giles	YFP marker, Hyg resistance, Pmyc1 promoter
pSS314	Giles	YFP marker, Hyg resistance
pSS624	Giles	pSS303 backbone B11 loop 1 complementary sequence insert (GGGGGGTC)
pSS625	Giles	pSS303 backbone B11 loop 2 complementary sequence insert (AGGGGGGGA)
pJEB402	L5	Kan resistance, MOP promoter
pSS626	L5	pJEB402 backbone with MOP promoter removed *M. smegmatis* B11 gene insert

Table 2 lists and describes all of the plasmids used to alter and construct *M. smegmatis* strains. It contains the plasmid name, its integration site, and any other significant features of the plasmid pertaining to its function.

Table 3: Primers

Primer	Sequence	Description
SSS1376	TAGGGCGTTGCCTCAATCG	Forward primer to amplify pSS303 backbone
SSS1377	GGCAACGCCCTAGTGATGGTGATGGTGATGAC	Reverse primer to amplify pSS303 backbone
SSS2410	CGGGGGGTCATGGCCAGCGATAGCACTGAGAGC	Forward primer to introduce B11 loop 1 complementary sequence into pSS303
SSS2411	GCCATGACCCCCCGCTCCTTCTTAATTAAGCATGCGGATCGTG	Reverse primer to introduce B11 loop 1 complementary sequence into pSS303
SSS2412	AGGGGGGGATCATGGCCAGCGATAGCACTGAGAG	Forward primer to introduce B11 loop 2 complementary sequence into pSS303
SSS2413	CCATGATCCCCCCCTCCTTCTTAATTAAGCATGCGGATCGTGCT	Reverse primer to introduce B11 loop 2 complementary sequence into pSS303
SSS1412	GGAAAAGAGGTCATCCAGGAAGAAATATTGGATCGTCGGC	Forward primer for sequencing B11 loop complementary sequences in pSS303
SSS248	ACGCCCCTCTAGCTGATCAC	Reverse primer for sequencing B11 loop complementary sequences in pSS303
SSS1172	CTCCGAACTCCTCCGAAACC	Forward primer for checking left junction integration in the Giles site
SSS1174	TGACGATCAACTCCGCGGGGCCGGGCCA	Reverse primer for checking left junction integration in the Giles site
SSS1175	CGGTGGATCCGCGCAACCTG	Forward primer for checking right junction integration in the Giles site
SSS1173	ACATATCTGTCGAAGCGCCC	Reverse primer for checking right junction integration in

		the Giles site
SSS2494	TGCAGAATTCGAAGCTTATC	Forward primer to amplify pJEB402 backbone without MOP promoter
SSS2495	GTCTAGATATGACGACAGGA	Reverse primer to amplify pJEB402 backbone without MOP promoter
SSS2492	TCCTGTCGTCATATCTAGACgtgttcctttgcttgtctgt	Forward primer to introduce B11 into pJEB402
SSS2493	GATAAGCTTCGAATTCTGCAgccgcgatcgattcctcagc	Reverse primer to introduce B11 into pJEB402
SSS132	CCTGATTCTGTGGATAACCG	Forward primer for sequencing B11 insert in pJEB402
SSS1103	TGGATTTGGTTTCAGCTCCC	Forward primer for checking left junction integration in the L5 site
SSS142	TAGAGCCGTGAACGACAGG	Reverse primer for checking left junction integration in the L5 site
SSS2527	tctgcaccacggtcgtgatc	Forward primer for checking right junction integration in the L5 site
SSS2528	CGGCGTCCATCTTGTTGTCG	Reverse primer for checking right junction integration in the L5 site

Table 3 lists the primers used to make the plasmids and any primers used for checking proper plasmid integration or sequencing. It contains the primer name, its sequence, and a brief description.

Table 4: B11 Loop Plasmid Constructs

Plasmid	Fragment	Primers	Fragment Size (bp)
pSS624	Fragment 1	SSS1377, SSS2410	762
	Fragment 2	SSS1376, SSS2411	5751
pSS625	Fragment 1	SSS1377, SSS2412	764
	Fragment 2	SSS1376, SSS2413	5749

Table 4 details the fragments used to make the complete B11 complementary loop plasmids with the pSS303 backbone.

Strain Construction

To make the *M. smegmatis* strains listed below in Table 5, 10 to 200 ng of the appropriate plasmid (see Table 2) was transformed into wild type or ΔB11 *M. smegmatis* competent cells using electroporation at 2.5 kV with an incubation time of three hours at 37℃ before plating. After three days of incubation at 37℃ on 7H10 antibiotic plates, colonies were screened using Taq PCR to check for integration of the plasmid. The total volume of Taq PCR reaction was 10 µL, containing 7.35 µL of H2O, 0.5 µL of DMSO, 0.2 µL of the respective forward primer at 10 mM, 0.2 µL of the respective reverse primer at 10 mM, 0.2 µL of 10 mM each dNTPs, 1 µL of 10x Standard Taq buffer, 0.05 µL of Taq polymerase, and 1 µL of template from a colony on the plate. In the thermocycler, the template was denatured at 95℃ for 5 minutes then, for 35 cycles, was denatured at 95℃ for 30 seconds, annealed at 55℃ or 54℃ for 30 seconds for Giles site or L5 site integration respectively, and elongated at 68℃ for 1 minute. After completion of the 35 cycles, the reaction elongated at 68℃ for 5 minutes and was then stored at 4℃. The PCR products were run through a 1% TAE gel to identify bands of the appropriate size signifying integration of the plasmid into the strain. The expected sizes of the left and right junction for

Giles site integration is 582 base pairs (bp) and 378 bp. The expected sizes of the left and right junction for L5 site integration is around 411 bp and 404 bp.

Table 5: Bacterial Strains

Strain	**Description**
WT	Wild-type *M. smegmatis*
SS-M_0786	ΔB11 *M. smegmatis*
SS-M_0889 SS-M_0890 SS-M_0944 SS-M_0945 SS-M_0946	SS-M_0786 transformed with pSS314
SS-M_0942 SS-M_0943	SS-M_0786 transformed with pSS303
SS-M_0940	Wild-type *M. smegmatis* transformed with pSS303
SS-M_0941	Wild-type *M. smegmatis* transformed with pSS314
SS-M_1075 SS-M_1076	Wild-type *M. smegmatis* transformed with pSS624
SS-M_1077 SS-M_1078	SS-M_0786 transformed with pSS624
SS-M_1079 SS-M_1080	Wild-type *M. smegmatis* transformed with pSS625
SS-M_1081 SS-M_1082 SS-M_1083	SS-M_0786 transformed with pSS625
SS-M_1147 SS-M_1148 SS-M_1149	Wild-type *M. smegmatis* transformed with pSS624 and pJEB402
SS-M_1153 SS-M_1154 SS-M_1155	Wild-type *M. smegmatis* transformed with pSS625 and pJEB402

SS-M_1150 SS-M_1151 SS-M_1152	SS-M_0786 transformed with pSS624 and pSS626
SS-M_1156 SS-M_1157 SS-M_1158	SS-M_0786 transformed with pSS625 and pSS626
SS-E_0195	*E. coli* Used to isolate and purify pSS303
SS-E_0215	*E. coli* Used to isolate and purify pSS314
SS-E_0062	*E. coli* Used to isolate and purify pJEB402
SS-E_0539	*E. coli* transformed with pSS624
SS-E_0540	*E. coli* transformed with pSS625
SS-E_0541	*E. coli* transformed with pSS626

Table 5 lists all *M. smegmatis* and *E. coli* strains used. It contains the strain name and its description. All *E. coli* strains were used to isolate and purify plasmids for use in the *M. smegmatis* strains.

Fluorescence Microscopy

Fluorescence microscopy was used to ensure that YFP fluorescence of wild type *M. smegmatis* and ΔB11 *M. smegmatis* integrated with pSS303 or a negative fluorescence control of pSS314, containing a YFP fluorescence marker, Hyg resistance, and no Pmyc1 promoter, was the same before alteration of their ribosomal binding sites. Cultures of SS-M_0940-0943 and SS-M_0945-0946 were grown to ODs of 0.6-0.8. 0.01% PBS Tween was used to dilute the cultures to an OD of 0.5. 5 μL of the diluted cultures were plated on 1% agar pads on microscope slides. The slides were observed using a 40x objective of a Zeiss Axio Imager Z1 Zeiss Apotome light microscope. Using the Zen software, the microscope was set to default GFP channel settings and a white value of 3500 for images.

Flow Cytometry

Flow cytometry was used to compare YFP fluorescence between wild-type and ΔB11 *M. smegmatis* strains with or without a B11 loop complementary sequence in the ribosomal binding site. Cultures of SS-M_1075-1082, SS-M_0940-0942, and SS-M_0944 were grown to an OD of roughly 0.8 and normalized with filtered 7H9 the day before analysis. 1 mL of the cultures were then diluted using the filtered 7H9 to an OD of 0.025, then filtered with 5 uM filter needles. The gain settings of the flow cytometer were FSC 500 and Violet SSC 50. The threshold settings were Violet SSC-H 100,000 and FSC-H 40,000. Roughly 3000 to 6000 event counts were collected for each strain.

All data collected from the flow cytometry were analyzed using FlowJo software. The fluorescence value in the software was set to FL1-H :: GFP-H for analysis. The data were processed twice, first with a gate containing roughly 6% of events for each strain and second with a gate containing roughly 3% of events for each strain.

Complementation

We attempted to complement the ΔB11 *M. smegmatis* strain by expression of B11 from the L5 site. A plasmid with a pJEB402 backbone, Kan resistance, and the native B11 promoter (as described earlier in the Plasmid Construction Section) was transformed into SS-M_0786 at the L5 site through the process described in the Strain Construction Section. Since transformation into the L5 site may have an impact on growth rate, an empty vector with a pJEB402 backbone, Kan resistance, and the MOP promoter was transformed into wild type *M. smegmatis*. The transformed strains were grown on 7H10 Kan agar. After confirming proper integration of both pJEB402 backboned plasmids into the L5 site, pSS624 and pSS625 were

transformed into the Giles site of both the complemented ΔB11 and WT strains and plated on 7H10 Hyg agar. All plasmids used and new strains made are listed above in Tables 2 and 5.

Upon integration of both the L5 and Giles plasmids into the ΔB11 and WT strains, fluorescence was measured by flow (see Flow Cytometry Section). Cultures SS-M_1075-1082 served as controls to be compared to duplicate cultures of the newly complemented strains.

We then constructed growth curves of the non-complemented and complemented wild-type and ΔB11 strains to compare their growth rates. Duplicate cultures of each strain were growth to an OD of between 0.4 and 0.5, then diluted to an OD of 0.001. 200 µL of each diluted culture were transferred to a 96 well plate, which was then placed in the BioTek Epoch 2 microplate reader to record the OD every ten minutes for 36 hours. The OD data from the plate reader were exported from the Gen5 software to be analyzed in GraphPad Prism.

Results

Insertion of sequences complementary to the B11 sRNA loops in the RBS decreases translation efficiency

The presence of sequences in the ribosomal binding site (RBS) that are complementary to sRNAs may provide a means for said sRNAs to bind to mRNA and regulate gene expression. If an sRNA is bound to an mRNA on the RBS, this may block the ribosome from binding and effectively initiating translation. We focused on testing the impact on fluorescence levels when the B11 sRNA could bind to B11 complementary sequences within a fluorescent reporter's RBS. Previous studies have found that B11 may regulate gene expression in *M. smegmatis* and *M. abscessus* by binding to sequences within a gene's RBS that are complementary to either of B11's two C-rich loops, thereby blocking the ribosome (Figure 1).

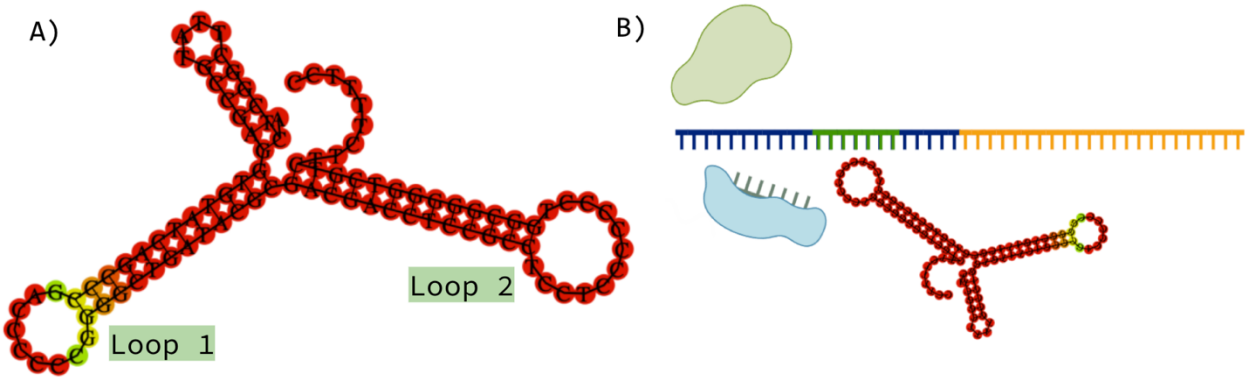
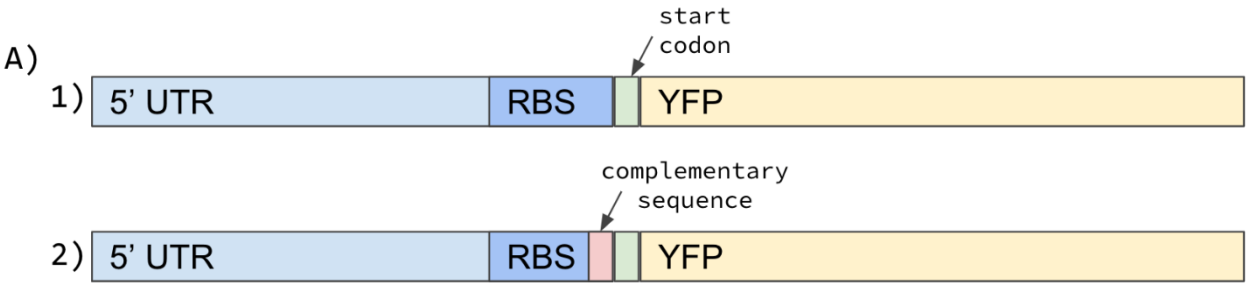


Figure 1: B11's loops may give the sRNA the ability to bind to mRNA's RBS. Figure 1 depicts the B11 sRNA. Panel A is the predicted structure of B11. Our designations for the two loops, loop 1 and loop 2, are labeled respectively. Panel B is a diagram of B11's predicted mechanism of binding to the RBS, thereby preventing ribosome binding. The blue portion of the mRNA strand represents the 5' UTR, the green portion is the RBS, and the yellow portion is the gene's coding sequence.

To investigate if the presence of sequences in the RBS that are complementary to B11's loops decreases translation efficiency, we replaced a portion of the Pmyc1 RBS in the pSS303 YFP reporter with a B11 loop complementary sequence (Ehrt et al, 2005; Nguyen et al, 2020). Previously, we investigated the presence of B11 complementary sequences in the 5' UTR of *M. abscessus* genes. We specifically looked at the last 50 nucleotides of each 5' UTR, which was then split into two 25 nucleotide regions. Between both these regions, we found that the longest sequence complementary to B11 loop 1 (GACCCCCCG) was an eight nucleotide sequence of CGGGGGGT and the longest sequence complementary to B11 loop 2 (UCCUCCCCCCCU) was a nine nucleotide sequence of AGGGGGGGA. All data pertaining to the presence of any B11 complementary sequence found in the *M. abscessus* genome annotation can be found in Supplemental Table S7. Given that the B11 loop sequences are conserved between *M. abscessus, M. tuberculosis,* and *M. smegmatis*, we used this finding in constructing our plasmids for this experiment (Figure 2). When designing our plasmids, we preserved the SD sequence of the

original fluorescent reporter's RBS (AAGAAGGAGA) (Figure 2B). The longest loop 1 and loop 2 sequences were used in these plasmid constructs, though plasmids containing different lengths of complementary sequences could be designed for future study.

A)



B)

Plasmid	Sequence
Pmyc1 RBS	TTAAGAAGGAGATATACATCATG
Pmyc1 RBS + Loop 1	TTAAGAAGGAGCGGGGGGTCATG
Pmyc1 RBS + Loop 2	TTAAGAAGGAGGGGGGGGATCATG

Figure 2: Design of fluorescent reporters with UTR sequences complementary to the loops of B11. Figure 2 depicts the B11 complementary sequence plasmid designs. Panel A displays a schematic of the plasmid design. A1 is the pSS303 plasmid without any alterations. A2 is represents the plasmids containing one of the B11-complementary sequences in the RBS. Panel B displays a table with a portion of the RBS sequence being altered in the different plasmids. The sequence highlighted in green represents the start codon of the YFP gene, while the sequence highlighted in pink represents the B11-complementary sequence that replaced a portion of the Pmyc1 RBS upstream of the start codon in pSS303 backbone.

We transformed pSS303 and pSS314 (a non-fluorescent negative control) into wild-type and ΔB11 strains of *M. smegmatis*. We then performed fluorescence microscopy to verify that

YFP was expressed from pSS303 at similar levels in both strain backgrounds (Figure 3). This confirmed that the pSS303 backbone did not intrinsically contain sequences that caused its expression to be affected by B11. It was noted that the ΔB11 strain grew slower and had a smaller cell size when compared to the wild-type strain. The ΔB11 strain was also clumpier than the wild-type strain.

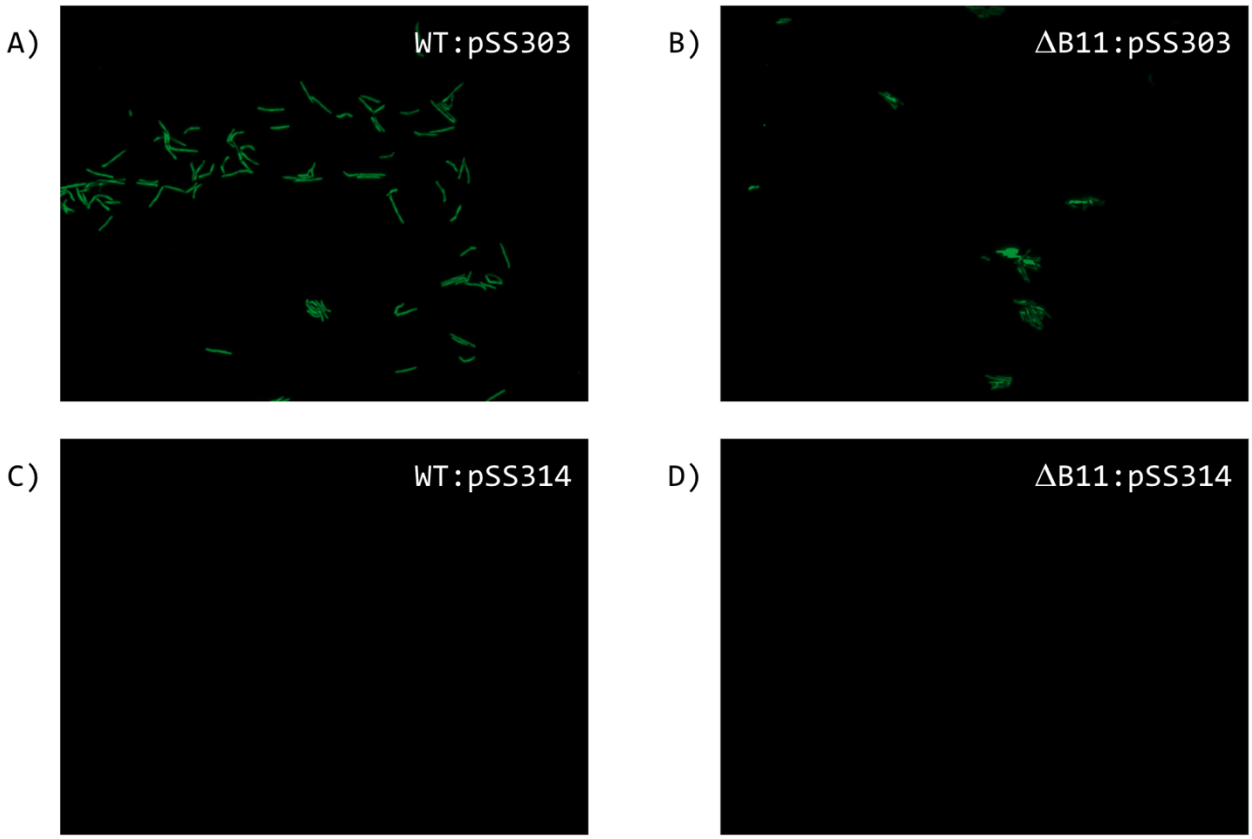


Figure 3: YFP fluorescence with a Pmyc1 RBS in *M. smegmatis* strains confirms the pSS303 backbone is not influenced by the presence of B11. Figure 3 displays images of the fluorescence microscopy of both wild-type and ΔB11 *M. smegmatis* strains to verify that expression of YFP from pSS303 was not affected by the absence of B11. pSS303 is therefore an appropriate backbone for testing the impact of RBS sequence changes. Panels A and B are the wild-type strain and ΔB11 strain transformed with pSS303 into the Giles site. The level of fluorescence for both strains was comparable.

Panels C and D are the wild-type strain and ΔB11 strain transformed with pSS314, which served as a negative fluorescence control. No fluorescence was observed in either strain.

After transformation of the B11 loop 1 and loop 2 plasmids into the wild-type and ΔB11 *M. smegmatis* strains, flow cytometry was performed to determine the level of fluorescence in each strain. Given the difference in cell size between the wild-type strains and the ΔB11 strains, the gates used in the analysis of the flow cytometry data were drawn to capture a significant portion of events for both strains. Additionally, the data were analyzed using two gates of different sizes, one which captured roughly 6% of events and another capturing 3% of events, to ensure that the trends in the data remained consistent even with more tightly gated populations. Examples of these gates can be seen in Figure 4.

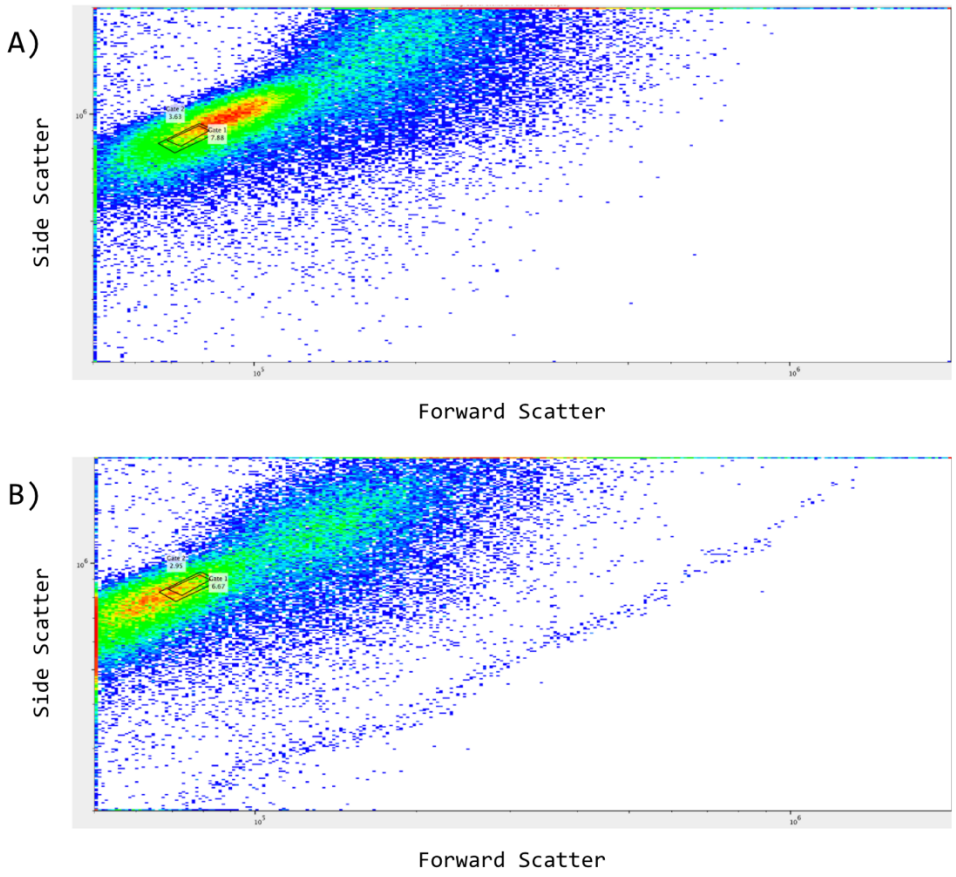


Figure 4: Gates for the flow cytometry data were chosen to allow for comparisons between the different cell sizes of wild-type and ΔB11 strains. Figure 4 depicts flow cytometry data for both the wild-type and ΔB11 strains. Forward scatter is on the x-axis and side scatter is on the y-axis. Forward scatter and side scatter both reflect cell size for bacterial cells. Panel A displays an example of the data used to determine the two gates in the wild-type strain transformed with the B11 loop 1 plasmid. Panel B displays a similar example but with the ΔB11 strain transformed with the B11 loop 1 plasmid.

We used flow cytometry to quantify the effect of the various RBS sequences on YFP levels. Both the wild-type and ΔB11 strains displayed a decrease in fluorescence when they contained either plasmid with a B11-complementary sequence in the RBS (see Figure 5). Strains containing the B11 loop 1-complementary sequence had a greater decrease in fluorescence than those containing the B11 loop 2-complementary sequence. For both B11 loop 1 and loop 2

plasmids, the wild-type strain had a greater decrease in fluorescence than the ΔB11 strain, though this difference between the strains was much more pronounced for B11 loop 2 than B11 loop 1.

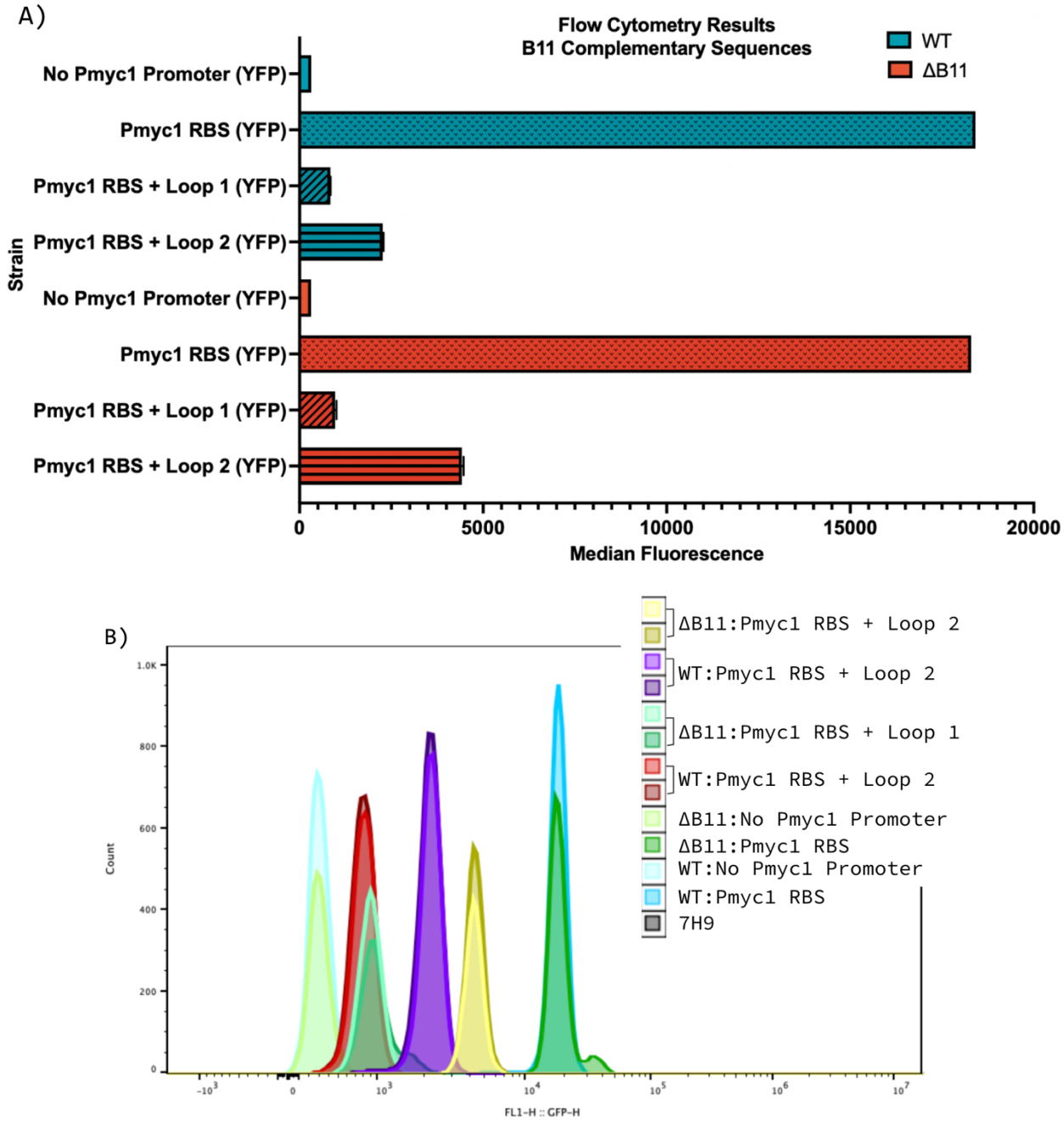


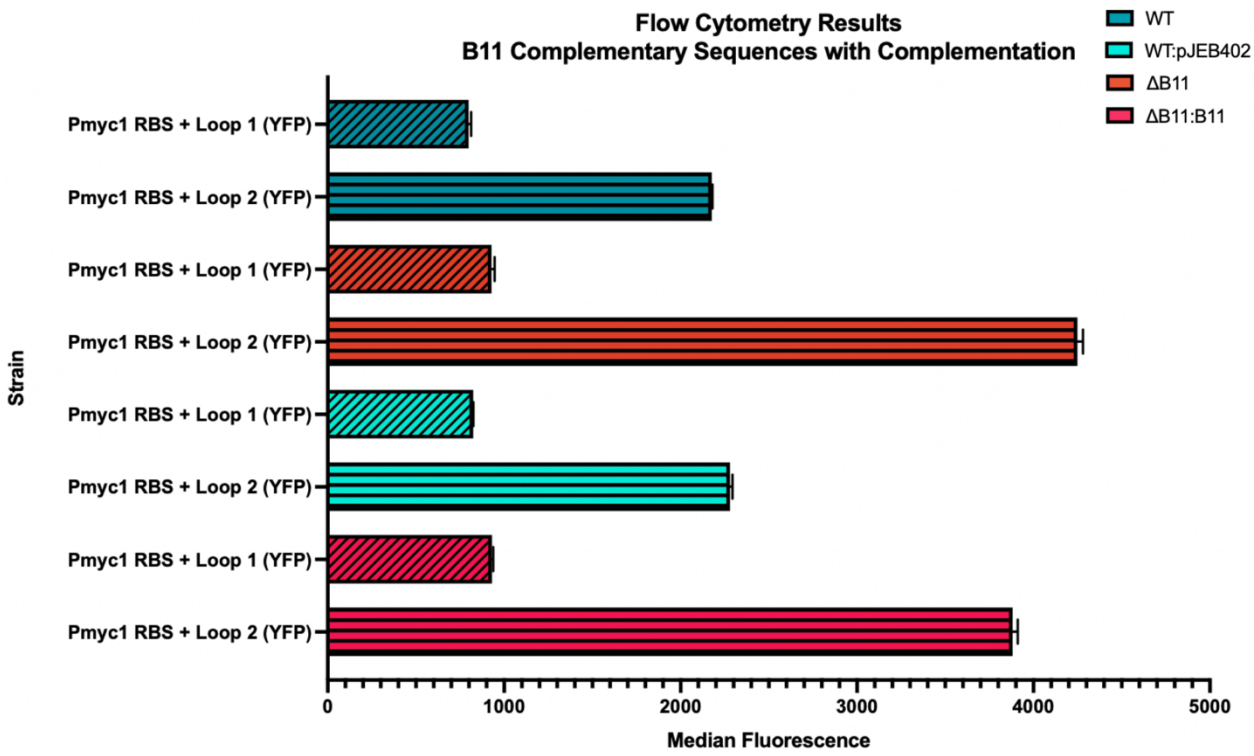
Figure 5: YFP fluorescence decreases in wild-type and ΔB11 *M. smegmatis* when the Pmyc1 RBS contains sequences complementary to the B11 sRNA loops. Figure 5 displays the flow cytometry results after transformation of the B11 loop 1 and loop 2 plasmids into wild-type and ΔB11 strains of *M. smegmatis*. Panel A displays a bar chart that contains the median fluorescence values for all wild-type (blue) and ΔB11 (red) strains using data from gate 1. Diagonal lines denote the B11 loop 1 construct. Horizontal lines denote the B11 loop 2 construct. Dots denote the unaltered Pmyc1 RBS. Error bars show

the standard deviation of the mean of the duplicate strains' median fluorescence values. Panel B displays the histogram generated by FlowJo of the fluorescence levels and event counts for each strain using data from gate 1.

From these results alone, we could not determine if B11 was directly causing the change in fluorescence by binding to the loop complementary sequences or if our alterations to the Pmyc1 5' UTR had created a less fit RBS. To address this, we attempted to complement the ΔB11 strain with *M. smegmatis* B11. If B11 was causing the decrease in fluorescence by binding to the complementary sequences in the RBS, we expected to see the YFP fluorescence levels of the complemented ΔB11 strains match those of the wild-type strains.

The suspected *M. smegmatis* B11 sequence and its predicted native promoter were inserted into a pJEB402 backbone in place of the MOP promoter. This construct was then transformed into the L5 site of the ΔB11 strain, while the original B11 loop constructs with the pSS303 backbone were transformed into the Giles site. Additionally, a pJEB402 empty vector was transformed into the wild-type strain to account for any potential impacts on growth rate that a plasmid in the L5 site may cause.

We repeated flow cytometry to quantify and compare YFP fluorescence of the complemented strains and the non-complemented strains. As with the first flow cytometry experiment, we analyzed our data using two different gates containing roughly 6% or 3% of events. We found that fluorescence levels of our complemented strains were equivalent to those of their non-complemented counterparts (Figure 6).



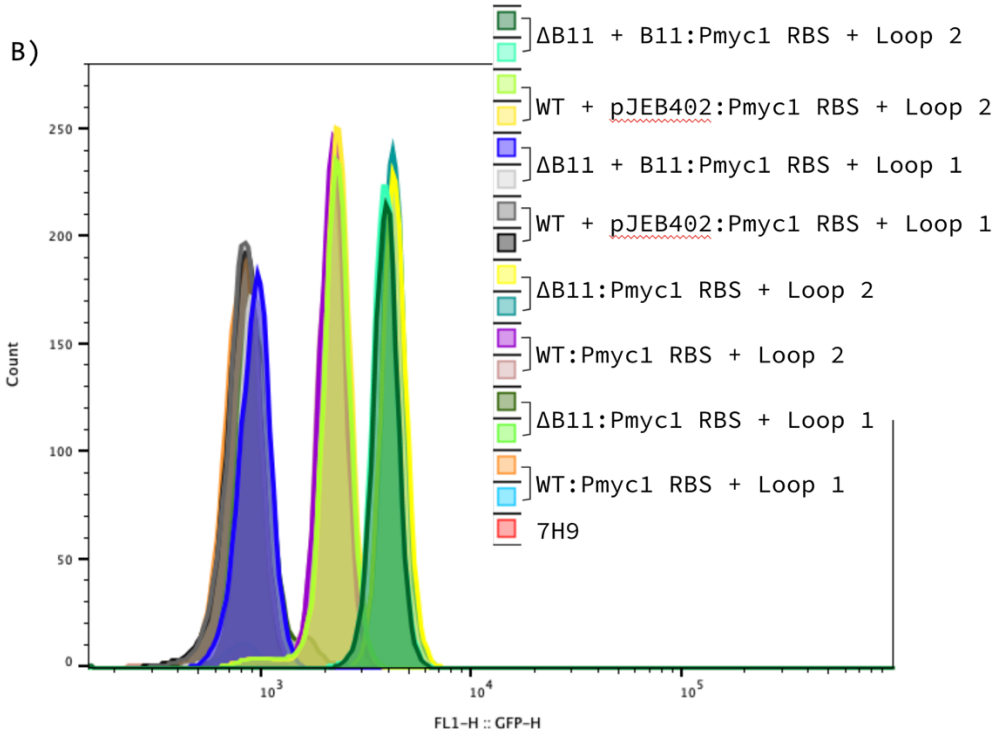


Figure 6: YFP fluorescence of complemented ΔB11 strains is comparable to the non-complemented strains, suggesting that our plasmid was unsuccessful in complementation. Figure 6 displays the flow cytometry results of complemented and non-complemented wild-type and ΔB11 *M. smegmatis* strains

after transformation of the B11 loop 1-complementary and loop 2-complementary reporter plasmids. Panel A displays a bar chart contains the median fluorescence values for all wild-type (blue) and ΔB11 (red) strains using data from gate 1. Diagonal lines denote the B11 loop 1 construct. Horizontal lines denote the B11 loop 2 construct. Error bars show the standard deviation of the mean of the duplicate strains' median fluorescence values. Panel B displays the histogram generated by FlowJo of the fluorescence levels and event counts for each strain using data from gate 1.

Given these results, we believe that our complementation plasmid did not actually express functional B11. When comparing the cell size data obtained through flow cytometry of the complemented wild-type and ΔB11 strains (Figure 7) with the non-complemented strains (Figure 4), we noted that the complemented and non-complemented ΔB11 strains both had small cells sizes. If complementation had been successful, the size of the complemented ΔB11 strains should be closer to the size of the complemented wild-type strain.

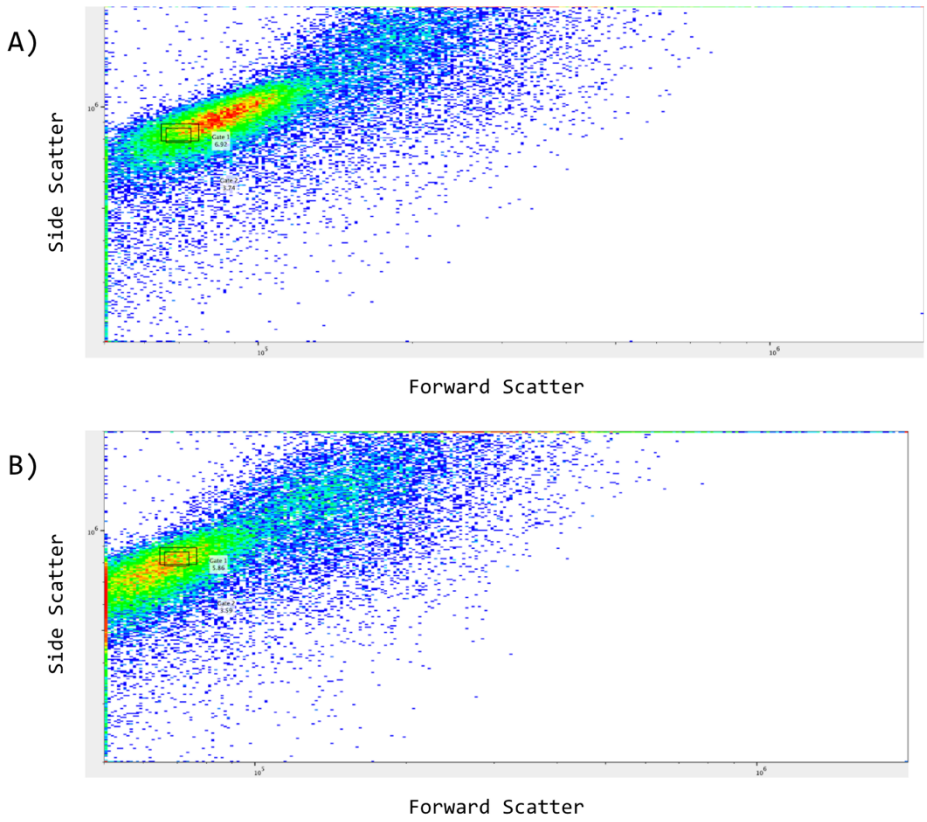


Figure 7: Cell size of the complemented strains determined from the flow cytometry data is comparable to the non-complemented strains. Figure 7 depicts flow cytometry data for cell size from both the complemented wild-type and ΔB11 strains. Forward scatter is on the x-axis and side scatter is on the y-axis. Panel A displays an example of the data used to determine the two gates in the complemented wild-type strain transformed with the B11 loop 1-complementary reporter plasmid. Panel B displays a similar example but with the complemented ΔB11 strain transformed with the B11 loop 1-complementary reporter plasmid.

To provide further evidence of that our flow cytometry results were caused by our L5 plasmid failing to express B11, we created a growth curve of all complemented and non-complemented wild-type and ΔB11 strains (Figure 8). The difference in growth rate between non-complemented WT and ΔB11 strains was comparable to that of complemented WT and ΔB11 strains, suggesting that there was no change in B11 expression in the ΔB11 strain after

attempted complementation. It should be noted that this analysis lacks data for the ΔB11 strain with a pJEB402 empty vector integrated in the L5 site. The presence of a plasmid in the L5 site can impact growth rate, which could mask any potential effects of B11 complementation in our complemented ΔB11 strain's fluorescence or cell size.

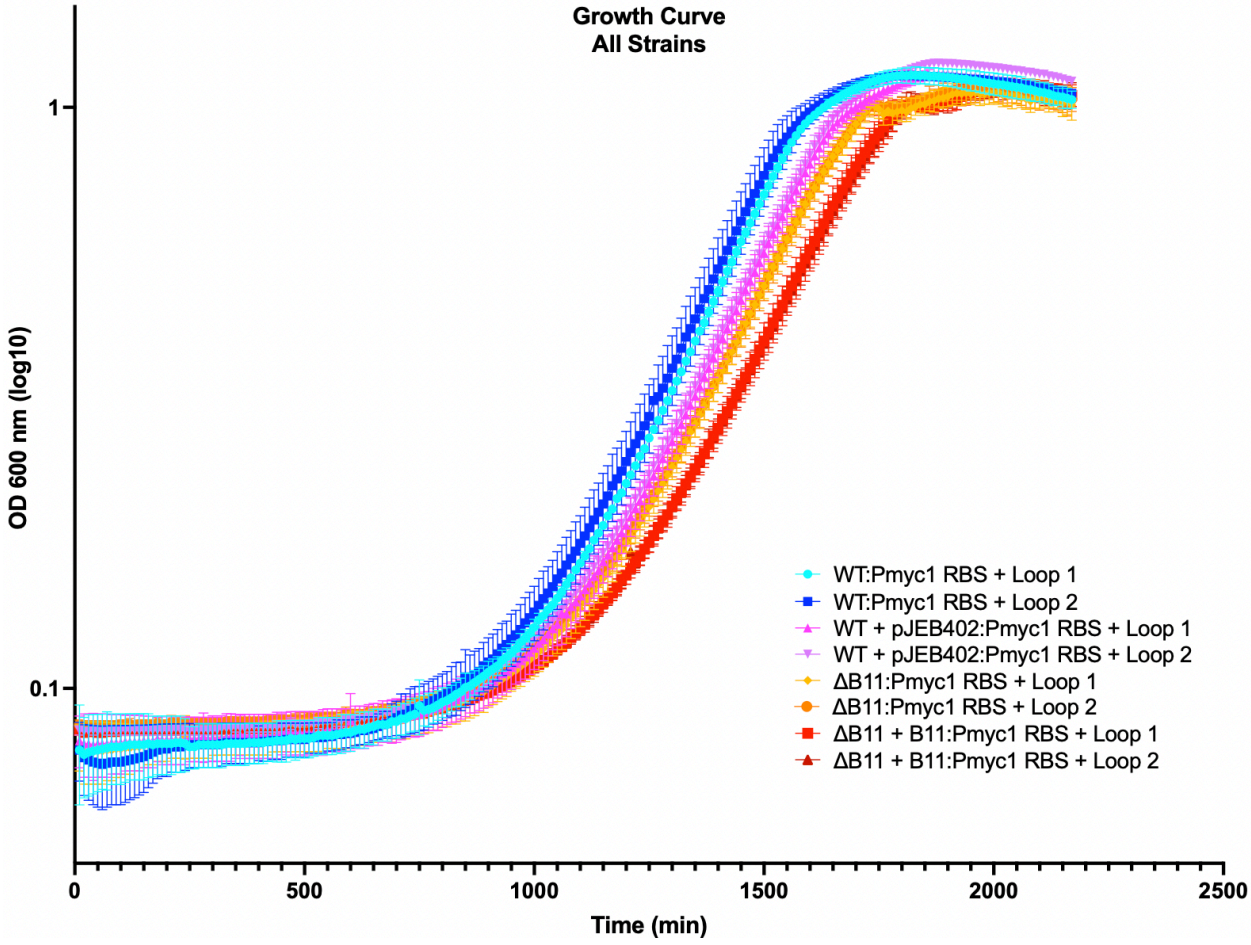


Figure 8: The L5 plasmid to complement the ΔB11 strain did not increase the strain's growth rate to match that of the WT as expected if complementation were successful. Figure 8 displays the growth curve of all complemented and non-complemented wild-type and ΔB11 strains. The time at which the OD was taken is on the x-axis in minutes, while the log of the OD taken at 600 nm is on the y-axis.

There may be higher G and A nucleotide content in the SDs of medium-fast degrading mRNAs

Efficient translation often protects mRNAs from degradation in bacteria (Caponigro & Parker, 1996; Mercante et al, 2009). Since strong SD sequences are typically G and A rich, we were therefore interested in the potential impact G and A nucleotide content in the SD region may have on mRNA stability. We examined the region of -17 to -4 upstream of the start codon in leadered *M. smegmatis* genes where we theorized the SD sequence may be and quantified both the percentage of this region made up of G and A nucleotides and a count of GA di-nucleotides in the region. We hypothesized that higher G and A nucleotide content by either quantification would correlate with higher mRNA stability. To investigate this relationship, we compared G and A nucleotide content between half-life classes and gene clusters. Other members of the lab have measured mRNA half-lives transcriptome-wide in log-phase *M. smegmatis*, divided genes into quartiles based on half-life, and designated each quartile as a class: fast, med-fast, med-slow, and slow. The fast class contains genes with the fastest mRNA decay rates, while the slow class contains genes with the slowest decay rates. We will refer to these as half-life classes. Separately, the mRNA degradation data were used to group genes by hierarchical clustering, resulting in four groups: fast-degrading, med-fast-degrading, med-slow-degrading, and slow-degrading. We will refer to these as degradation pattern clusters.

Kruskal-Wallis tests followed by Dunn's Multiple Comparison tests were used to identify any significant differences in GA nucleotide percentage or GA di-nucleotide count between half-life classes or degradation pattern clusters. There was no significant difference in GA nucleotide percentage for either degradation pattern clusters or half-life classes (Figure 9A-B). There was a significant difference in GA di-nucleotide counts among gene clusters (p-value = 0.0481) (Figure 9C), but not among half-life classes (Figure 9D). Despite the significant p-values from the

Kruskal-Wallis test of degradation pattern clusters and GA di-nucleotide counts, a Dunn's Multiple Comparison test could not determine between which clusters the significant relationship came from. However, it appears that the med-fast-degrading cluster had the highest count while the slow-degrading cluster had the lowest count. This result is not consistent with our hypothesis that greater G and A content in the SD region would promote mRNA stability.

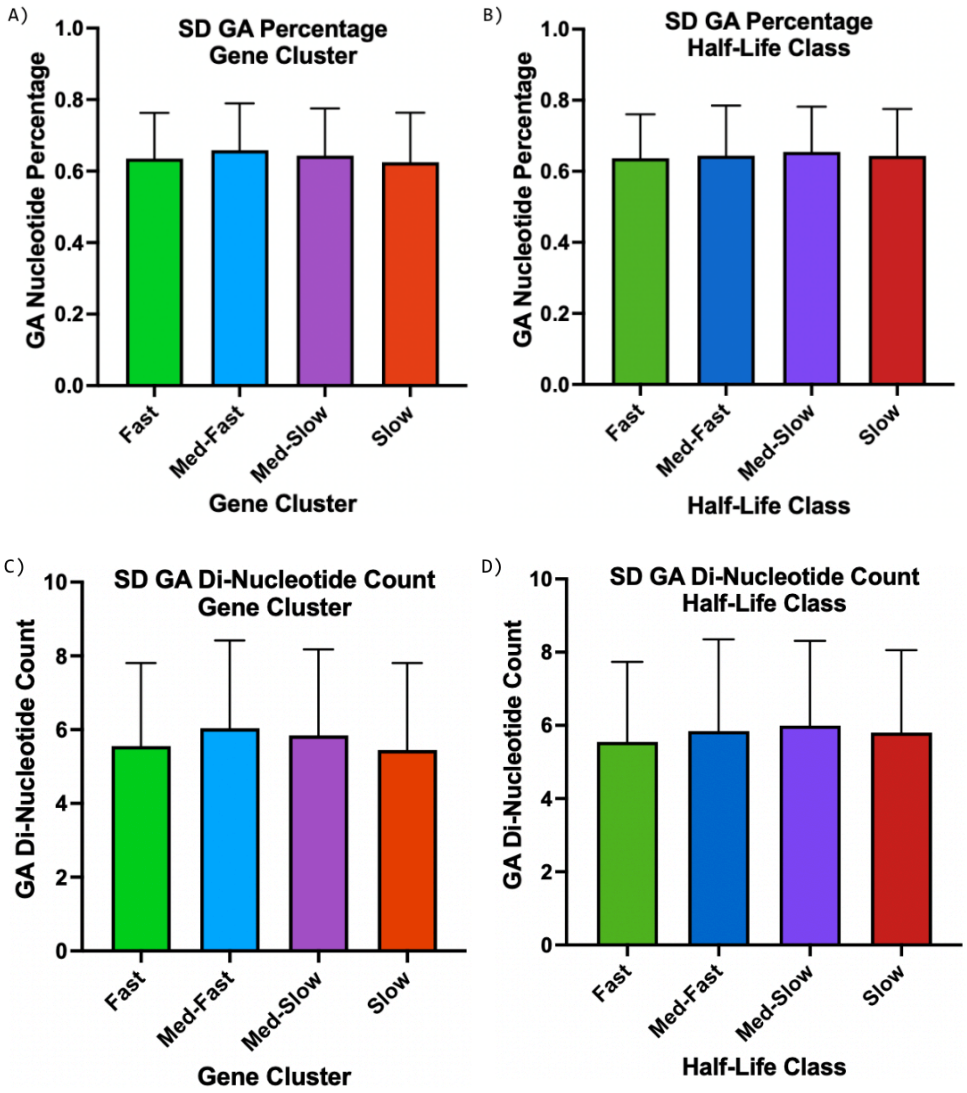


Figure 9: There may be a modest difference in GA di-nucleotide content among degradation pattern gene clusters. Figure 9 depicts G and A nucleotide content of degradation pattern gene clusters (fast-degrading, med-fast-degrading, med-slow-degrading, and slow-degrading) and half-life classes (fast,

med-fast, med-slow, and slow) in the region of -17 to -4 upstream of the start codon. Panels A and B contain data on GA nucleotide percentage for each degradation pattern gene cluster or half-life class respectively. Panels C and D contain data on GA di-nucleotide counts for degradation pattern gene clusters and half-life classes respectively.

There is a higher SD motif frequency in genes belonging to the med-fast half-life class or degradation pattern gene cluster.

We compared the presence of SD motifs among degradation pattern gene clusters and half-life classes. SD motif occurrence is a count of how often particular motifs appear within a region of 25 nucleotides upstream of the start codon of leadered *M. smegmatis* genes. For both questions, a Kruskal-Wallis test followed by a Dunn's Multiple Comparison test were run to identify a potential relationship between these features.

When comparing motif occurrence among gene clusters, the Kruskal-Wallis test showed that there were significant differences in the percentage of genes with the motifs AGGA (p-value = 0.0027), AGGAA (p-value = 0.0084), and GGAA (p-value = 0.0369) (Figure 10). For all three motifs, there was a higher frequency of motif occurrence in the med-fast-degrading and med-slow-degrading gene clusters.

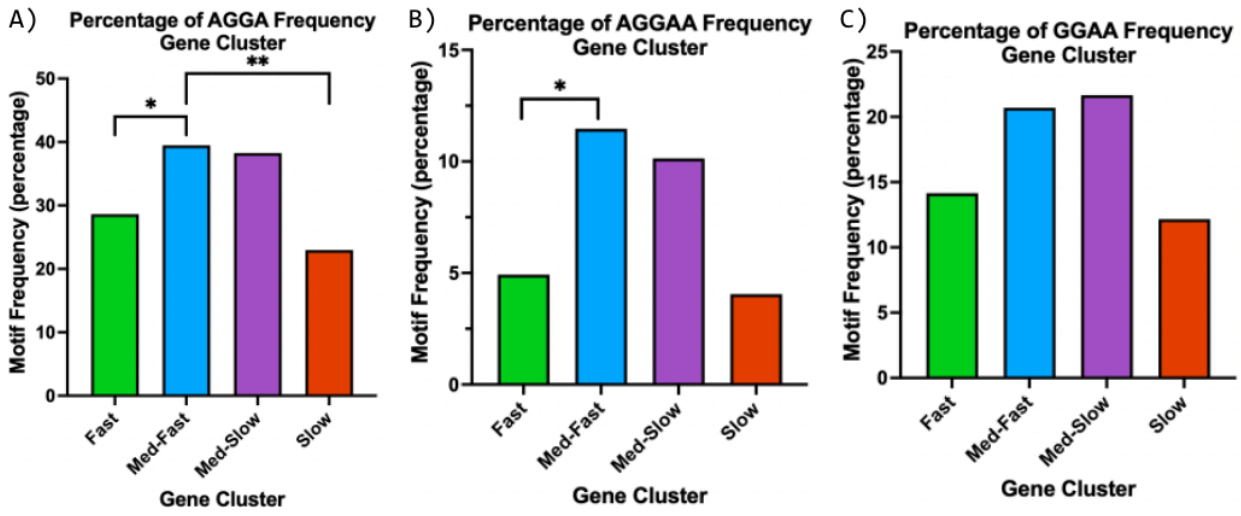


Figure 10: Med-fast-degrading and med-slow-degrading gene clusters have significantly higher frequency of SD motif occurrence. Figure 10 displays the percentage of genes within a cluster (fast-degrading, med-fast-degrading, med-slow-degrading, and slow-degrading) containing an SD motif in the region of 25 nucleotides upstream the start codon. Panel A depicts the data for the motif AGGA. The Dunn's Multiple Comparison test found a significant difference between the fast and med-fast clusters (p-value = 0.0225) and the med-fast and slow clusters (p-value = 0.0422). Panel B depicts the data for the motif AGGAA. The Dunn's Multiple Comparison test found a significant difference between the fast and med-fast clusters (p-value = 0.0163). Panel C depicts the data for the motif GGAA. While the Kruskal-Wallis test had a significant p-value, the Dunn's Multiple Comparison test did not identify significant differences between clusters.

When comparing the occurrence of SD motifs among half-life class, the Kruskal-Wallis test showed that there were significant differences in motif occurrence for AGGAA (p-value = 0.0216) and GGAA (p-value = 0.0091) (Figure 11). There was a significantly higher frequency of motif occurrence within the med-fast class.

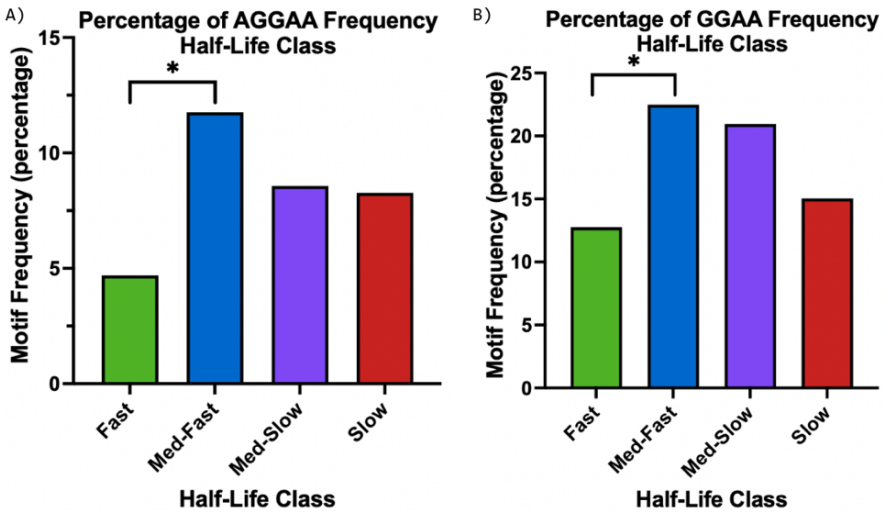


Figure 11: Med-fast half-life class has significantly higher frequency of SD motif occurrence. Figure 11 displays the percentage of genes within a half-life class (fast, med-fast, med-slow, and slow) containing each SD Motif. Panel A depicts the data for the motif AGGAA. The Dunn's Multiple Comparison test found a significant difference between the fast and med-fast clusters (p-value = 0.0114). Panel B depicts the data for the motif GGAA. The Dunn's Multiple Comparison test found a significant difference between the fast and med-fast clusters (p-value = 0.0128).

The presence of any SD motif is associated with higher ribosome occupancy, although no motifs are significantly better than others

To determine the impact that the presence of particular motifs may have on translation efficiency, we investigated the relationship between presence of SD motif sequences and ribosome occupancy as a measure of translation efficiency. The ribosome occupancy data used in our analysis was compiled by other members of the lab and originally published by Chen *et al* (Chen et al, 2020). We focused specifically on the ribosome occupancy of the coding sequence (CDS) excluding 18 nucleotides at the 5' end. This measure of ribosome occupancy was chosen given that it decreases the chance that the ribosomes bound to the gene that we are observing are stalled at the SD or start codon and not actually representative of active translation initiation.

We ran a Kruskal-Wallis test to determine if there was a significant difference in ribosome occupancy based on the SD motif, followed by a Dunn's Multiple Comparison test to determine which motifs were significantly different from each other. We found that while some motifs tended to have a higher ribosome occupancy than others, no motifs were significantly different than another (Figure 12). However, there were multiple motifs that had ribosome occupancy levels significantly higher than genes containing no motif.

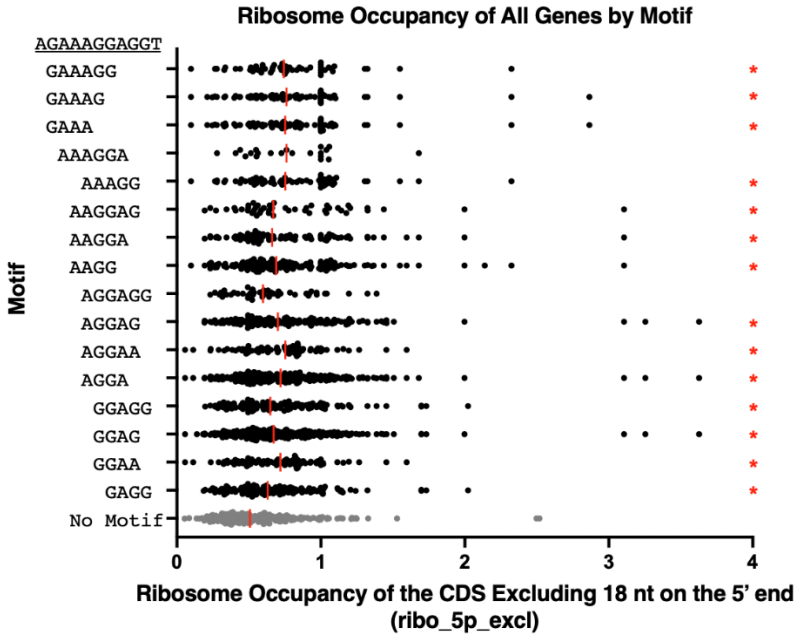


Figure 12: SD motifs cause significantly higher ribosome occupancy, but none are significantly better than another. Figure 12 displays the ribosome occupancy of genes with various SD motifs. The red lines denote the median ribosome occupancy for the genes containing the motif. The red stars denote that genes containing the motif have significantly greater median ribosome occupancy compared to genes containing no motif.

Controlling for a lower level of secondary structure around the RBS did not reveal any particular SD motifs associated with higher ribosome occupancy

We sought to repeat our analysis of the relationship between SD motif and ribosome occupancy without the potentially confounding variable of secondary structure, which may mask the relationship between motif presence and ribosome occupancy. To achieve this, we used data only from genes predicted to have low levels of secondary structure in the SD region. Other members of the lab have compiled data pertaining to the predicted secondary structure of *M. smegmatis* mRNA. While there are many different secondary structure measurements, we focused on secondary structure that directly impacted the region around the SD as that may influence the SD's accessibility to ribosomes. A higher level of secondary structure in the region around the SD may increase the likelihood of the ribosome being blocked from binding to the area, thereby lowering translation efficiency.

In our analysis we used data from two such secondary structure features. "Start Codon + RBS" (fpr_UTR_last30_start_codon_2nt_unpaired_prob_SD) is the probability of nucleotides -6 to -14 upstream of the start codon being unpaired when computationally predicting secondary structure by folding the last 30 nucleotides of the 5' UTR plus the start codon (Kelly, 2021). "CDS 5' 20 Nucleotides + RBS" (fpr_UTR_last30_20CDS_2nt_unpaired_prob_SD) is the probability of nucleotides -6 to -14 upstream of the start codon being unpaired when computationally predicting secondary structure by folding the last 30 nucleotides of the 5' UTR plus the first 20 nucleotides of 5' end of the CDS (Kelly, 2021) (Figure 13). If an SD motif is present, it is most likely to occur in the -6 to -14 region. Both features were measured on a scale of 0 to 1, with 0 being a 0% probability that the region -6 to -14 upstream of the start codon is unpaired (more secondary structure) and 1 being a 100% probability that the region -6 to -14 upstream of the start codon is unpaired (less secondary structure).

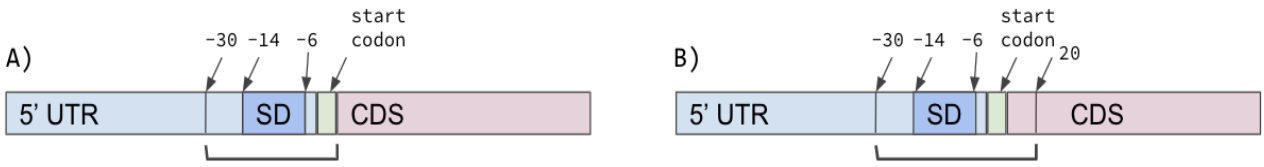


Figure 13: RBS regions that were computationally folded (Kelly, 2021) to predict secondary structure that would block SD accessibility. Brackets show the region of each transcript that was computationally folded. The probability of nucleotides -6 to -14 being unpaired within these structures was then determined. Panel A depicts Start Codon + RBS. Panel B depicts CDS 5' 20 Nucleotides + RBS.

When determining the relationship between specific motif occurrence and ribosome occupancy, we controlled for the level of secondary structure. For both features of interest, Start Codon + RBS and CDS 5' 20 Nucleotides + RBS, genes were divided into quartiles based on the probability of their SD region being unpaired. Quartile 1 contained genes with the lowest probability of being unpaired (more secondary structure), while Quartile 4 contained genes with the highest probability of being unpaired (less secondary structure) (Table 1). There were substantial numbers of genes with lower levels of secondary structure measured by both features (Figure 14 and 15), confirming that we would be able to control for secondary structure in further analysis by testing only these genes.

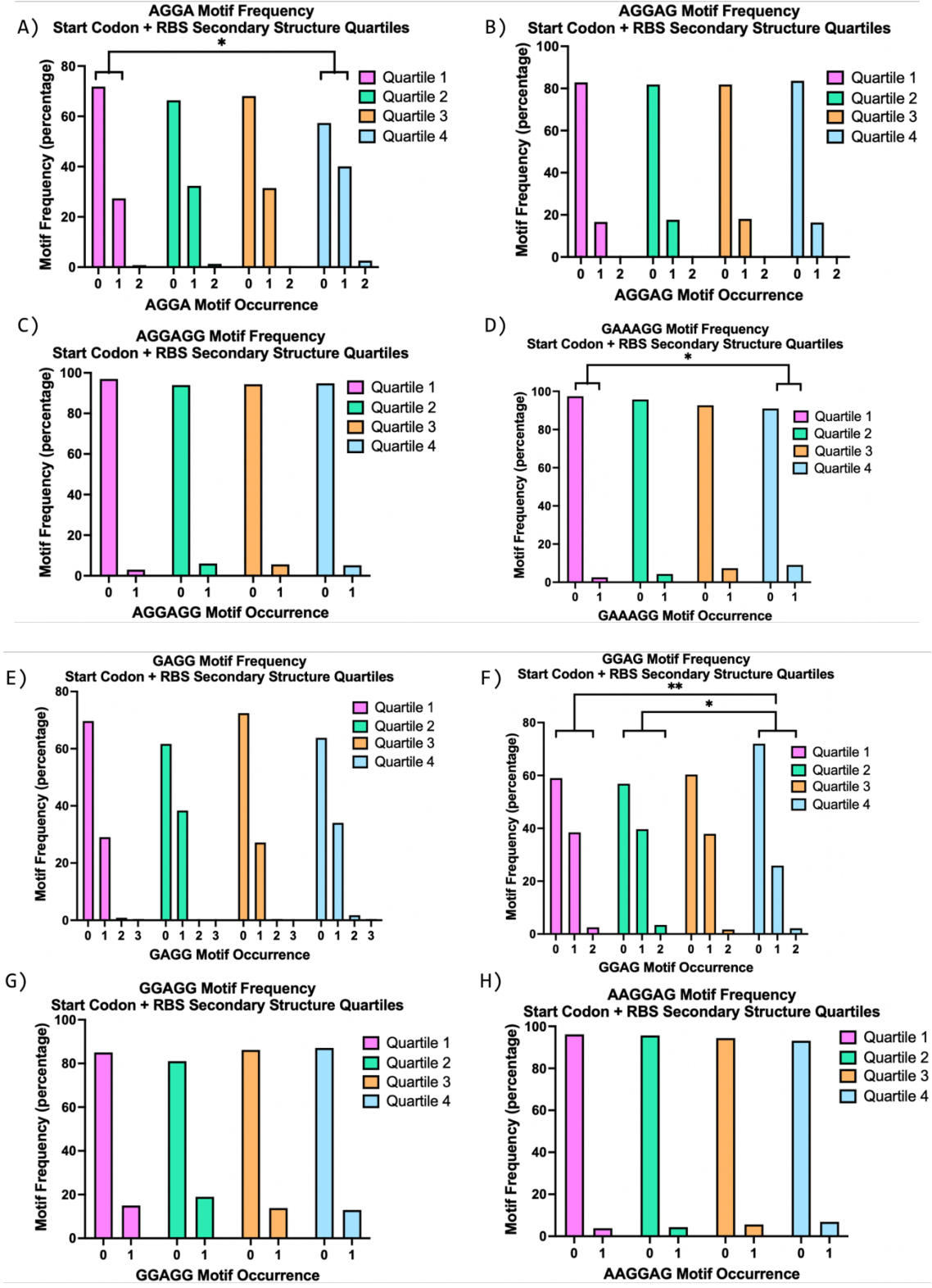
Table 1: Secondary Structure Feature Quartile Groups

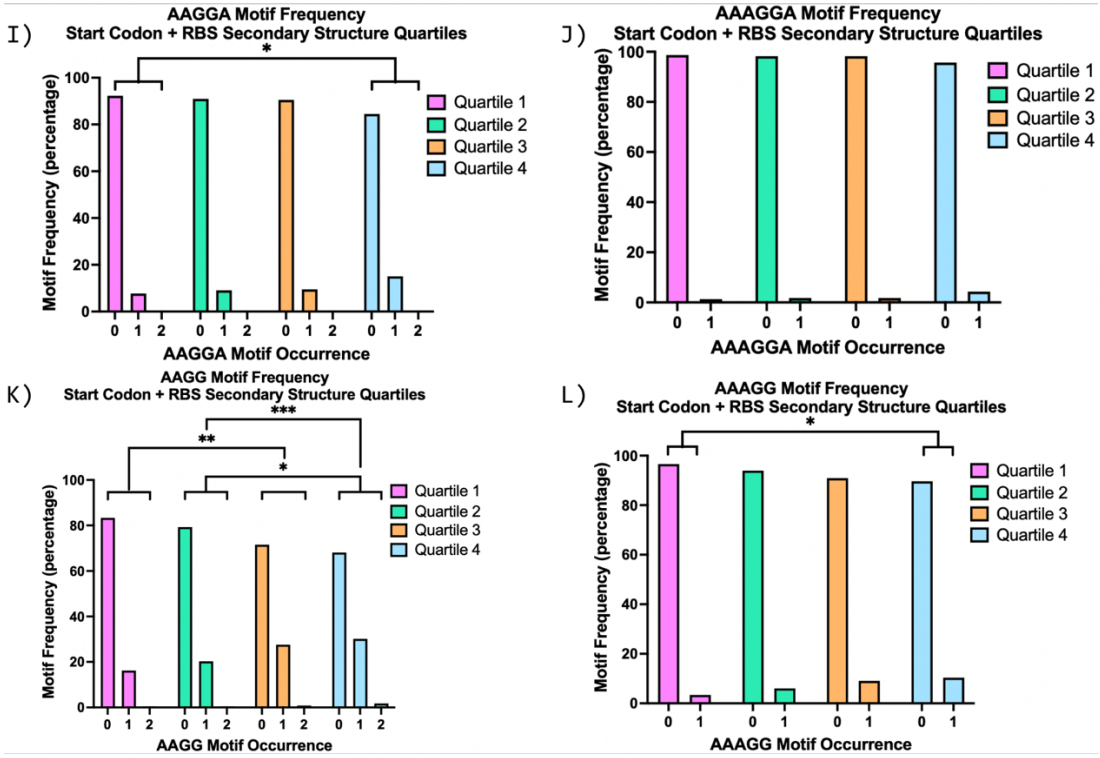
Feature	Quartile	Unpaired Probability Range
Start Codon + RBS	1	0.0020-0.3210
	2	0.3210-0.4300
	3	0.4300-0.5553

	4	0.5553-0.9950
CDS 5’ 20 Nucleotides + RBS	1	0.0000-0.2320
	2	0.2320-0.3555
	3	0.3555-0.5008
	4	0.5008-0.9840

Table 1 lists the quartile groups of Start Codon + RBS and CDS 5’ 20 Nucleotides + RBS.

We first ran Kruskal-Wallis tests followed by Dunn’s Multiple Comparison tests to determine if there were significant differences in motif occurrence based on the level of estimated secondary structure for these two features. We determined that there were significant differences in motif occurrence among the secondary structure quartiles. Motifs that had a higher A nucleotide content appeared more likely to be unpaired, as we see with the motif GAAA for example (Figure 14N and 15N). This means that A-rich motifs may have more genes with lower secondary structure levels in our analysis.


A)
AGGA Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
*
Motif Frequency (percentage)
AGGA Motif Occurrence
B)
AGGAG Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
AGGAG Motif Occurrence
C)
AGGAGG Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
AGGAGG Motif Occurrence
D)
GAAAGG Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
*
GAAAGG Motif Occurrence
E)
GAGG Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
GAGG Motif Occurrence
F)
GGAG Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
**
*
GGAG Motif Occurrence
G)
GGAGG Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
GGAGG Motif Occurrence
H)
AAGGAG Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
AAGGAG Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4


I)
AAGGA Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
AAGGA Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4
J)
AAAGGA Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
AAAGGA Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4
K)
AAGG Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
AAGG Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4
L)
AAAGG Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
AAAGG Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4

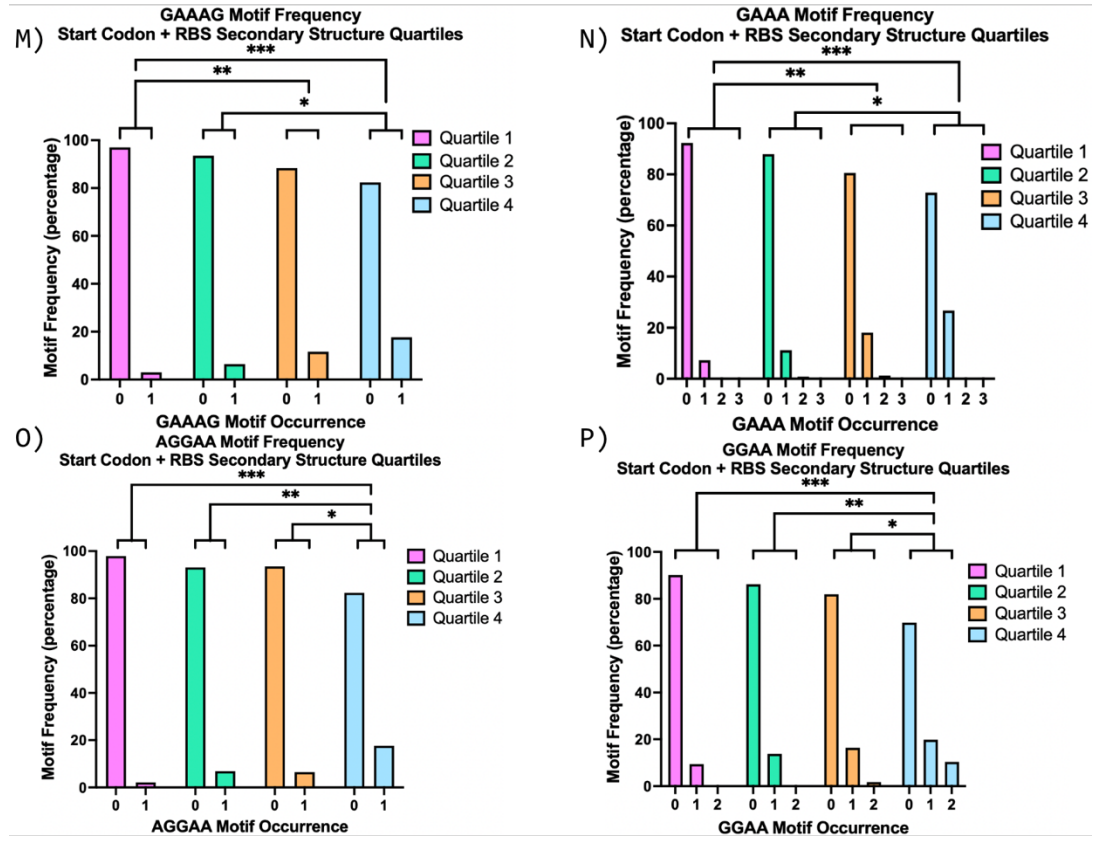
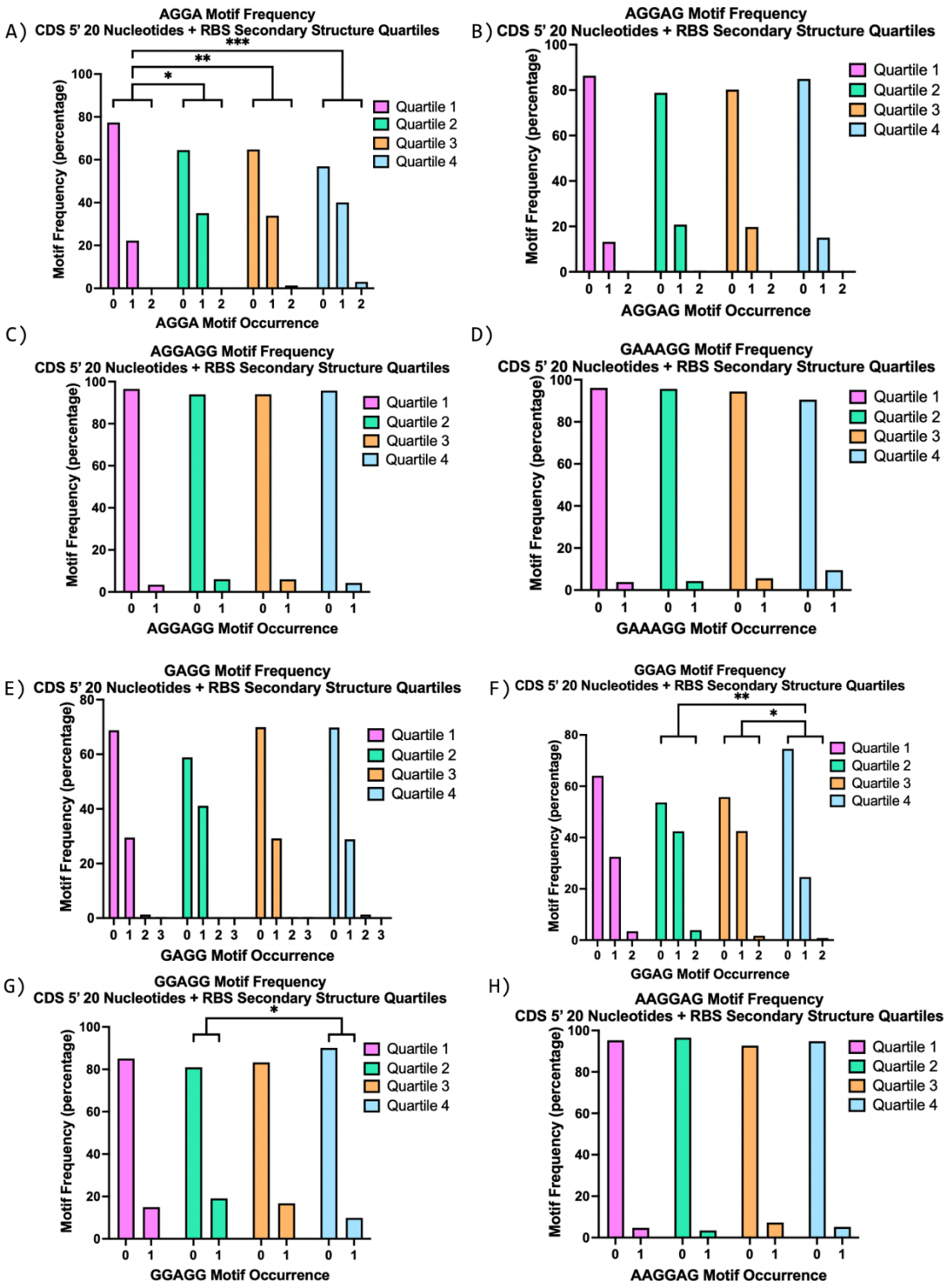
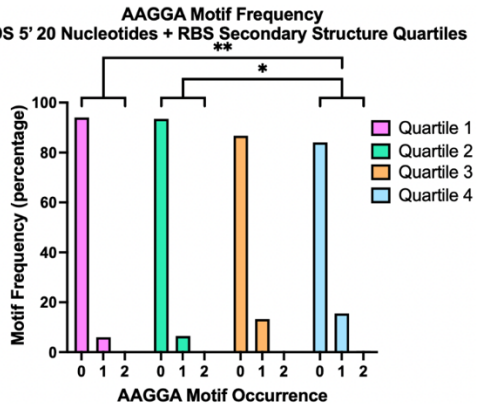

M)
GAAAG Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
GAAAG Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4
N)
GAAA Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
GAAA Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4
O)
AGGAA Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
AGGAA Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4
P)
GGAA Motif Frequency
Start Codon + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
GGAA Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4

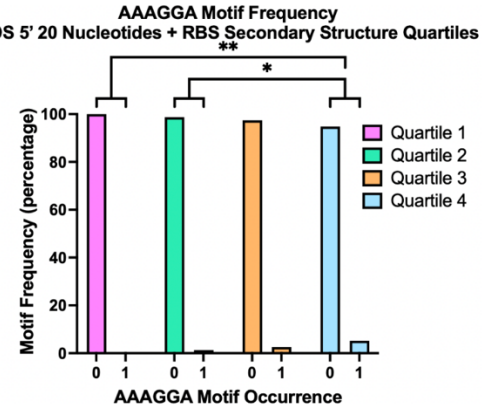
Figure 14: A-richness likely explains differences in SD motif frequency among Start Codon + RBS secondary structure quartiles. Motif occurrence for each Start Codon + RBS secondary structure quartile is shown. The motifs in the figure were all found to have significant differences in motif frequency among secondary structure quartiles through the Kruskal-Wallis test. Lines and stars indicate significant differences in the frequency of motif occurrence between different secondary structure quartiles in the Dunn's Multiple Comparisons Test.


A)
AGGA Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
AGGA Motif Occurrence
B)
AGGAG Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
AGGAG Motif Occurrence
C)
AGGAGG Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
AGGAGG Motif Occurrence
D)
GAAAGG Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
GAAAGG Motif Occurrence
E)
GAGG Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
GAGG Motif Occurrence
F)
GGAG Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
GGAG Motif Occurrence
G)
GGAGG Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
GGAGG Motif Occurrence
H)
AAGGAG Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
AAGGAG Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4

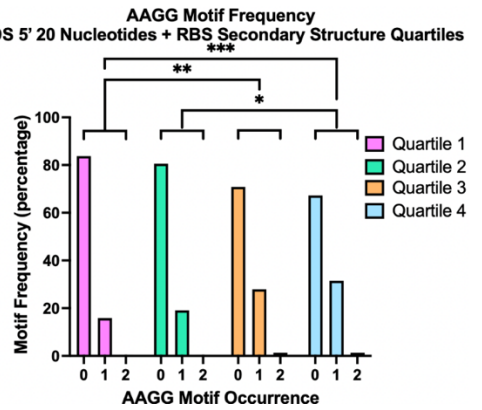
I)


AAGGA Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
AAGGA Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4

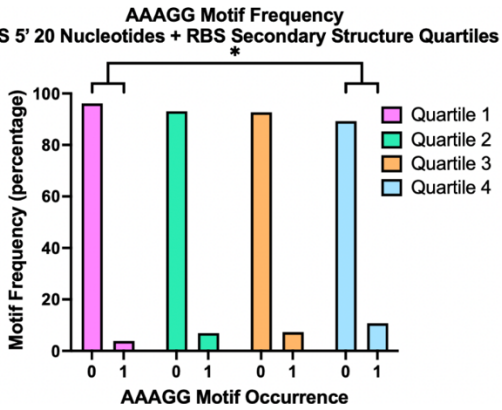
J)


AAAGGA Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
AAAGGA Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4

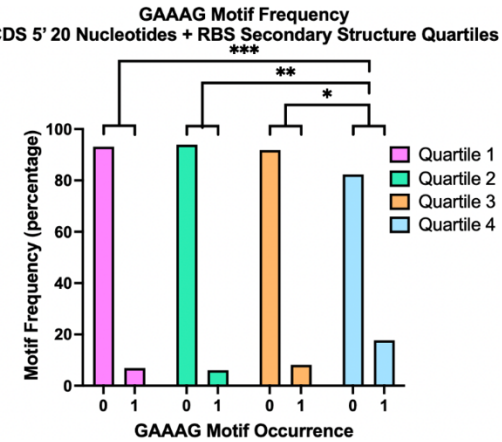
K)


AAGG Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
AAGG Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4

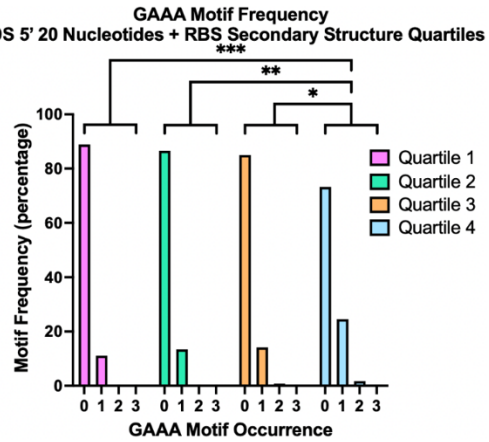
L)


AAAGG Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
AAAGG Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4

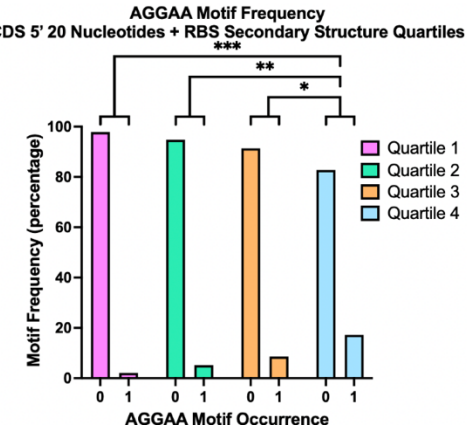
M)


GAAAG Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
GAAAG Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4

N)


GAAA Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
GAAA Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4

O)


AGGAA Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
AGGAA Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4

P)

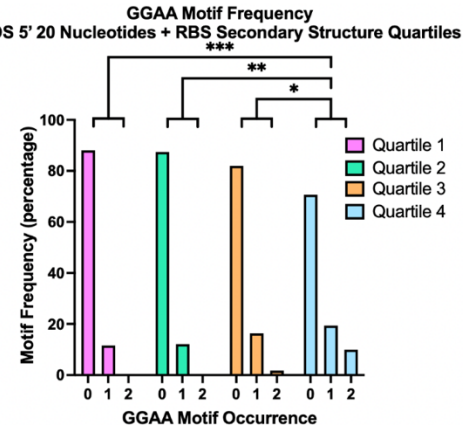

GGAA Motif Frequency
CDS 5' 20 Nucleotides + RBS Secondary Structure Quartiles
Motif Frequency (percentage)
GGAA Motif Occurrence
Quartile 1
Quartile 2
Quartile 3
Quartile 4

Figure 15: A-richness likely explains differences in SD motif frequency among CDS 5' 20 Nucleotides + RBS secondary structure quartiles. Motif occurrence for each CDS 5' 20 Nucleotides + RBS secondary structure quartile is shown. The motifs in the figure were all found to have significant differences in motif frequency among secondary structure quartiles through the Kruskal-Wallis test. Lines and stars indicate significant differences in the frequency of motif occurrence between different secondary structure quartiles in the Dunn's Multiple Comparisons Test.

To re-examine the relationship between motif occurrence and ribosome occupancy without the potential influence of secondary structure levels, we repeated the Kruskal-Wallis and Dunn's Multiple Comparison tests using only the genes that fell within Quartile 4 for each secondary structure feature. The results of this analysis paralleled those of the original tests in which secondary structure was not controlled for. Multiple motifs had ribosome occupancy levels that were significantly higher than genes containing no motif, but no particular motifs had significantly higher ribosome occupancies than the others (Figure 16).

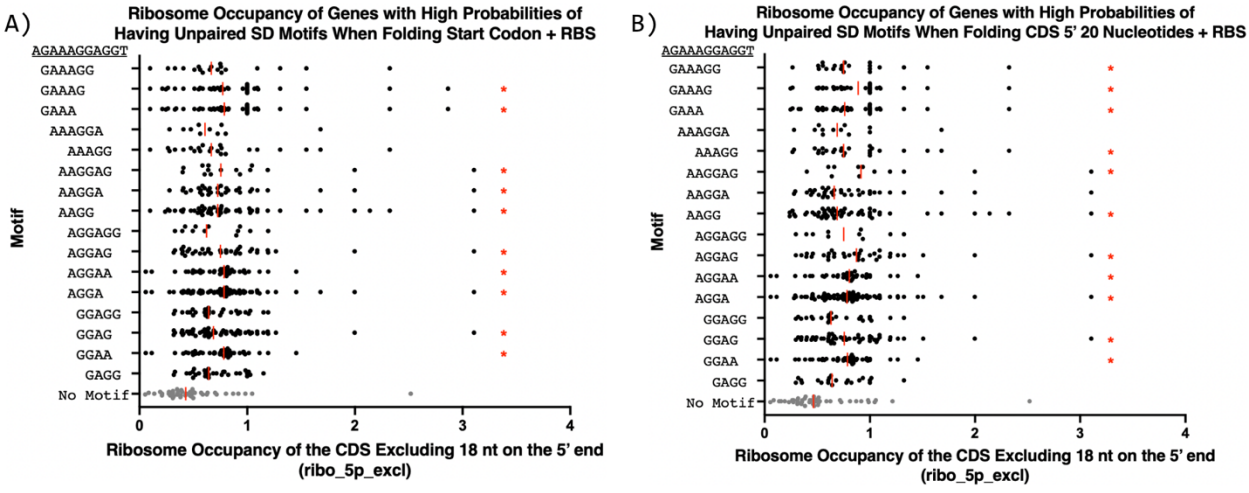


Figure 16: Even when controlling for lower secondary structure levels, no motifs were significantly better than others. Figure 16 displays the analysis of differences between ribo_5p_excl as a measure of ribosome occupancy based on the occurrence of specific motifs. The red lines denote the median

ribosome occupancy for the motif. The red stars denote that the motif's ribosome occupancy was significantly different than the ribosome occupancy of genes containing no motif. Panel A used gene data from quartile four of Start Codon + RBS. Panel B used gene data from quartile four of CDS 5' 20 Nucleotides + RBS.

Ribosome occupancy appears to be affected by SD motif location

We were interested in examining the impact of SD motif location on ribosome occupancy. We theorized that the anti-SD would bind to SD motifs in an optimal location based on where particular motifs fell in alignment with a theoretically perfect SD sequence. In this model, motifs located in the upstream portion of the theoretically perfect SD sequence should be found farther upstream from the start codon than motifs located in the downstream portion of the theoretically perfect SD sequence. To determine motif location based on this alignment metric, we calculated the number of nucleotides between a nucleotide aligned with the theoretically perfect SD sequence and the start codon (Figure 17).

Figure 17: Motif location was determined based on an alignment with the theoretically perfect SD sequence. Figure 17 displays the motifs of interest aligned with the theoretically perfect SD sequence. The nucleotide bolded in red was used to calculate motif location.

We focused on motifs that were associated with higher ribosome occupancy compared to genes containing no motif after controlling for both secondary structure features of interest (Figure 16). With motif locations determined through the metric shown in Figure 17, we used Kruskal-Wallis and Dunn's Multiple Comparison tests to determine significant differences in the location of particular motifs. While there were significant differences in location for specific motifs, we observed that the median locations of the motifs did not match our alignment predictions. For example, motifs we predicted to find farther upstream, such as GAAAG and GAAA, had common nucleotides that were downstream of the common nucleotides of motifs that aligned with the downstream portion of the theoretically perfect SD sequence (Figure 18).

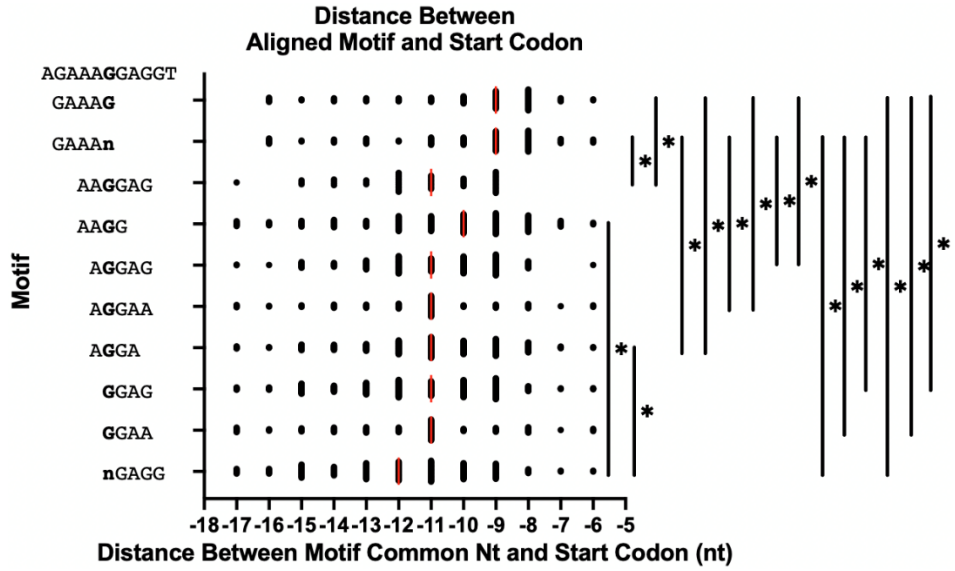


Figure 18: Location of some SD motifs are significantly different from each other, but do not match the locations predicted by alignment to a theoretically perfect SD motif. The distance between a common nucleotide within the theoretically perfect SD sequence and the start codon are shown. The red lines denote the median distance for all genes containing the motif. Lines to the right of the graph depict significant differences in the motif location by Dunn's post-test.

Given these results, we theorized that the optimal location for anti-SD binding was not dependent on motif alignment with the theoretically perfect SD sequence as previously thought. Perhaps rather than a specific optimal location for each motif, there is an optimal location in the RBS for the anti-SD to bind. Based on this idea, we repeated our analysis with a different metric of motif location obtained by calculating the number of nucleotides between a motif's 3' end nucleotide and the start codon (Figure 19).

Figure 19: Motif location determined based on each motif's 3' end nucleotide. Motifs of interest were aligned based on their 3' ends. The nucleotide bolded in red is the 3' end nucleotide used to calculate motif location.

Through another Kruskal-Wallis test, followed by a Dunn's Multiple Comparison test, we found there were still significant differences between the median locations of different motifs using this location metric (Figure 20). Motifs complementary to the 3' region of the anti-SD were located slightly upstream of motifs complementary to 5' region of the anti-SD. However, the differences were smaller than would be predicted if alignment with the anti-SD were the primary determinant of SD motif location. This is consistent with a model in which there is an optimal spacing between SD motifs and start codons that is somewhat affected by where within a theoretically perfect SD the motif lies.

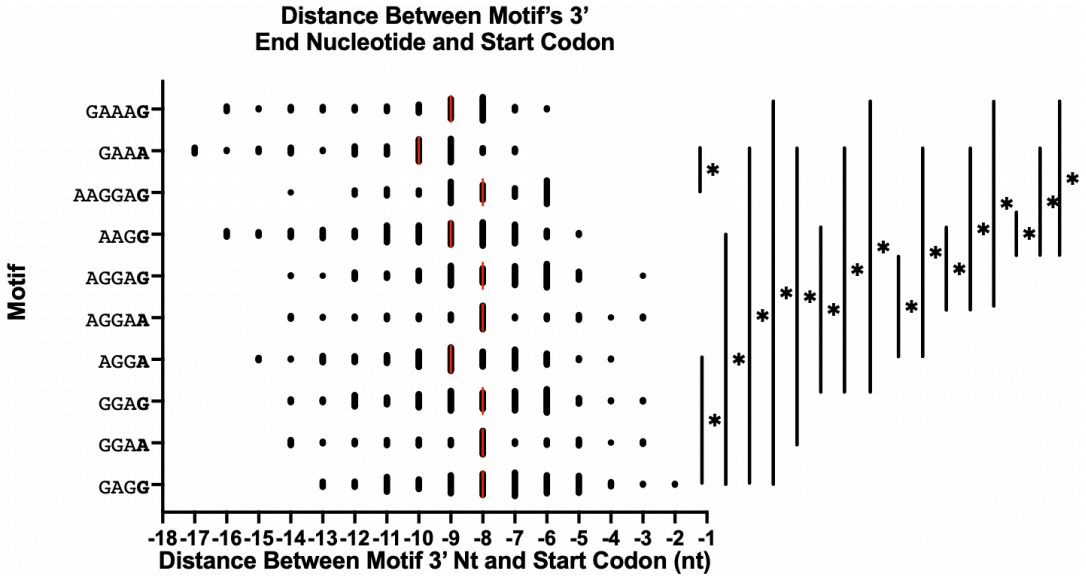


Figure 20: When motif location is determined by the absolute distance between the motif and the start codon, there is a modest upstream skew for motifs aligning further upstream in a theoretically perfect SD compared to motifs aligning further downstream. Distances between a motif's 3' end nucleotide and the start codon are shown. Red lines denote the median distance for each motif. Lines to the right of the graph depict significant differences in the motif location by Dunn's post-test.

When motif location is determined based on 3' end nucleotide and controlling for a lower level of secondary structure, ribosome occupancy is higher for motifs farther upstream

The range of motif locations for further analysis was narrowed based on the general trend in the peak containing the largest amount of genes for each location metric. When location was determined based on alignment to a theoretically perfect SD (Figure 17), the range of interest was narrowed to -13 to -8. When location was determined based on a motif's 3' end nucleotide (Figure 19), the range of interest was narrowed to -12 to -5.

With the narrowed ranges, we looked at the difference in ribosome occupancy based on motif location. For this analysis, all genes containing any of the previously chosen motifs were grouped together. We did the analyses both with and without controlling for secondary structure.

When using the theoretically perfect SD alignment as a metric of location and not controlling for secondary structure, there were no significant differences in ribosome occupancy among motif locations (Figure 21A). However, when using the same location metric and also using the Start Codon + RBS feature to control for secondary structure, the genes with SD motifs further upstream from the start codon had lower median ribosome occupancy than the genes with SD motifs at intermediates distances from the start codon (Figure 21B). When controlling for secondary structure with the CDS 5' 20 Nucleotides + RBS feature, there were no statistically significant differences between distance groups (Figure 21C).

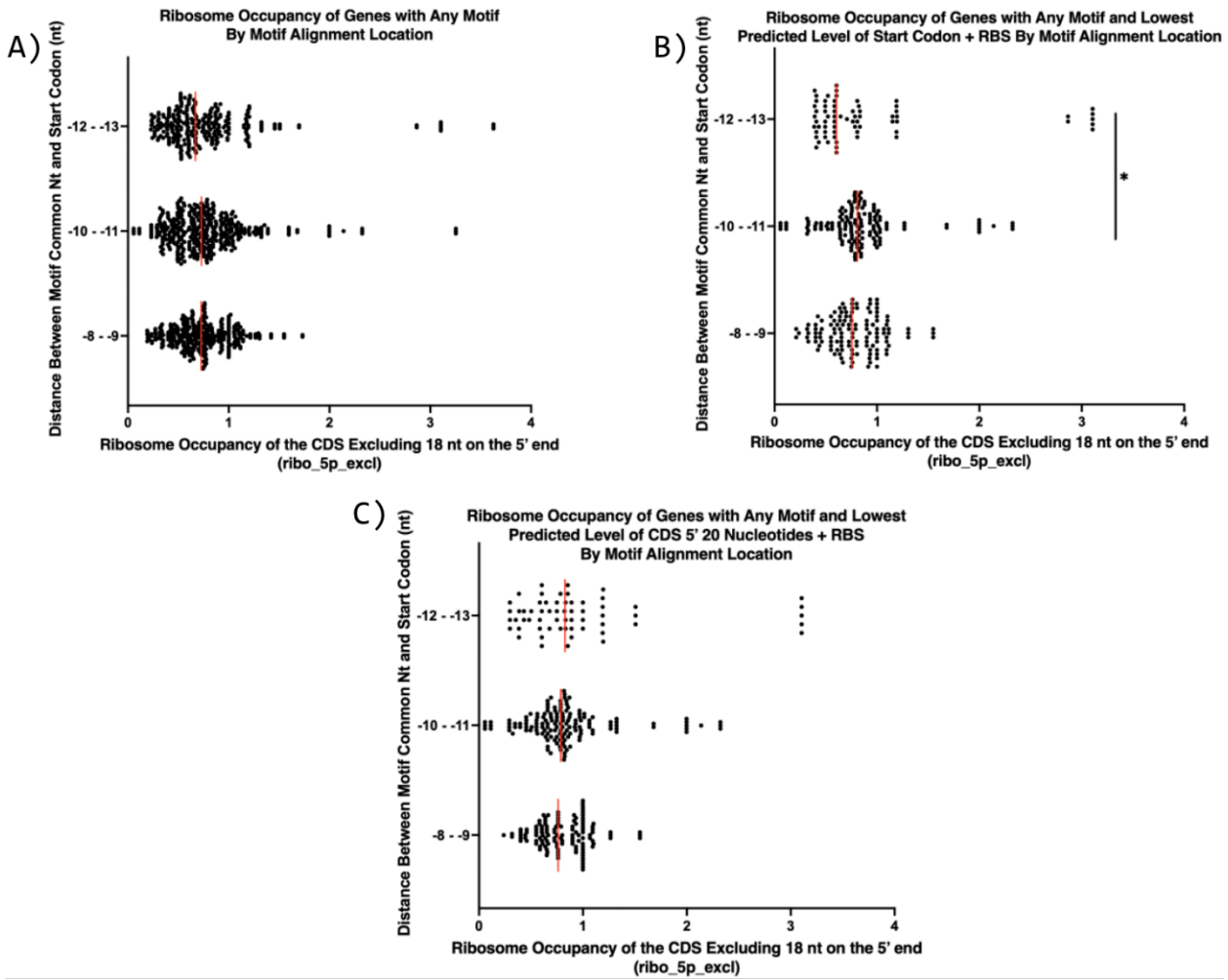


Figure 21: Differences in ribosome occupancy by motif location are modest when determining location based on alignment to a theoretically perfect SD. Figure 21 displays the analysis of ribosome occupancy based on motif location determined by the first location metric. Panel A displays the analysis without controlling secondary structure. Panel B displays the analysis using only genes that fell within Quartile 4 of Start Codon + RBS (lowest base-pairing probability for the SD region). Panel C displays the analysis using only genes that fell within Quartile 4 of CDS 5' 20 Nucleotides + RBS. The red lines denote the median ribosome occupancy for the distance group. Lines to the right of the graph depict significant differences in the ribosome occupancy by Dunn's post-test. The motifs included in this analysis were GAAAG, GAAA, AAGGAG, AAGG, AGGAG, AGGAA, AGGA, AGGA, GGAG, GGAA, and GAGG.

When determining motif location based on the motif's 3' end nucleotide, there was a significantly lower median ribosome occupancy for genes with SD motifs in the region closest to the start codon compared to the regions farther upstream when secondary structure was controlled for with either the Start Codon + RBS or CDS 5' 20 Nucleotides + RBS features (Figure 22 B and C). This difference was not seen when secondary structure was not controlled for (Figure 22A). Overall, the difference in ribosome occupancy based on motif location was more pronounced when determining location through this metric.

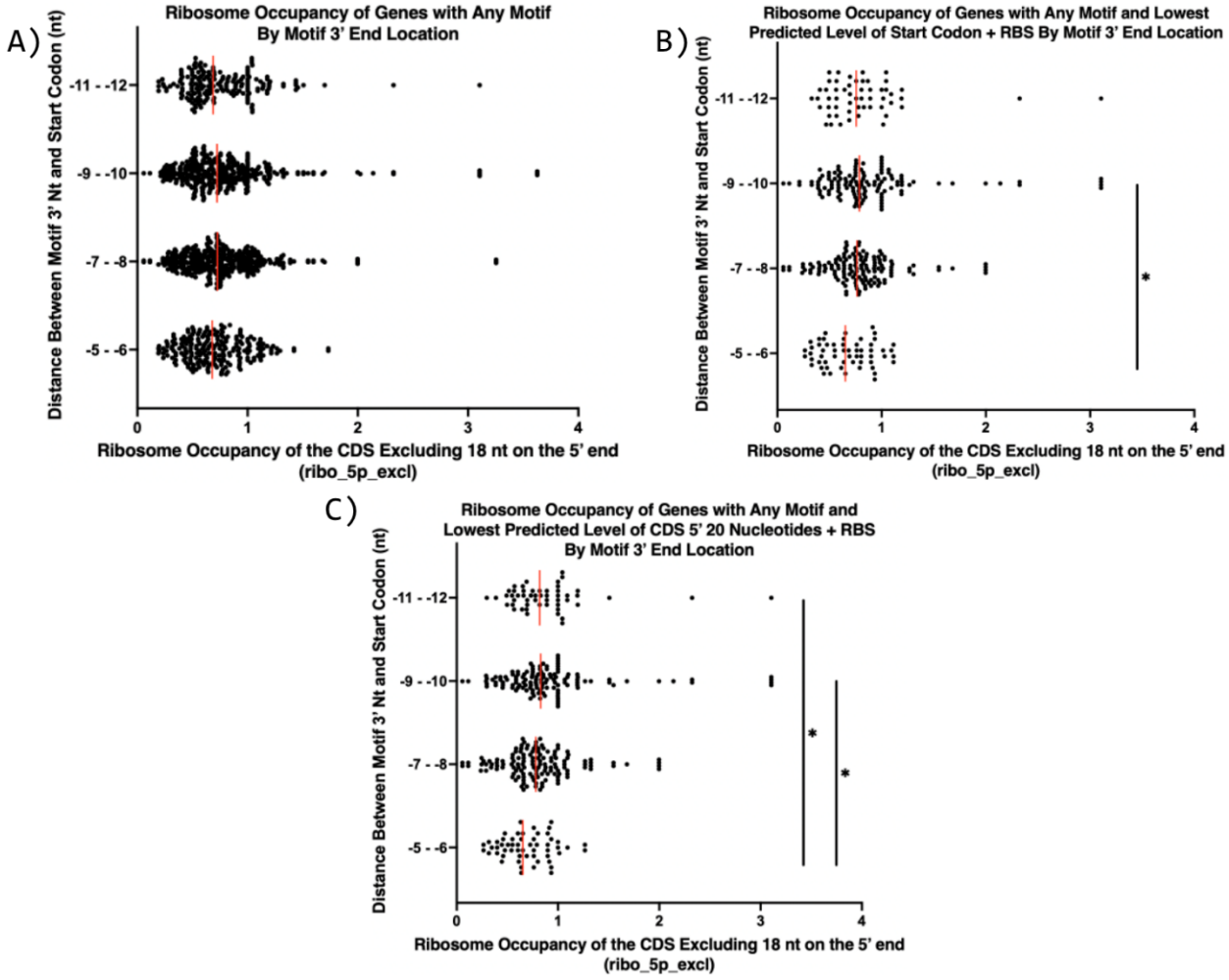


Figure 22: When controlling for secondary structure and motif location is determined by the 3' end nucleotide, motifs farther upstream have higher ribosome occupancy. Figure 22 displays the analysis

of ribosome occupancy based on motif location determined by the second location metric. Panel A displays the analysis without controlling secondary structure. Panel B displays the analysis using only genes that fall within Quartile 4 of the Start Codon + RBS secondary structure metric (highest predicted probability of the SD region being unpaired). Panel C displays the analysis using only genes that fall within Quartile 4 of the CDS 5' 20 Nucleotides + RBS secondary structure metric. The red lines denote the median ribosome occupancy for genes with motifs in each distance group. Lines to the right of the graph depict significant differences in the ribosome occupancy by Dunn's post-test. The motifs included in this analysis were GAAAG, GAAA, AAGGAG, AAGG, AGGAG, AGGAA, AGGA, AGGA, GGAG, GGAA, and GAGG.

This analysis was repeated for each motif of interest individually to see if the trend would also be seen for individual motifs. Given the low number of genes containing each motif, there was less statistical power behind this approach and it should be noted that may decrease the accuracy of the analysis. We performed this analysis without controlling for secondary structure as well to see potential trends that may be detectable only with a larger number of genes in the analysis. When not accounting for secondary structure levels, there was no significant difference in the level of ribosome occupancy based on motif location for the majority of the motifs. Even when some difference was observed, there were too few genes within the analysis to confidently draw a conclusion.

There was one motif, GGAG, for which we felt the number of genes was large enough to investigate further with more confidence. When not controlling for secondary structure levels, there was no significant difference in ribosome occupancy based on the motif's location. When either Start Codon + RBS or CDS 5' 20 Nucleotides + RBS were used to control for secondary structure, there did appear to be a difference between motif locations. While the trend was not

statistically significant, it appeared that motifs farther upstream generally had higher ribosome occupancy than those closer to the start codon. (Figure 23)

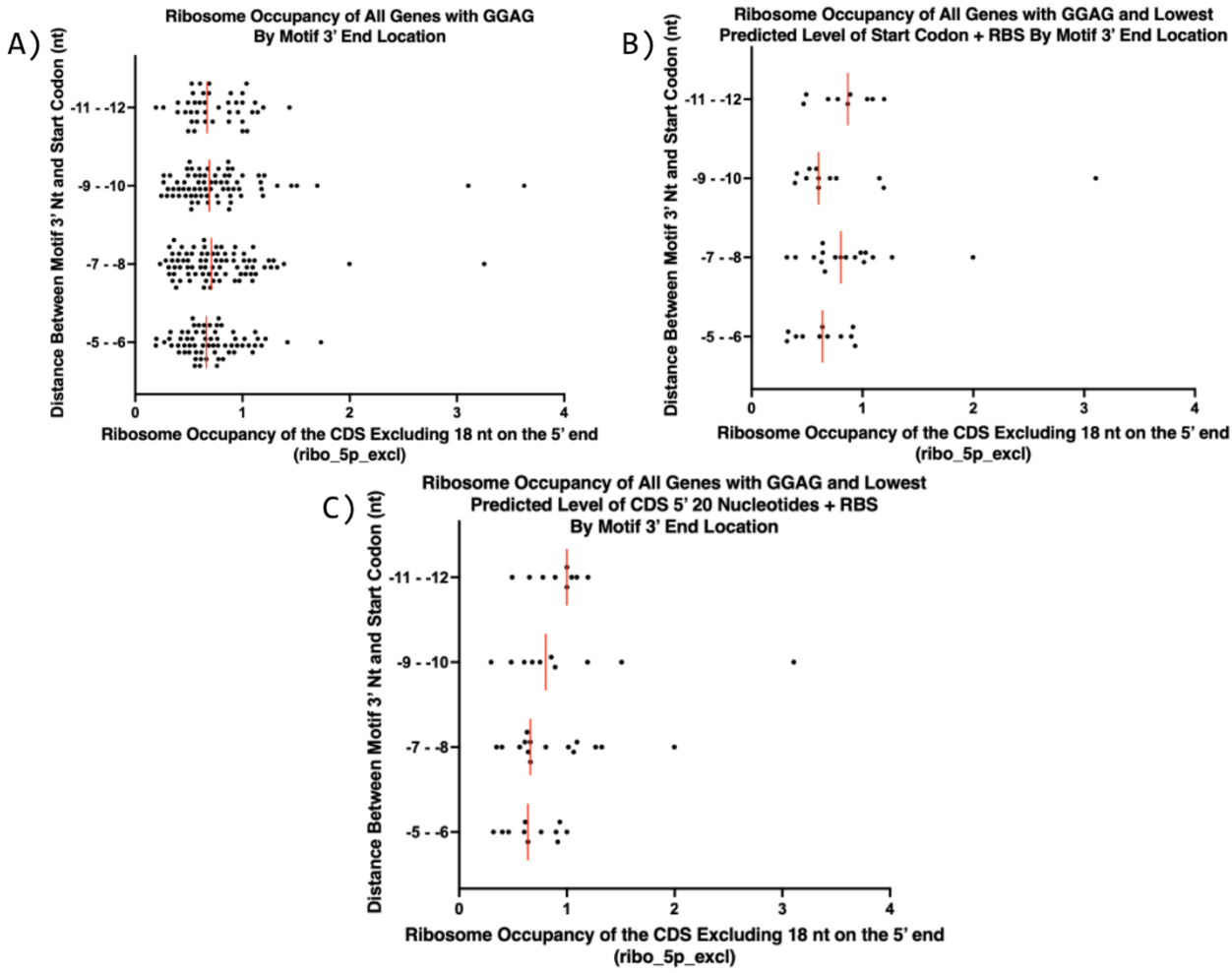


Figure 23: Though not statistically significant, the presence of GGAG farther upstream showed a trend towards higher ribosome occupancy. Figure 23 displays the analysis of ribosome occupancy based on motif location using the second location metric data for genes containing the GGAG motif. Panel A displays the analysis without controlling for secondary structure. Panel B displays the analysis using only genes that fell within Quartile 4 of the Start Codon + RBS secondary structure metric. Panel C displays the analysis using only genes that fell within Quartile 4 of the CDS 5' 20 Nucleotides + RBS secondary structure metric. The red lines denote the median ribosome occupancy for genes with SD motifs in each distance group.

Discussion

We sought to investigate the impact of RBS characteristics on translation efficiency in *M. smegmatis*. We theorized that the presence of sequences within the RBS that are complementary to the loops of the B11 sRNA may result in lower translation efficiency when B11 is present. Though we were unable to determine a direct relationship between B11 and changes in gene expression of our fluorescent reporter, we found that there was a greater decrease in fluorescence in *M. smegmatis* strains containing B11 and these complementary sequences compared to strains without B11. Additionally, we examined the influence of the SD sequence and it's position relative to the start codon. When controlling for the potential confounding characteristic of secondary structure, we discovered that the presence of any SD motif appeared to increase ribosome occupancy, though no one motif was significantly better than another. We also found that our initial determination of SD motif location for optimal anti-SD binding may not accurately represent the determinants of efficient ribosome binding. We originally determined motif location through an alignment of SD motifs with a theoretically perfect SD sequence, but we found that median SD motif locations were farther upstream or downstream than expected based on our alignment.

We found that replacing a portion of the Pmyc1 RBS with a B11-complementary sequence decreased expression of our YFP fluorescent reporter in both in the presence of B11 in the wild-type strain and the absence of B11 in the ΔB11 strain. If the binding of B11 to complementary sequences in the RBS influenced translation initiation, we expected that the strains in which B11 was absent would not show a decrease in YFP expression, determined through fluorescence level, after introduction of the complementary sequences. While we did

preserve the apparent SD sequence, it is possible that the B11-complementary sequences used in our constructs and the location they were inserted at may have decreased the efficiency of the Pmyc1 RBS. This could explain the decreased fluorescence seen in both strains.

We also observed that YFP expression was much higher in the ΔB11 strain than the wild-type strain when the B11-complementary sequence obtained from loop 2 was inserted in the Pmyc1 RBS. Despite the overall decrease in fluorescence of both strains, this difference between them suggests that B11 binds to the loop 2-complementary sequence when present in the wild-type strain, thereby blocking the ribosome and decreasing translation efficiency. However, when the loop 1-complementary sequence was inserted in the Pmyc1 RBS there was only a slight difference in fluorescence between the wild-type and ΔB11 strains. Though the ΔB11 strain had a slightly higher fluorescence, we could not confidently conclude that there was a substantial difference in gene expression between the two strains. It is possible that B11 could not recognize and bind to this complementary sequence as efficiently as to the loop 2-complementary sequence, perhaps given its shorter stretch of G nucleotides. Overall gene expression for both strains containing the loop 1-complementary sequence was substantially lower than that of strains with the loop 2-complementary sequence. This may suggest that the loop 1-complementary sequence chosen for our construct may have heavily decreased the efficiency of the Pmyc1 RBS compared to the loop 2 sequence, potentially masking the effects of B11 presence on YFP expression.

Despite the loop 2-complementary sequence appearing to allow for more effective B11 binding, seen through the substantially greater YFP expression in the ΔB11 strain than the wild-type strain after integration of this construct, we cannot conclude which B11 loop is actually binding to the sequence. This complementary sequence was based on the loop 2 sequence, but both B11 loops have long stretches of C nucleotides that could bind to the G nucleotides present

in the complementary sequence. In order to determine if there is a true difference in the function of the loops, further experimentation could be done in which the loops are individually mutated and the impact on fluorescence assessed. If a greater decrease in fluorescence is seen when only one of the loops is able to bind to the mRNA, this may suggest that loop has a greater influence in B11's function as a translation regulator.

Previous research by Mai *et al* provided evidence that the binding of B11 to mRNA does decrease translation efficiency of some *M. smegmatis* genes, so our results may demonstrate a similar mechanism (Mai et al, 2019). Given that our attempt to complement the ΔB11 strain with *M. smegmatis* B11 was unsuccessful, we cannot definitively conclude that the binding of B11 to these complementary sequences is directly causing the impact on YFP expression. Further work is needed to construct a plasmid that will cause B11 expression in the ΔB11 strain, which will allow us to be more confident in our current results.

Additionally, the question of B11's role in gene regulation when complementary sequences are present in the RBS warrants further investigation. The *M. smegmatis* genome annotation could be searched for the presence of naturally occurring B11-complementary sequences in 5' UTRs, similar to an investigation previously done with the *M. abscessus* genome. Several *M. abscessus* genes containing B11-complementary sequences in the 5' UTR were associated with the ESX-4 secretion system, which is related to the pathogen's virulence (Bar-Oz, SUBMITTED). Genes containing B11-complementary sequences in *M. smegmatis* may also be associated with particular functional categories. Such an analysis would provide further evidence for B11's role in mycobacteria gene and virulence regulation.

In analyzing the relationship between SD motifs and translation efficiency, we determined that no single motif caused a significantly greater rate of translation than any other motif. Though there were trends in the level of ribosome occupancy between motifs, these results

were not related to the length of the motif, the motif's alignment to the theoretically perfect SD sequence, or otherwise statistically significant. Rather, the presence of any motif with four or more nucleotides of complementarity to the anti-SD seemed to correlate with greater translation efficiency compared to genes that lacked any such motif. The positive impact of SD motif presence on translation has been shown in previous studies in other organisms, so our findings further support the significance of SD motifs for effective gene expression (Saito et al, 2020; Weyens et al, 1988; Ma et al, 2020).

To provide more evidence to support our finding that no one motif is significantly better than another, a set of fluorescent reporters containing different motifs could be constructed. This set could contain a plasmid with no motif present in its RBS and plasmids with motifs found to have either low or high levels of ribosome occupancy in our analysis. Any differences in the levels of fluorescence between these plasmids would provide more evidence to show the importance of motif presence and the impact of specific motifs present in the RBS.

Our analysis of motif location revealed that there are significant differences in ribosome occupancy depending on the location of different SD motifs within 5' UTRs. This suggests that specific motifs may be more efficient when located at a certain distance from the start codon. This relationship was difficult to determine through bioinformatic analysis alone, as the small sample size of genes containing a specific motif at a specific location lowered the statistical power of the analysis. With this in mind, the relationship between motif location and translation efficiency could be further investigated with sets of plasmid constructs of the same motifs with different spacing from a fluorescent reporter's start codon or different motifs with the same spacing from the start codon. When pooling genes with different SD motifs and controlling for secondary structure, we found a general trend in which genes with motifs farther upstream from

the start codon had greater ribosome occupancy. For the one motif in which we had a high enough sample size to control for secondary structure, we found a very similar trend.

We also noted that for motifs identified in our analysis, the general trends in their location were not as we had originally expected to find. Originally, we expected that the alignment of motifs with the theoretically perfect SD sequence would play a key role in a ribosome's ability to bind to the SD motif. In this model of ribosome binding, a motif would need to be located at an optimal location relative to the theoretically perfect SD sequence in order to be most efficiently recognized and bound by the anti-SD (Figure 24A-B). However, when we looked at motif location based on this model we did not see this trend. Motifs aligned with the upstream portion of the theoretically perfect SD sequence were found farther downstream than expected while motifs aligned downstream in the theoretically perfect SD sequence were found farther upstream than expected. This leads us to believe that our assumptions regarding the motif alignment's relevance may be incorrect. Instead, the anti-SD may have more flexibility in where it binds the RBS. We propose a new model for the binding of the anti-SD to the RBS, in which there is an optimal location for SD motifs relative to the start codon that is largely independent of which part of the anti-SD it must base-pair with (Figure 24B). In this model, the anti-SD has more flexibility in its ability to base-pair with these motifs and is not constrained by the theoretically perfect SD sequence.

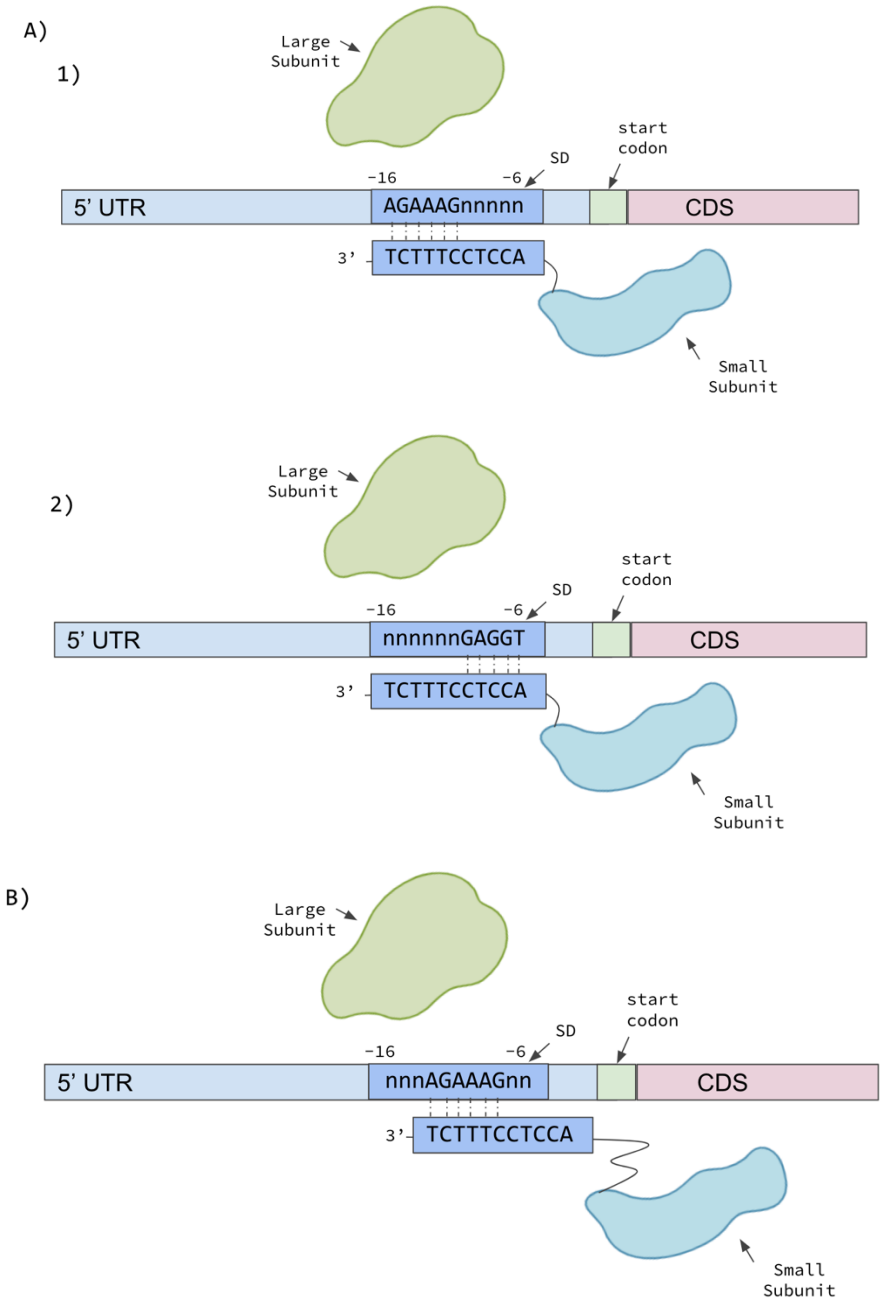


Figure 24: We propose a new model of anti-SD binding in which optimal SD motifs location is based primarily on distance from the start codon and only somewhat affect by which part of the anti-SD it base-pairs with. Two models of how the ribosome may recognize and bind to the RBS are shown. Panel A illustrates our original model in which SD motifs are aligned with a theoretically perfect SD sequence and optimal motif location depends on which part of the anti-SD is bound. A1 and A2 show two examples of optimally located SD motifs in this scenario, in which the anti-SD can bind to upstream

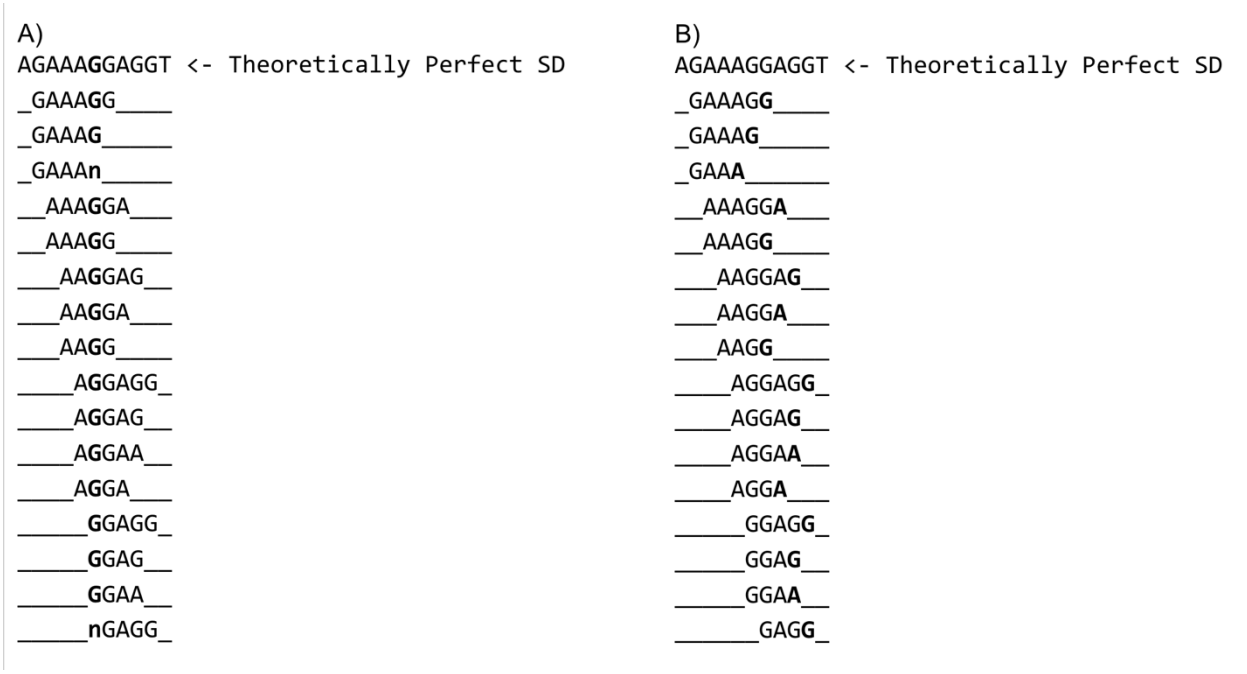
or downstream complementary SD sequences that specifically fall within the alignment. Part B illustrates the new model in which the optimal SD motif location is largely independent of which part of the anti-SD sequence base-pairs with the SD motif.

Our investigation provided insight on RBS characteristics that may increase the SD's efficiency in promoting translation, such as the presence of SD motifs, SD motif location, and lower levels of secondary structure. These characteristics may be beneficial in designing plasmids that maximize translation of a gene of interest in experimental studies using mycobacterial species. Our findings here can also be used as a basis for constructing more plasmids to further explore the impact of specific motif spacings and secondary structure levels on translation efficiency in future work.

Appendix

AGAAAGGA**G**GT	Theoretically Perfect SD
____AGGA**G**__	Motif 1
_____GGA**G**__	Motif 1a
_GAAAGGn**n**__	Motif 2
_GAAnGGn**n**__	Motif 2a
_GAAnGTn**n**__	Motif 2b
____AGGA**n**__	Motif 6
_____GAA**G**__	Motif 7
_____GGC**G**__	Motif 7a
_____GGT**G**__	Motif 7b
_____TCA**G**__	Motif 10a
_____TGA**G**__	Motif 10b

Appendix A: Supplemental Figure S1, 11 significant motifs were identified within the region of 20 nucleotides upstream of the start codon in leadered *M. smegmatis* genes. 11 motifs were identified in the 5' UTR region of 20 nucleotides upstream of the start codon in leadered *M. smegmatis* genes. These motifs were aligned with a theoretically perfect SD sequence. The bolded nucleotide was used to determine the location of these motifs relative to each other.



Appendix B: Supplemental Figure S2, Motifs of interest were aligned with a theoretically perfect SD sequence and their locations were determined through two metrics of a common nucleotide position. The figure lists all motifs of interest aligned with a theoretically perfect SD sequence. In Panel A, the bolded nucleotide signifies a commonly aligned nucleotide used to determine motif location relative to the motif alignment. In Panel B, the bolded nucleotide represents the motif's 3' end nucleotide used to determine motif location without consideration of the motif alignment.

References

1. Arnvig, K. B., & Young, D. B. (2009). Identification of small RNAs in Mycobacterium tuberculosis. *Molecular microbiology, 73*(3), 397–408. https://doi.org/10.1111/j.1365-2958.2009.06777.x
2. Baker, C. S., Eöry Lél A., Yakhnin, H., Mercante, J., Romeo, T., & Babitzke, P. (2020). CsrA inhibits translation initiation of Escherichia coli hfq by binding to a single site overlapping the shine-dalgarno sequence. *Journal of Bacteriology, 189*(15), 5472–5481. https://doi.org/10.1128/jb.00529-07
3. Baker, C. S., Morozov, I., Suzuki, K., Romeo, T., & Babitzke, P. (2002). CsrA regulates glycogen biosynthesis by preventing translation of glgC in Escherichia coli. *Molecular Microbiology, 44*(6), 1599–1610. https://doi.org/10.1046/j.1365-2958.2002.02982.x
4. Bar-Oz, M., Martini, M.C., Meir, M., Riva, C., Lore, N.I., Masiello, C.S., Sun, H., Moy, J.K., Cirillo, D.M., Shell, S.S., & Barkan, D. The small non-coding RNA B11 is a master regulator of Mycobacterium abscessus virulence. SUBMITTED.
5. Caponigro, G., & Parker, R. (1996). Mechanisms and control of mrna turnover in Saccharomyces cerevisiae. *Microbiological Reviews, 60*(1), 233–249. https://doi.org/10.1128/mr.60.1.233-249.1996
6. Casali, N., Nikolayevskyy, V., Balabanova, Y., Harris, S. R., Ignatyeva, O., Kontsevaya, I., Corander, J., Bryant, J., Parkhill, J., Negentsev, S., Horstmann, R. D., Brown, T., & Drobniewski, F. (2014) Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet 46*, 279–286. https://doi.org/10.1038/ng.2878
7. Center for Disease Control and Prevention. (2016). *Adverse Events*. Center for Disease Control and Prevention. Retrieved October 6, 2021, from https://www.cdc.gov/tb/topic/treatment/adverseevents.htm.
8. Center for Disease Control and Prevention. (2010). *Mycobacterium abscessus in Healthcare Settings*. Center for Disease Control and Prevention. Retrieved October 6, 2021, from https://www.cdc.gov/hai/organisms/mycobacterium.html.
9. Chen, Y. X., Xu, Z. Y., Ge, X., Sanyal, S., Lu, Z. J., & Javid, B. (2020). Selective translation by alternative bacterial ribosomes. *Proceedings of the National Academy of Sciences of the United States of America, 117*(32), 19487–19496. https://doi.org/10.1073/pnas.2009607117
10. de Smit, M. H., & van Duin, J. (1990). Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proceedings of the National Academy of Sciences of the United States of America, 87*(19), 7668–7672. https://doi.org/10.1073/pnas.87.19.7668
11. Del Campo, C., Bartholomäus, A., Fedyunin, I., & Ignatova, Z. (2015). Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function. *PLOS Genetics, 11*(10). https://doi.org/10.1371/journal.pgen.1005613
12. Heidrich, N., Moll, I., & Brantl, S. (2007). In vitro analysis of the interaction between the small RNA SR1 and its primary target ahrC mRNA. *Nucleic acids research, 35*(13), 4331–4346. https://doi.org/10.1093/nar/gkm439
13. Hockenberry, A. J., Pah, A. R., Jewett, M. C., & Amaral, L. A. (2017). Leveraging genome-wide datasets to quantify the functional role of the anti-shine–dalgarno sequence in regulating translation efficiency. Open Biology, 7(1), 160239. https://doi.org/10.1098/rsob.160239
14. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using

Ribosome Profiling. *Science, 324*(5924), 218-223. https://doi.org/10.1126/science.1168978
15. Iyer, S., Le, D., Park, B. R., & Kim, M. (2018). Distinct mechanisms coordinate transcription and translation under carbon and nitrogen starvation in escherichia coli. *Nature Microbiology, 3*(6), 741–748. https://doi.org/10.1038/s41564-018-0161-3
16. Jagodnik, J., Chiaruttini, C., & Guillier, M. (2017). Stem-Loop Structures within mRNA Coding Sequences Activate Translation Initiation and Mediate Control by Small Regulatory RNAs. *Molecular cell, 68*(1), 158–170.e3. https://doi.org/10.1016/j.molcel.2017.08.015
17. Kelly, J. (2021). Investigating the relationship between mRNA degradation rates and secondary structure in mycobacteria. (Undergraduate Major Qualifying Project No. E-project- 050321-104915). Retreived from Worcester Polytechnic Institute Electronic Projects Collection: https://digital.wpi.edu/concern/student_works/5138jh642?locale=en
18. Komarova, E. S., Chervontseva, Z. S., Osterman, I. A., Evfratov, S. A., Rubtsova, M. P., Zatsepin, T. S., Semashko, T. A., Kostryukova, E. S., Bogdanov, A. A., Gelfand, M. S., Dontsova, O. A., & Sergiev, P. V. (2020). Influence of the spacer region between the Shine-Dalgarno box and the start codon for fine-tuning of the translation efficiency in Escherichia coli. *Microbial biotechnology, 13*(4), 1254–1261. https://doi.org/10.1111/1751-7915.13561
19. Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. *Science, 324*(5924), 255–258. https://doi.org/10.1126/science.1170160
20. Laencina, L., Dubois, V., Le Moigne, V., Viljoen, A., Majlessi, L., Pritchard, J., Bernut, A., Piel, L., Roux, A. L., Gaillard, J. L., Lombard, B., Loew, D., Rubin, E. J., Brosch, R., Kremer, L., Herrmann, J. L., & Girard-Misguich, F. (2018). Identification of genes required for Mycobacterium abscessus growth in vivo with a prominent role of the ESX-4 locus. *Proceedings of the National Academy of Sciences of the United States of America, 115*(5), E1002–E1011. https://doi.org/10.1073/pnas.1713195115
21. Lee, M. H., Pascopella, L., Jacobs, W. R., Hatfull, G. F. (1991). Site-specific integration of mycobacteriophage L5: integration-proficient vectors for Mycobacterium smegmatis, Mycobacterium tuberculosis, and bacille Calmette-Guérin. *ProcNatl Acad Sci USA 88*(8), 3111–3115. https://doi.org/10.1073/pnas.88.8.3111
22. Liu, M. Y., Yang, H., & Romeo, T. (1995). The product of the pleiotropic Escherichia coli gene csrA modulates glycogen biosynthesis via effects on mRNA stability. *Journal of Bacteriology, 177*(10), 2663–2672. https://doi.org/10.1128/jb.177.10.2663-2672.1995
23. Ma, J., Campbell, A., & Karlin, S. (2002). Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *Journal of Bacteriology, 184*(20), 5733–5745. https://doi.org/10.1128/jb.184.20.5733-5745.2002
24. Mai, J., Rao, C., Watt, J., Sun, X., Lin, C., Zhang, L., & Liu, J. (2019). Mycobacterium tuberculosis 6C sRNA binds multiple mRNA targets via C-rich loops independent of RNA chaperones. *Nucleic acids research, 47*(8), 4292–4307. https://doi.org/10.1093/nar/gkz149
25. Man, O., & Pilpel, Y. (2007) Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat Genet, 39*, 415–421. https://doi.org/10.1038/ng1967

26. Mao, Y., Liu, H., Liu, Y., & Tao, S. (2014). Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in Saccharomyces cerevisiae. *Nucleic acids research, 42*(8), 4813–4822. https://doi.org/10.1093/nar/gku159
27. Mercante, J., Edwards, A. N., Dubey, A. K., Babitzke, P., & Romeo, T. (2009). Molecular geometry of CSRA (RSMA) binding to RNA and its implications for regulated expression. *Journal of Molecular Biology, 392*(2), 511–528. https://doi.org/10.1016/j.jmb.2009.07.034
28. Mvubu, N. E., Pillay, B., Gamieldien, J., Bishai, W., & Pillay, M. (2016). Canonical pathways, networks and transcriptional factor regulation by clinical strains of mycobacterium tuberculosis in pulmonary alveolar epithelial cells. *Tuberculosis, 97*, 73–85. https://doi.org/10.1016/j.tube.2015.12.002
29. Nakagawa, S., Niimura, Y., Miura, K., & Gojobori, T. (2010). Dynamic evolution of translation initiation mechanisms in prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America, 107*(14), 6382–6387. https://doi.org/10.1073/pnas.1002036107
30. Nguyen, T. G., Vargas-Blanco, D. A., Roberts, L. A., & Shell, S. S. (2020). The Impact of Leadered and Leaderless Gene Structures on Translation Efficiency, Transcript Stability, and Predicted Transcription Rates in Mycobacterium smegmatis. *Journal of bacteriology, 202*(9), e00746-19. https://doi.org/10.1128/JB.00746-19
31. Olson, P. D., Kuechenmeister, L. J., Anderson, K. L., Daily, S., Beenken, K. E., Roux, C. M., Reniere, M. L., Lewis, T. L., Weiss, W. J., Pulse, M., Nguyen, P., Simecka, J. W., Morrison, J. M., Sayood, K., Asojo, O. A., Smeltzer, M. S., Skaar, E. P., & Dunman, P. M. (2011). Small molecule inhibitors of Staphylococcus aureus RnpA alter cellular mRNA turnover, exhibit antimicrobial activity, and attenuate pathogenesis. *PLoS pathogens, 7*(2), e1001287. https://doi.org/10.1371/journal.ppat.1001287
32. Proshkin, S., Rahmouni, A. R., Mironov, A., & Nudler, E. (2010). Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science, 328*(5977), 504–508. https://doi.org/10.1126/science.1184939
33. Qvist, T., Taylor-Robinson, D., Waldmann, E., Olesen, H. V., Hansen, C. R., Mathiesen, I. H., Høiby, N., Katzenstein, T. L., Smyth, R. L., Diggle, P. J., & Pressler, T. (2016). Comparing the harmful effects of nontuberculous mycobacteria and Gram negative bacteria on lung function in patients with cystic fibrosis. *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society, 15*(3), 380–385. https://doi.org/10.1016/j.jcf.2015.09.007
34. Saito, K., Green, R., Buskirk, A. R. (2020). Translation initiation in E. coli occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing. *eLife, 9*, e55002. https://doi.org/10.7554/eLife.55002
35. Shell, S. S., Wang, J., Lapierre, P., Mir, M., Chase, M. R., Pyle, M. M., Gawande, R., Ahmad, R., Sarracino, D. A., Ioerger, T. R., Fortune, S. M., Derbyshire, K. M., Wade, J. T., & Gray, T. A. (2015). Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLoS genetics, 11*(11), e1005641. https://doi.org/10.1371/journal.pgen.1005641
36. Solans, L., Gonzalo-Asensio, J., Sala, C., Benjak, A., Uplekar, S., Rougemont, J., Guilhot, C., Malaga, W., Martín, C., & Cole, S. T. (2014). The PhoP-dependent ncRNA Mcr7 modulates the TAT secretion system in Mycobacterium tuberculosis. *PLoS pathogens, 10*(5), e1004183. https://doi.org/10.1371/journal.ppat.1004183
37. Streicher, E. M., Müller, B., Chihota, V., Mlambo, C., Tait, M., Pillay, M., Trollip, A., Hoek, K. G. P., Sirgel, F. A., Gey van Pittius, N. C., van Helden, P. D., Victor, T. C., &

Warren, R. M. (2012). Emergence and treatment of multidrug resistant (MDR) and extensively drug-resistant (XDR) tuberculosis in South Africa. Infection, *Genetics and Evolution, 12*(4), 686–694. https://doi.org/10.1016/j.meegid.2011.07.019

38. Studer, S. M., & Joseph, S. (2006). Unfolding of mrna secondary structure by the bacterial translation initiation complex. *Molecular Cell, 22*(1), 105–115. https://doi.org/10.1016/j.molcel.2006.02.014
39. Tholstrup, J., Oddershede, L. B., & Sørensen, M. A. (2012). mRNA pseudoknot structures can act as ribosomal roadblocks. *Nucleic acids research, 40*(1), 303–313. https://doi.org/10.1093/nar/gkr686
40. Thomason, M. K., Voichek, M., Dar, D., Addis, V., Fitzgerald, D., Gottesman, S., Sorek, R., & Greenberg, E. P. (2019). A rhlI 5′ UTR-derived sRNA regulates RhlR-dependent quorum sensing in Pseudomonas aeruginosa. *MBio, 10*(5). https://doi.org/10.1128/mbio.02253-19
41. Udekwu, K. I., Darfeuille, F., Vogel, J., Reimegård, J., Holmqvist, E., & Wagner, E. G. (2005). Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA. *Genes & development, 19*(19), 2355–2366. https://doi.org/10.1101/gad.354405
42. Vargas-Blanco, D. A., Zhou, Y., Zamalloa, L. G., Antonelli, T., & Shell, S. S. (2019). mRNA Degradation Rates Are Coupled to Metabolic Status in Mycobacterium smegmatis. *mBio, 10*(4), e00957-19. https://doi.org/10.1128/mBio.00957-19
43. Wang, X., Dubey, A. K., Suzuki, K., Baker, C. S., Babitzke, P., & Romeo, T. (2005). CsrA post-transcriptionally represses pgaABCD, responsible for synthesis of a biofilm polysaccharide adhesin of Escherichia coli. *Molecular Microbiology, 56*(6), 1648–1663. https://doi.org/10.1111/j.1365-2958.2005.04648.x
44. Weyens, G., Charlier, D., Roovers, M., Piérard, A., & Glansdorff, N. (1988). On the role of the shine-dalgarno sequence in determining the efficiency of translation initiation at a weak start codon in the car operon of escherichia coli K12. *Journal of Molecular Biology, 204*(4), 1045–1048. https://doi.org/10.1016/0022-2836(88)90061-7
45. World Health Organization. (2021). *Tuberculosis*. World Health Organization. Retrieved October 3, 2021, from https://www.who.int/health-topics/tuberculosis#tab=tab_1.