January 2017

# Role of the 5' UTR in Expression of an Essential Heat Shock Protein in M. smegmatis

David M. Morgan
*Worcester Polytechnic Institute*

# Role of the 5' UTR in Expression of an Essential Heat Shock Protein in *M. smegmatis*

A Major Qualifying Project

Submitted to the Faculty of

Worcester Polytechnic Institute

In partial fulfillment of the requirements for the

Degree in Bachelor of Science

In

Biology and Biotechnology

By

David Morgan

Date: 25JAN16

Project Advisor:

Dr. Scarlet Shell

**Acknowledgements**

The author would like to acknowledge the assistance provided by the members of the Shell Lab, notably Paula Bertuso, Ying Zhou, and Diego Vargas. Their advice and aid was indispensable, as was the welcoming and inclusive environment they created in the lab.

**Abstract**

Tuberculosis (TB) is a major worldwide health concern, infecting an estimated one third of the global population and responsible for greater than one million deaths each year. Populations of the causative agent, *Mycobacterium tuberculosis*, are genetically identical but phenotypically heterogeneous within an infected individual. This phenotypic variation can confer antibiotic tolerance or otherwise make the infection difficult to treat. We hypothesize that one source of phenotypic heterogeneity is the 5' untranslated regions (5' UTRs), which can modulate both transcript stability and translation efficiency. The relatively long length of the 5' UTR of *groES* suggests that its functionality may extend beyond ribosome binding. Heat shock proteins, one of which is encoded by *groES*, are part of bacterial stress response, indicating that heterogeneity in expression of these genes could lead to differences in stress tolerance within the population. In order to investigate this hypothesis, the long 5' UTR of the *groES* gene from *Mycobacterium smegmatis* (a nonpathogenic relative of *M. tuberculosis*) was used to create a number of strains utilizing the native 5' UTR and several modified variants. The native *groES* promoter and 5' UTR variants were placed in front of a fluorescent reporter and single-cell expression analysis was performed. Native 5' UTR components altered the mean fluorescence levels as well as the distribution of fluorescence levels within the populations relative to a synthetic 5' UTR with a strong ribosome binding site. Our results indicate that 5' UTR composition dramatically affects expression levels and contributes to phenotypic heterogeneity.

## Introduction

Tuberculosis (TB) is an infectious disease affecting millions worldwide each year, with an estimated one-third of the global population suffering a latent infection (WHO 2016). Infections of *Mycobacterium tuberculosis*, the causative agent of TB, are often genotypically homogeneous but phenotypically heterogeneous (De Boer et al. 2000). Phenotypic heterogeneity has been demonstrated in protein production; furthermore, such heterogeneity increases when the cells are under stress (Manina et al. 2015). Phenotypic heterogeneity has also been seen in growth rate and growth state, response to host immune system, and stringent response (Aldridge et al. 2012, Corper and Cohn 1933, Boom et al. 1991, Shamputa et al. 2004, Ghosh et al. 2011). The heterogeneity in phenotypic state is of special concern in *M. tuberculosis* as it can lead to the presence of non-growing cells in a population, which can confer phenotypic antibiotic tolerance as most antibiotics require actively growing bacteria to be effective (Stewart et al 2003). Non-growing cells may persist after a course of treatment, allowing a resurgence of the disease. This is one reason why treatment of TB takes six months or more (WHO 2016).

The GroE chaperone system, which is comprised of GroEL and its cochaperone GroES, is made up of two large oligomeric barrels (GroEL) and an oligomeric cap structure (GroES) (Walter and Storz 2002). The GroE system mediates protein folding by encapsulating the missfolded protein and providing an environment where it can fold correctly (Horwich et al. 2006, Grallert and Buchner 2001). The genomes of *M. tuberculosis* and *M. smegmatis* contain two copies of the *groEL* gene (*groEL1* and *groEL2*), only one of which is essential (Ojha et al. 2002). In *M. smegmatis groES* is essential (Eric Rubin, personal communication), and is directly upstream of the nonessential copy of *groEL*. *groES* is essential in *M. tuberculosis* as well, and the sequence is highly conserved throughout mycobacteria (Sasseti et al 2003, Sasseti and Rubin 2003, Altschul et al 1997, DeJesus et al 2017). GroES and GroEL are heat shock proteins, meaning they are upregulated when stressors, canonically heat, are applied to the organism. They have been shown to be upregulated in response to increased temperatures and multiple other stresses, including macrophage infection (Stewart et al 2002, Aravindham et al 2009). There are

indications that the GroE system is under regulation of the repressor HrrA, though results are inconclusive (Steward et al 2002). The *groES* 5' untranslated region (UTR) is 130 nt long, based on unpublished transcription start site data from our lab. The relatively long length of this 5' UTR suggested the possibility that it has a regulatory function.

5' UTRs house ribosome binding sites and may also serve as sites of regulation of translation and control over the degradation of the transcript. Secondary structure, or lack thereof, in the 5' UTR can influence transcript stability. Hairpin structures in the 5' UTR have been shown to have a stabilizing effect on mRNA, possibly by protecting the transcript from degradation by 5'-to-3' exonuclease activity, or sterically inhibiting endonuclease activity (Chen et al. 1991, Emory et al. 1992, Unniraman et al. 2001). For example, the endonuclease RNase E binds to mRNAs at the 5' end before degrading them, an activity potentially prohibited by a hairpin structure at the binding site (Bouvet and Belasco 1992). Hairpin structures change the binding dynamics of mRNA and RNase E and alter the rate and duration at which such associations occur, which the catalytic activity is contingent upon. The binding of RNase E is stimulated by the dephosphorylation of the 5' triphosphate to a monophosphate, which in turn requires the activity of RppH (Deana et al. 2008). The ability of RppH to dephosphorylate can be modulated by 5' structures, which can stabilize or block a conformation suitable for this activity (Rauhut and Klug 1999). All of this is predicated on the conformation of the 5' UTR and its interactions with cellular machinery.

The 5' UTR can affect translation efficiency through its effects on ribosomal binding. Specifically, higher affinity for the ribosome leads to higher translation efficiency (Hall 1996). The Shine-Dalgarno sequence within the 5' UTR binds the 30S ribosomal subunit during initiation of translation (Shine and Dalgarno 1974). This is the first step in the recruitment of the ribosome and translation initiation factors. It should be noted however that Mycobacteria have an extensive array of leaderless transcripts, which have no 5' UTR and thus no Shine-Dalgarno sequence, which are none the less translated, using an alternate mechanism to that described above (Shell 2015). When present, the 5' UTR can serve as a site for regulation. These regulatory 5' UTRs are also known as riboswitches, which change conformation in accordance with cellular stimuli (Livny et al. 2010, Waters and Storz 2009). Riboswitches can alter conformation based on

the presence or absence of metabolites or sRNAs, or in response to changes in temperature (Waters and Storz 2009, Krajewski and Narberhaus 2014). The binding of ligands or change in temperature causes a conformational change in the mRNA, such as the formation or disassociation of a hairpin structure, thus changing the availability of the RBS or altering translation dynamics (Waters and Storz 2009). This allows for agile regulation in response to environmental conditions at the level of the transcript.

Heterogeneity in gene expression within a clonal population is a result of the combined effects of intrinsic and extrinsic noise (Elowitz et al 2002). Extrinsic noise arises from variability in the cellular concentrations of proteins and other molecules involved in gene expression, while intrinsic noise refers to the inherent randomness of molecular interactions during the processes of gene expression (Swain et al 2002). Thus the inherent randomness in processes such as transcription, degradation, and translation produce intrinsic noise. Intrinsic noise is also a function of the gene sequence and the sequence of the gene product (Swain et al 2002). Therefore, as the 5' UTR can influence transcript stability and translation efficiency as stated above it becomes a contributor to stochasticity in expression. Because the 5' UTR can affect transcript stability and translation efficiency in different ways depending on the sequence, we hypothesize that different 5' UTR sequences will differentially affect noise, or heterogeneity in expression.

Here we sought to elucidate the role of the long 5' UTR of *groES* from *M. smegmatis.* We created reporter constructs to examine the contributions of the 5' UTR to transcript stability, translation efficiency, and heterogeneity of gene expression. Our results indicate that 5' UTR composition can dramatically affect gene expression levels. We found that the native 5' UTR decreases mean expression relative to a short synthetic 5' UTR containing a strong RBS and produces an altered distribution of gene expression levels in the population. We found that the full-length native 5' UTR confers lower protein expression levels than a truncated variant of the native 5' UTR containing only the RBS, as well as lower deviation of expression within the population; however, more of that deviation arises from extreme values.

## Methods and Materials

### Bacteria and Growth Conditions

All experiments were conducted using a derivative of *Mycobacterium smegmatis* strain $mc^2155$ that has a transposon insertion in MSMEG_2952 and does not clump together as much as wildtype $mc^2155$. The strain was obtained from Anil Ojha and in the Shell lab is designated SS-M_0023. *M. smegmatis* was grown in 7H9 broth (Middlebrook) or on 7H10 agar (Middlebrook), with 100 mL/L 10x ADC and 150 µg/mL Hygromycin. 10x ADC is comprised of BSA Fraction V (50 g/L), dextrose (20 g/L), NaCl (8.5 g/L), and catalase (30 mg/L), combined with $H_2O$ and filter sterilized. Cells were grown at 37°C with agitation of 175 RPM for ~24 hrs, then normalized and grown to $OD_{600}$ 0.60 over a further ~24 hour period before use.

### Plasmid and Strain Construction

Plasmids were constructed using restriction digest cloning. The putative native promoter (the region 418 to 131 nt upstream of the *groES* start codon) and 5' UTR (the 130 nt region directly upstream of the *groES* start codon) of the *M. smegmatis groES* were amplified from genomic DNA by PCR. The enhanced green fluorescent protein (eGFP) codon-optimized (co)-mEmerald sequence was obtained from the laboratory of Christopher Sassetti. The Native RBS strain (See Table 1) utilizes the putative native *groES* ribosome binding site (RBS) with sequence CCTTTAACTAGTGGAGGGCTCCATC, a region 25 nt directly upstream of the *groES* start codon. The Synthetic RBS strain utilizes a synthetic RBS with the sequence GAAGGAGAT, which is widely used in mycobacterial expression vectors, in the context of a 25 nt 5' UTR with the sequence TGCAGAATTCGAAGGAGATATACAT. Each fragment was constructed with overlapping sequence and terminal restriction digest sites. This allowed them to be stitched together using PCR, where each fragment was combined in the proper order through PCR amplification using primers that added a region of overlapping sequence to one end in one reaction, then a subsequent reaction was used to combine the fragments into a single insert for each plasmid. Inserts were cut using

endonucleases HindIII and EcoRI and ligated into the vector pSS047. pSS047 is an integrating plasmid which inserts into the genome at the Giles phage integration site via site specific integration (Morris et al2007). Due to issues with these strains during qPCR the plasmids and strains were recreated in a new vector using yellow fluorescent protein (YFP) instead of co-mEmerald. The previously created plasmids were used as templates to amplify the inserts which were then inserted into a new vector by Gibson assembly according to the manufacturer's instructions (Synthetic Genomics kit). A new plasmid was used as a backbone, pSS076, a variant of pMV762 with additional terminators upstream and downstream of several inserted genes. This plasmid is episomal. Using PCR and Gibson assembly, an empty version of pSS076 was cloned without the eccC1b, MSMEG_0063, MSMEG_0064, MSMEG_0065, MSMEG_0066 genes, a 2.5 kb region (2579 bp). A 20 bp fragment of eccC1b was left directly following the upstream terminator to make assembly easier, as terminators can inhibit this. Fragments from the first series of plasmids (Table 1) were then inserted. Strains were constructed through electroporation of competent *M. smegmatis* cells with the plasmids described above.

RNA Extraction

Each *M. smegmatis* strain was grown in 5 mL cultures of 7H9 and normalized to reach an $OD_{600}$ of 0.60 simultaneously, then frozen using liquid nitrogen and stored at -80°C. Samples were thawed on ice then centrifuged for 5 min at 20°C. The supernatant was decanted and the pellet resuspended in 1 mL TRIzol (Invitrogen) and transferred to bead-beating tubes (MP Biochemicals), then beating commenced in a FastPrep 5G (MP Biochemicals) in two instillations for 40 sec each at 9 msec, with 5 min incubation on ice intervening. 300 µL of chloroform were added, then the tubes were centrifuged for 15 min at 4°C. The aqueous layer was then pipetted off and added to 600 µL chilled isopropanol. Tubes were incubated at -20°C for 30 min, then centrifuged for 10 min at 4°C. The supernatant was decanted and 1 mL 75% ethanol was added. The tubes were centrifuged for 1 minute and the supernatant decanted, with the remainder pipetted off. Tubes were incubated at room temperature for 10 min to dry the pellets. The pellet was then resuspended in 100 µL RNase-free $H_2O$. RNA was then treated to degrade DNA using

DNase Turbo (Ambion). Cleanup was conducted according to the RNeasy Mini Kit instructions (Qiagen), except that a new collection column was used at each wash step. A third wash step was added, and elution was done using 50 µL $H_2O$. The resulting purified RNA was stored at -80°C.

cDNA Synthesis

When making cDNA everything is done in duplicate, with one set containing reverse transcriptase while the other acts as a control and contains an equivalent amount of $H_2O$ instead. To synthesize cDNA, 1 µg purified RNA was combined with 0.5 µL 100 mM Tris pH 7.5 and 0.5 µL random hexamers (Invitrogen) at 1 mg/mL and enough $H_2O$ for the total reaction volume to reach 6.25 µL. The reaction was then heated at 70°C for 10 min, then snap-cooled in an ice-water bath. With the addition of 2 µL 5X ProtoScript II reaction buffer (NEB), 0.5 µL 10 mM each dNTPs, 0.5 µL 100 mM DTT, 0.25 µL Murine RNase Inhibitor (NEB), and 0.5 µL ProtoScript II reverse transcriptase (NEB) (or $H_2O$ for the no RT controls) the reaction volume was 10 µL. The samples were then incubated at 25°C for 10 min, then 42°C for 2 hrs. Then 5 µL 0.5 µM EDTA and 5 µL 1 N NaOH were added and the samples were incubated at 65°C for 15 min. 12.5 µL of 1M Tris-Hcl pH 7.5 was added, and from this point forward cleanup was carried out using the PCR and DNA Cleanup Kit (Monarch) with the following exceptions. When binding buffer was added 325 µL was used, wash buffer was added with the samples still inside the centrifuge to reduce moving them, and elution was done using 30 µL.

Quantitative PCR (qPCR)

Previous data indicated that the qPCR primers SSS833 (GATAGCACTGAGAGCCTGTT) and SSS834 (CTGAACTTGTGGCCGTTTAC) were appropriate for YFP, with cDNA at 80 pg/µL (de Rivera, 2016). Primers were added at a final concentration of 0.25 µM along with iTaq SYBR supermix (BioRad) and $H_2O$. Once samples were transferred to a 96 well plate (Axygen), qPCR was run on an Applied Biosystems 7500 qPCR machine with the following parameters. The reaction volume was 10 µL

and samples were incubated at 50°C for 2 min, then 95°C for 10 min, and 40 cycles of 95°C for 15 sec followed by 61°C for 1 min, then 95°C for 15 sec, and finally 60°C for 1 min, with the assay type standard curve. The process was the same as above with the following modifications. Triplicate reactions were done for each cDNA with 80 pg cDNA per reaction.

Quantitative Microscopy and Image Analysis

Three isolates of each *M. smegmatis* strain were grown in 5 mL cultures of 7H9 overnight then each culture was split into four subcultures and normalized to reach an $OD_{600}$ of 0.60 simultaneously. This was done to counteract issues with normalization and loss of fluorescence of some cultures that had been observed in earlier experiments. Cultures were then stored at 4°C for 24 to 36 hrs. Immediately prior to microscopy all samples were incubated at 37°C for 30 min and then at 4°C for at least 15 min. Images were taken using a Zeiss AX10 at 400x magnification using an ApoTome module and settings. ApoTome functions by superimposing a grid over the acquisition area and shifting the grid as multiple focal planes are captured, keeping bright regions from bleeding into less bright regions and distorting the fluorescence readings. Z-stack images were taken to image fluorescence in multiple planes within the cells. The Z-stack images were processed to the same display settings and channels, using a sample of SS-M_0060 (a strain that expresses YPF from a strong promoter) as a benchmark for fluorescence readings, such that all images were processed to have the same black, white, and gamma levels in all imaging channels. Using ImageJ software the Z-stacks were converted to grey scale (Schneider et al 2012). A Z Projection was then created for each image by combining the Z-stack slices into a single image using slice summing. Cells were then isolated using polygon selection and the mean intensity measured. Clumped cells were disregarded in the cases where individual cells could not be distinguished. Cells that moved during imaging were disregarded as the slices did not sum to the same location, making analysis impossible.

## Results

### Design and Construction of YFP Reporter Strains

To investigate the role of the 5' UTR of *groES* in *M. smegmatis*, a number of strains were created, using YFP to report expression levels (Figure 1 and Table 1). Using YFP allowed the use of fluorescence microscopy to ascertain gene expression levels. To determine the effect of the full 5' UTR a plasmid was made which included the native *groES* promoter and 5' UTR along with YFP (Full 5' UTR strain). To study the extent of the impact of the 5' UTR beyond housing the RBS a strain was created using the native promoter and the RBS from the native 5' UTR, without the remainder of the 5' UTR sequence (Native RBS strain). In order to determine the relative strength of the native RBS another strain was created using a synthetic RBS (Synthetic RBS strain). The untranslated regions of the native and synthetic RBS strains was kept the same length to eliminate the relative position of the RBS to the start codon as a confounding factor. As a negative control a strain was created using the full native 5' UTR but no promoter, demonstrating that the reporter system was in fact under the control of the native promoter (No Promoter strain). An empty vector control was also created, which contained no *groES* components nor YFP, allowing the determination of autofluorescence levels (Empty Vector strain). All plasmids were sequence confirmed and all strains were confirmed to contain their respective plasmids. The systematic strain name was used in internal lab documentation, while the shorthand designation will be used throughout this report.

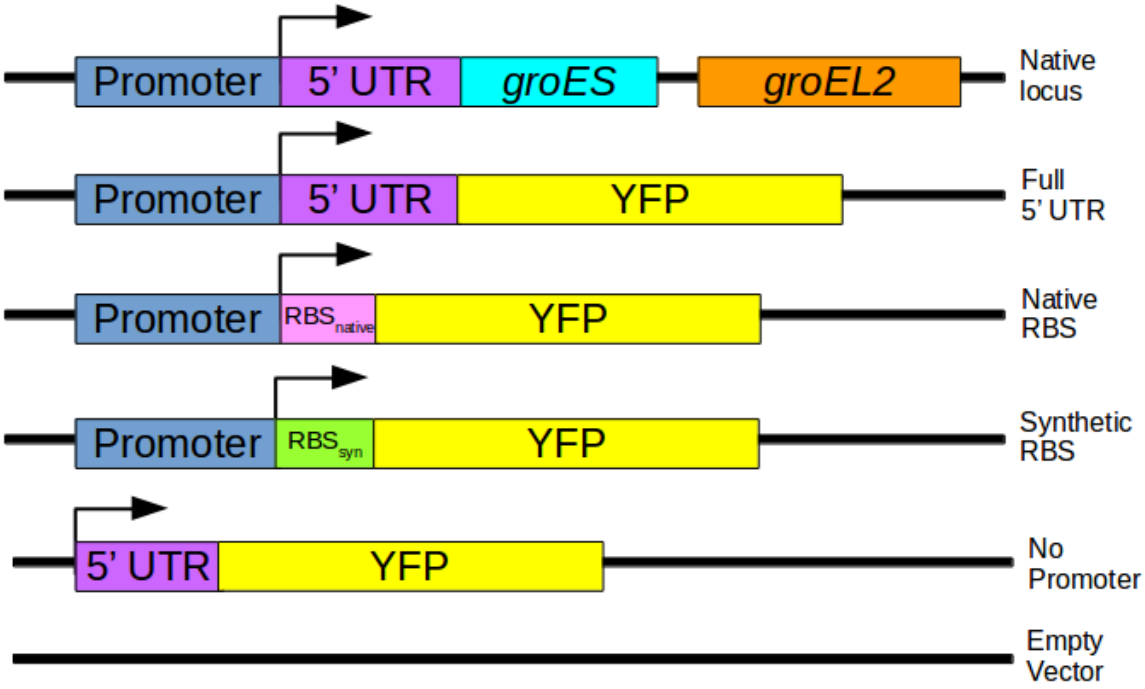


Figure 1. Diagram of the plasmid inserts in each reporter strain. The arrow indicates the transcription start site (TSS). Sizes of elements in the diagram are not to scale. *groES* promoter (288 bp, positions -418 to -131 relative to the *groES* start codon), *groES* 5' UTR (130 bp), *groES* coding sequence (303 bp), *groEL2* coding sequence (1623 bp), $RBS_{native}$ (25 bp) denotes RBS from *groES* 5' UTR, $RBS_{syn}$ (25 bp) denotes synthetic RBS, YFP (720 bp). Diagram denotes relevant linearized fragments of circular DNA for ease of display.

Table 1. Strain Key indicating strain designation, plasmid used in each strain, and a brief description of the relevant contents of each plasmid.

| Shorthand Designation | Systematic Strain Name | Plasmid | Contents |
|---|---|---|---|
| Full 5' UTR | SS-M_0180 | pSS196 | Vector, Native Promoter, Native 5' UTR, YFP |
| Native RBS | SS-M_0182 | pSS198 | Vector, Native Promoter, Native RBS, YFP |
| Synthetic RBS | SS-M_0183 | pSS199 | Vector, Native Promoter, Synthetic RBS, YFP |
| No Promoter | SS-M_0184 | pSS200 | Vector, Native 5' UTR, YFP |
| Empty Vector | SS-M_0185 | pSS201 | Vector |

5’ UTR Composition Dramatically Affects YFP Expression

To understand protein expression levels individual cell fluorescence was measured using fluorescence microscopy. Twelve cultures per strain were normalized to reach an approximate OD of 0.6 simultaneously, with the exception of the Empty Vector Strain, of which only six cultures were made because loss of fluorescence was not an issue. In order to determine if fluorescence levels were dependent on optical density the samples were taken out of incubation at a range of ODs (Table 2). Mean fluorescence was used as the reported value as this controls for cell size. Data are presented in arbitrary fluorescence units (AFU).

Table 2. Optical Density of cultures used in fluorescence microscopy, where each value is the OD of a culture†.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Full 5’ UTR | 0.50 | 0.53 | 0.53 | 0.56 | 0.57 | 0.62 | 0.73 | 0.74 | | |
| Native RBS | 0.57 | 0.63 | 0.63 | 0.65 | 0.65 | 0.68 | 0.86 | | | |
| Synthetic RBS | 0.51 | 0.53 | 0.53 | 0.55 | 0.60 | 0.60 | 0.67 | 0.70 | 0.80 | 0.86 |
| No Promoter | 0.53 | 0.60 | 0.60 | 0.66 | 0.67 | | | | | |
| Empty Vector | 0.62 | 0.68 | | | | | | | | |

†Differences in the number of strains is due to difficulties in getting the strains to normalize and the loss of fluorescence in some strains.

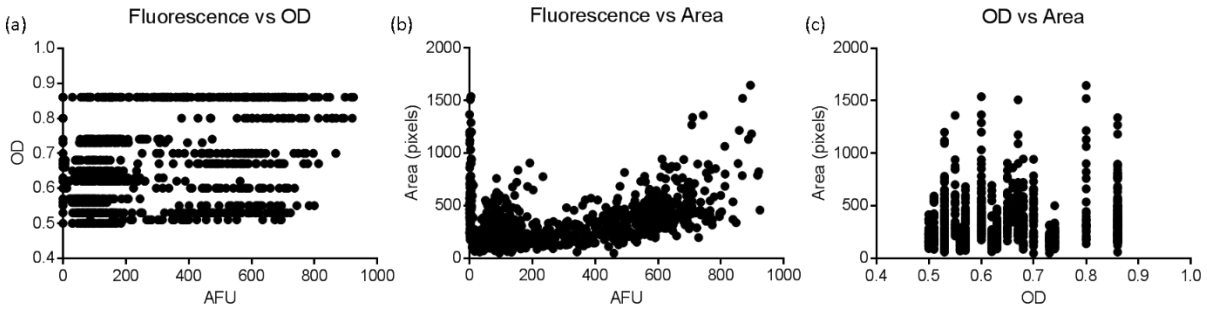


Figure 2. (a) Relationship between mean fluorescence per cell and culture OD at time of harvest. Spearman’s ρ =0.1385, $p < 0.0001$. (b) Relationship between mean fluorescence per cell and cell area. Spearman’s ρ = 0.3672, $p < 0.0001$. (c) Relationship between OD and area. Spearman’s ρ = 0.2319, $p < 0.0001$.

Table 3. Correlation analysis of the relationship between mean fluorescence per cell, Optical Density (OD) at harvest, and cell area. Area correlations for No Promoter or Empty Vector strains not shown as area was not measured for non-fluorescing cells.

| | | Spearman's ρ | *p* value |
|---|---|---|---|
| Full 5' UTR | Fluorescence vs OD | 0.2794 | < 0.0001 |
| | Fluorescence vs Area | 0.07343 | 0.1258 |
| | OD vs Area | 0.02649 | 0.5811 |
| Native RBS | Fluorescence vs OD | 0.1410 | 0.0004 |
| | Fluorescence vs Area | 0.1863 | 0.0095 |
| | OD vs Area | -0.0005915 | 0.9935 |
| Synthetic RBS | Fluorescence vs OD | 0.2610 | < 0.0001 |
| | Fluorescence vs Area | 0.7005 | < 0.0001 |
| | OD vs Area | 0.1902 | 0.0002 |
| No Promoter | Fluorescence vs OD | -0.1592 | 0.0074 |
| Empty Vector | Fluorescence vs OD | -0.05695 | 0.5265 |

To ensure that differences in fluorescence between the strains was due to differences in YFP levels and not a confounding variable, correlation analysis was conducted to establish the relationships between fluorescence and optical density, fluorescence and area, and area and optical density. While all three relationships were positive (Figure 2), none was great enough to be considered responsible for confounding the data. The relationships were further parsed by strain (Table 3) and demonstrated the variability in relationship by strain. A Spearman correlation was used because the data were assumed to be non-parametric. Given that all strains were samples over similar OD ranges, the magnitude of the Spearman's rho values was not sufficient to cast doubt on interpretations of the data disregarding area and optical density.

While the strains used in this study were made from a non-clumping strain of *M. smegmatis* (SS-M_0023) some clumps still persisted, and were ignored for analysis as individual cells could not be accurately outlined for measurement. There were fewer Native RBS strain cells

than the other two strains because the individual images contained fewer cells and two of the biological replicate cultures had uniformly lost fluorescence and were therefore excluded from analysis.

Levels of autofluorescence were defined as those displayed by the Empty Vector Strain (Figure 3f) as it contained no YFP gene. Similar values can be seen in the No Promoter strain (Figure 3e), indicating that detectable YFP levels were not produced in the No Promoter strain. All strains produced some cells that fluoresced with an intensity of 10 AFU or lower, which we defined as non-fluorescing cells. Background fluorescence, defined as the lowest fluorescence value in each image, ranged from 0 to 10 AFU after processing, indicating that the background level may be a result of the conversion undertaken for image analysis. The background level for each image was subtracted from the fluorescence value for each cell in that image to obtain the values reported here.

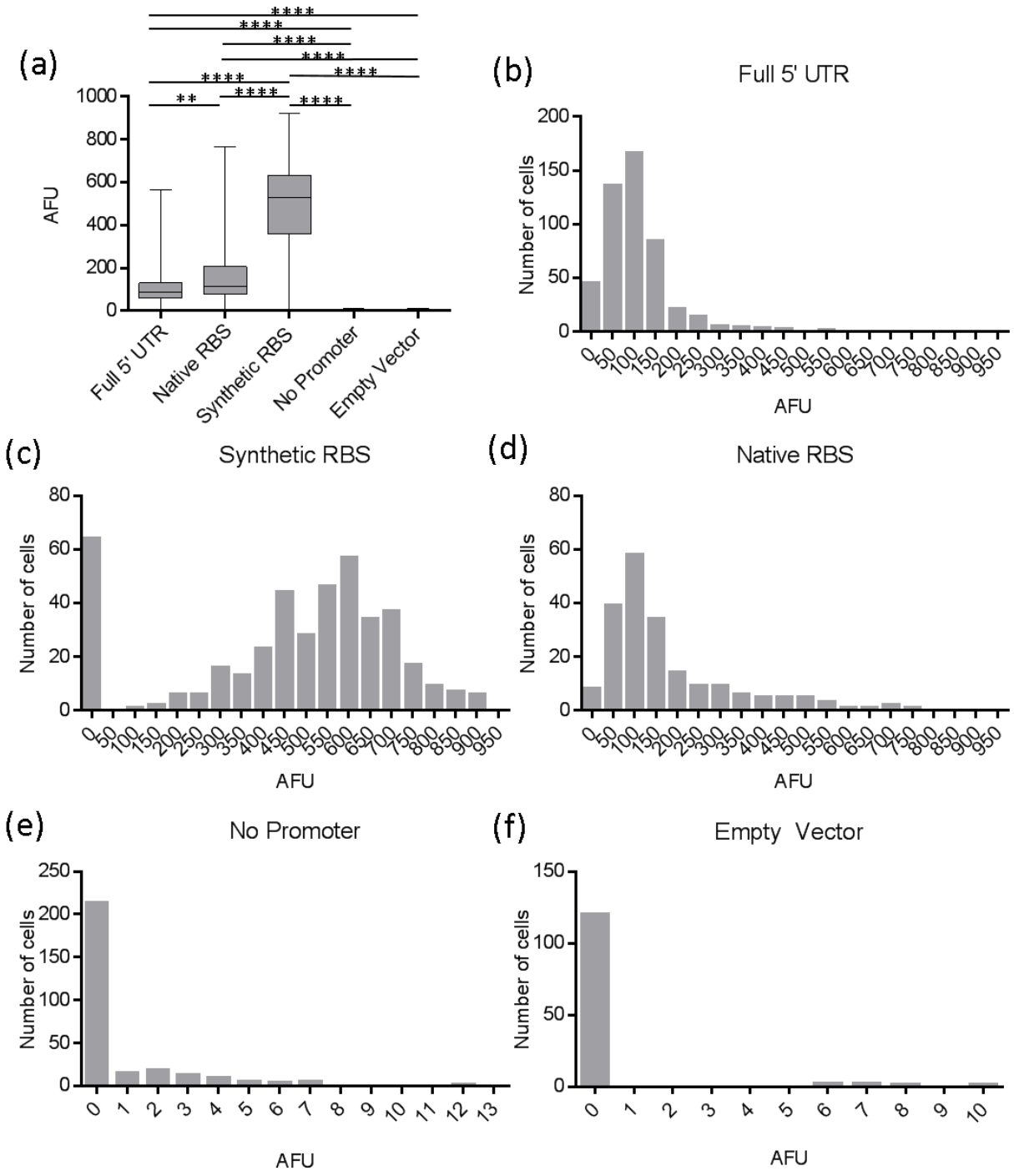


Figure 3. (a) Box and whisker plot indicating intensity of fluorescence by strain, where the grey box indicates the interquartile range, the line within indicates the median, and the lines extending from the box indicate maximum and minimum values. Significance according to a Kruskal-Wallis test followed by Dunn's multiple comparisons test, ** $p$ <0.01, **** $p$ < 0.0001. (b-f) Histograms of strains sorted into bins according to arbitrary fluorescence units (AFU) of each cell.

The mean rank difference in fluorescence between the strains was significant for all pairwise comparisons, with the exception of the No Promoter and Empty Vector strains which were indistinguishable (Figure 3A). Among fluorescing strains, the Synthetic RBS strain was brightest, followed by the Native RBS and then Full 5' UTR strains. While the difference in median fluorescence was significant between the Full 5' UTR and Native RBS strains, the difference between those two strains was less than the difference between either of them and the Synthetic RBS strain.

A distinct subpopulation of non-fluorescing cells appeared within the Synthetic RBS strain, comprising 10.6% of the total population. The Full 5' UTR and Native RBS strains had 9.4% and 3.5% non-fluorescing cells, respectively; however, these population were less readily distinguishable from the major population than in the Synthetic RBS strain. Interestingly, the Full 5' UTR and Synthetic RBS strains had similarly sized populations of non-fluorescing cells while the Native RBS strain had fewer proportionally.

Table 4. Analysis of fluorescence microscopy of *M. smegmatis* strains.

| | Full 5' UTR | Native RBS | Synthetic RBS | No Promoter | Empty Vector |
|---|---|---|---|---|---|
| Number of Cells | 481 | 200 | 416 | 282 | 126 |
| Average Optical Density | 0.60 | 0.71 | 0.66 | 0.63 | 0.65 |
| Median | 87.99 | 113.5 | 529.2 | 0 | 0 |
| Interquartile Range | 71.75 | 130.41 | 273.7 | 0.1205 | 0 |
| Mean | 102.7 | 170.3 | 467.7 | 0.7777 | 0.3497 |
| Standard Deviation | 74.98 | 150 | 242.4 | 1.718 | 1.593 |
| Coefficient of Variation | 0.730088 | 0.880799 | 0.518281 | 2.209078 | 4.555333 |
| Skewness | 1.86 | 1.699 | -0.7197 | 2.88 | 4.514 |
| Excess Kurtosis | 6.241 | 2.657 | -0.351 | 10.33 | 19.41 |

Median fluorescence give an incomplete picture, so to determine the effect of the 5' UTR on the distribution of fluorescence within each population we compared the strains using a number of metrics, including skew. The Full 5' UTR strain is the furthest right skewed, and the Native RBS strain also exhibits a relatively high degree of right skewing (Table 4). Skewing in the Synthetic RBS strain is slightly left skewed, and this skew is not diminished when the non-fluorescing population is excluded. Skewing in the No Promoter and Empty Vector strains can be ignored, as the fluorescence in these groups is considered indistinguishable from autofluorescence.

Interquartile range and standard deviation suggest that the Synthetic RBS strain had much of its elements distributed across a large range, whereas the Full 5' UTR and Native RBS strains were more confined around their median values, with the exception of their greater tailing. According to a Shapiro-Wilk normality test none of the strains exhibited normal distributions. The coefficient of variation, an indication of the degree of noise in the data, serves as a useful metric of heterogeneity in data sets with largely divergent means as it reflects standard deviations normalized to the mean. The Synthetic RBS strain has the lowest coefficient of variation, suggesting that it has relatively lower heterogeneity than the strains with native 5' UTR components (Table 4). The data suggest that the Native RBS strain had the greatest noise, or randomness in its data, followed by the Full 5' UTR strain. This indicates that the full 5' UTR decreases stochasticity in expression relative to the native RBS alone, decreasing heterogeneity. The native 5' UTR confers a degree of stochasticity that is less than the RBS alone, while increasing the population of non-expressing cells.

<u>The Native RBS and Synthetic RBS strains appear to have similar transcript abundance</u>

To measure transcript abundance, quantitative PCR was used to determine the relative expression of YFP in each strain. Each strain was analyzed in biological triplicate, with the exception of the Empty Vector strain, which was not analyzed by qPCR. YFP expression was expressed relative to expression of *sigA*, which encodes an essential sigma factor in the $\sigma^{70}$ family. As *sigA* is expressed at a stable level within the cell it is a useful benchmark for relative expression

profiling. Controls indicated that genomic DNA contamination and amplicon contamination of the reagents were not issues for this experiment.

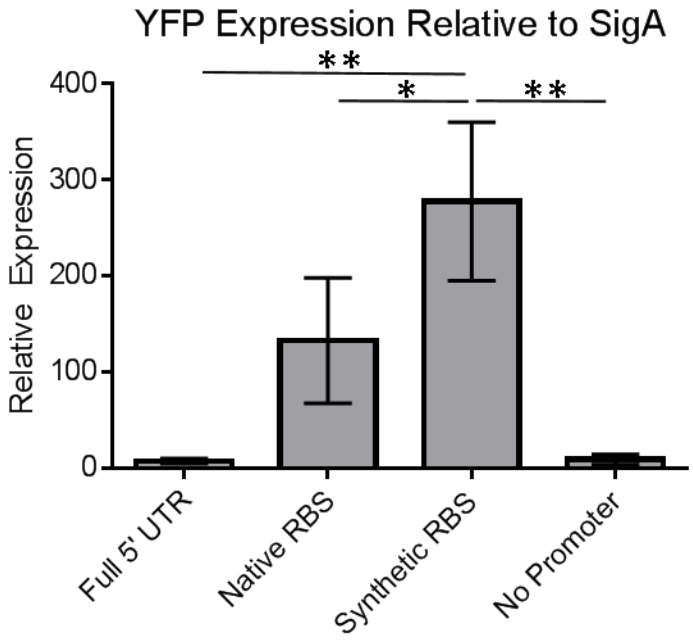


Figure 4. Mean YFP expression relative to *sigA*, with values presented as the mean of biological triplicates for each strain and error bars of standard deviation. Significance according to ANOVA followed by Tukey's multiple comparison. * $p < 0.05$, ** $p < 0.01$

The Synthetic RBS strain displayed the highest transcript abundance and was significantly different from all other strains (Figure 4). The Full 5' UTR, Native RBS, and No Promoter strains were not significantly different, however this relationship requires further testing to validate. Given the fluorescence microscopy results in Figure 3B the relative expression values in Figure 4 are viewed with caution, as fluorescence in the Full 5' UTR and Native RBS strains were comparable and significantly higher than that of the no promoter strain. If true, these data indicate that transcripts with native 5' UTR components have lower stability, as they are under the control of the same promoter as the Synthetic RBS strain but have significantly lower transcript levels. The impact on translational efficiency is unclear as native 5' UTR containing strains had both lower fluorescence and lower abundance levels than the Synthetic RBS strain.

This experiment should be repeated to determine if this is the result of an actual discrepancy in the relationship between transcript and protein levels or the result of technical errors.

**Discussion**

A 5' UTR can have a regulatory role for its transcript, potentially affecting transcript stability, translation efficiency, and heterogeneity in expression. The 5' UTR of the mycobacterial *groES* gene contains the RBS and over one hundred additional nucleotides. This extended sequence suggests that it has some functionality beyond simply housing the RBS. In this study we sought to elucidate the role of the 5' UTR of *groES* from *M. smegmatis*. To accomplish this a number of reporter strains were created using YFP, the native *groES* promoter, and variations of 5' UTR. The strains were grown and transcript and protein levels were measured with the intent of examining and correlating the data to determine the effect of the 5' UTR component.

Protein levels were measured by fluorescence microscopy, allowing the determination of fluorescence levels of individual cells. While this limited the number of cells it was possible to analyze, it was the most sensitive and accurate single cell analysis method available. In the Full 5' UTR and Native RBS strains the median fluorescence was closer to zero than was the case in the Synthetic RBS strain, such that the non-fluorescing cells are distinguishable as a separate population only in the Synthetic RBS strain. The sizes of these populations differed between the strains, with the Native RBS strain having fewer than half the proportional population of non-fluorescing cells than did the other two strains. This indicates that the RBS alone is not responsible for this degree of heterogeneity in phenotypic state, that the extended 5' UTR sequence of the Full 5' UTR strain had a dramatic effect on the number of non-fluorescent cells. While it is unclear what these non-fluorescing cells represent, it seems that the sequence of the 5' UTR can affect the stochastic production of cells with very low gene expression within a population. On an individual gene level this produces a subset of the population that expresses

at very low levels, however if expression was similarly low across the whole genome this could be a factor responsible for cellular quiescence, or total lower metabolic activity, the cause of phenotypic antibiotic tolerance. One limitation of this method is that it does not distinguish between live and dead cells. We do not have any indication of the decay rate of YFP and cannot determine which cells had expired but not yet lysed.

Fluorescence levels between the strains differed significantly, indicating that the 5' UTR component did have an effect on translation efficiency and/or transcript stability. The distributions of fluorescence values for the Full 5' UTR and Native RBS strains had shapes that were similar to each other and distinctly different from that of the Synthetic RBS strain, defined by their skewness, standard deviation, and interquartile range. Strains with native 5' UTR components had lower fluorescence than the Synthetic RBS strain, but relatively similar levels to each other. Excess kurtosis for this data set is difficult to interpret because the standard deviation differs greatly between strains. The coefficient of variation was higher in the Native RBS strain than in the Full 5' UTR strain, indicating that the full native 5' UTR decreased stochasticity in translation or mRNA degradation relative to the RBS alone. In addition lower interquartile range and lower standard deviation between the Full 5' UTR and Native RBS strains indicate that the 5' UTR decreases heterogeneity. However, this interpretation may change if other metrics, such as kurtosis, are used to evaluate heterogeneity. The higher proportion of non-fluorescent cells in the Full 5' UTR strain relative to the Native RBS strain may be explained by the lower mean fluorescence, as the left shoulder of the Full 5' UTR distribution is closer to zero and therefore will contain more low values. Protein levels could be approximated through flow cytometry as well, which would allow for the analysis of more cells, but with reduced sensitivity. Alone, data on protein levels provides an incomplete picture; qPCR was conducted to provide complimentary transcript level data.

The qPCR results were surprising as the Full 5' UTR strain YFP transcript level did not significantly differ from the No Promoter strain while fluorescence levels for these strains did differ significantly. If these results are correct, then it appears that the full *groES* 5' UTR greatly increases the degradation rate of the transcript. However, further testing is required to accept these results, as they could be the result of technical errors, or the result of the cells losing the

plasmid, as we sometimes observed when analyzing cultures by microscopy. Loss of fluorescence may be explained by the drug used for selection. It was later determined that the hygromycin used in this experiment was old and partially degraded, diminishing its efficacy. We expect that the issue of plasmid loss would not persist using a fresher batch of hygromycin. If the Full 5' UTR strain results are indicative of technical errors, doubts are raised on the rest of the data from this experiment. This experiment must be repeated before meaningful conclusions can be drawn with confidence.

If the qPCR results are accurate, the transcript abundance difference between the Native RBS and Synthetic RBS strains is too small to explain the difference in protein abundance, confirming the hypothesis that the difference in YFP protein levels is due to a difference in translation efficiency. The Synthetic RBS is predicted to have a stronger Shine-Dalgarno sequence than the native RBS, suggesting it is more efficient at binding the ribosome, driving up translation rates.

Because the Full 5' UTR, Native RBS, and Synthetic RBS strains use the same promoter it is likely that transcripts are being produced at a similar rate in each strain. The apparent differences in transcript abundance therefore suggest that the stability of the transcripts differs between the strains. Direct measurement of transcript half-lives is required to explore this; however, there was insufficient time in this study to undertake that experiment. The discrepancy could be explained by 5' protection from degradation by the RBS sequence, either by altering ribosome occupancy or by blocking ribonuclease binding sites, or a combination of the two. Another explanation could be that differences in ribosome occupancy are occurring in positions along the transcript other than the 5' UTR, causing differential protection from RNases. Translation efficiency directly affects transcript stability through ribosome binding, which can block access to the transcript by RNases thus increasing stability (Rauhut and Klug 1999, Deana and Belasco 2005, Belasco and Brawerman 2012). Further study of the RBS sequence and its interactions with ribosomes and ribonucleases would be required to elucidate the cause of this phenomenon.

Future work using these strains could focus on the stress response nature of *groES*, and determine what effect heat shock or other stresses have on transcript and protein levels, and what degree of influence the 5' UTR has on these factors. Additionally, the completion of the Leaderless strain, SS-M_0181, would allow investigations into differences between leadered and leaderless transcripts. Fluorescence activated cell sorting (FACS) could be used to divide the cells into distinct populations and allowing them to regrow to see if there were differences between them, or if the parameter used to sort them was maintained after one or more division cycles. It would be interesting to see if the non-fluorescing cells had completely lost the plasmid or acquired a mutation, or if the highest fluorescing cells in a strain would maintain high expression levels a generation or more later.

**References**


Aldridge, Bree B., Marta Fernandez-Suarez, Danielle Heller, Vijay Ambravaneswaran, Daniel Irimia, Mehmet Toner, and Sarah M. Fortune. "Asymmetry and Aging of Mycobacterial Cells Lead to Variable Growth and Antibiotic Susceptibility." *Science* 335, no. 6064 (January 6, 2012): 100–104. doi:10.1126/science.1216166.

Belasco, Joel G., and George Brawerman. *Control of Messenger RNA Stability*. Elsevier, 2012.

Boer, Annette S. de, Kristin Kremer, Martien W. Borgdorff, Petra E. W. de Haas, Herre F. Heersma, and Dick van Soolingen. "Genetic Heterogeneity in Mycobacterium Tuberculosis Isolates Reflected in IS6110 Restriction Fragment Length Polymorphism Patterns as Low-Intensity Bands." *Journal of Clinical Microbiology* 38, no. 12 (December 2000): 4478–84.

Boom, W. H., R. S. Wallis, and K. A. Chervenak. "Human Mycobacterium Tuberculosis-Reactive CD4+ T-Cell Clones: Heterogeneity in Antigen Recognition, Cytokine Production, and Cytotoxicity for Mononuclear Phagocytes." *Infection and Immunity* 59, no. 8 (August 1, 1991): 2737–43.

Bouvet, Philippe, and Joel G. Belasco. “Control of RNase E-Mediated RNA Degradation by 5’-Terminal Base Pairing in E. Coil.” *Nature* 360, no. 6403 (December 3, 1992): 488–91. doi:10.1038/360488a0.

Chen, L. H., S. A. Emory, A. L. Bricker, P. Bouvet, and J. G. Belasco. “Structure and Function of a Bacterial mRNA Stabilizer: Analysis of the 5’ Untranslated Region of ompA mRNA.” *Journal of Bacteriology* 173, no. 15 (August 1, 1991): 4578–86.

Deana, Atilio, and Joel G. Belasco. “Lost in Translation: The Influence of Ribosomes on Bacterial mRNA Decay.” *Genes & Development* 19, no. 21 (November 1, 2005): 2526–33. doi:10.1101/gad.1348805.

DeJesus, Michael A., Elias R. Gerrick, Weizhen Xu, Sae Woong Park, Jarukit E. Long, Cara C. Boutte, Eric J. Rubin, et al. “Comprehensive Essentiality Analysis of the Mycobacterium Tuberculosis Genome via Saturating Transposon Mutagenesis.” *mBio* 8, no. 1 (March 8, 2017): e02133-16. doi:10.1128/mBio.02133-16.

Elowitz, Michael B., Arnold J. Levine, Eric D. Siggia, and Peter S. Swain. “Stochastic Gene Expression in a Single Cell.” *Science* 297, no. 5584 (August 16, 2002): 1183–86. doi:10.1126/science.1070919.

Emory, S. A., P. Bouvet, and J. G. Belasco. “A 5’-terminal Stem-Loop Structure Can Stabilize mRNA in Escherichia Coli.” *Genes & Development* 6, no. 1 (January 1, 1992): 135–48. doi:10.1101/gad.6.1.135.

Ghosh, Sayantari, Kamakshi Sureka, Bhaswar Ghosh, Indrani Bose, Joyoti Basu, and Manikuntala Kundu. “Phenotypic Heterogeneity in Mycobacterial Stringent Response.” *BMC Systems Biology* 5 (2011): 18. doi:10.1186/1752-0509-5-18.

Grallert, Holger, and Johannes Buchner. “Review: A Structural View of the GroE Chaperone Cycle.” *Journal of Structural Biology* 135, no. 2 (August 1, 2001): 95–103. doi:10.1006/jsbi.2001.4387.

Hall, Michael N., Joëlle Gabay, Michel Débarbouillé, and Maxime Schwartz. “A Role for mRNA Secondary Structure in the Control of Translation Initiation.” *Nature* 295, no. 5850 (February 18, 1982): 616–18. doi:10.1038/295616a0.

Horwich, Arthur L., George W. Farr, and Wayne A. Fenton. “GroEL−GroES-Mediated Protein Folding.” *Chemical Reviews* 106, no. 5 (May 2006): 1917–30. doi:10.1021/cr040435v.

Krajewski, Stefanie Sandra, and Franz Narberhaus. “Temperature-Driven Differential Gene Expression by RNA Thermosensors.” *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, Riboswitches, 1839, no. 10 (October 2014): 978–88. doi:10.1016/j.bbagrm.2014.03.006.

Livny, Jonathan, and Matthew K. Waldor. “Mining Regulatory 5’UTRs from cDNA Deep Sequencing Datasets.” *Nucleic Acids Research* 38, no. 5 (March 1, 2010): 1504–14. doi:10.1093/nar/gkp1121.

Manina, Giulia, Neeraj Dhar, and John D. McKinney. “Stress and Host Immunity Amplify Mycobacterium Tuberculosis Phenotypic Heterogeneity and Induce Nongrowing Metabolically Active Forms.” *Cell Host & Microbe* 17, no. 1 (January 14, 2015): 32–46. doi:10.1016/j.chom.2014.11.016.

Ojha, Anil, Mridula Anand, Apoorva Bhatt, Laurent Kremer, William R. Jacobs Jr., and Graham F. Hatfull. “GroEL1: A Dedicated Chaperone Involved in Mycolic Acid Biosynthesis during Biofilm Formation in Mycobacteria.” *Cell* 123, no. 5 (December 2, 2005): 861–73. doi:10.1016/j.cell.2005.09.012.

Rao, Srinivasa P. S., Sylvie Alonso, Lucinda Rand, Thomas Dick, and Kevin Pethe. “The Protonmotive Force Is Required for Maintaining ATP Homeostasis and Viability of Hypoxic, Nonreplicating Mycobacterium Tuberculosis.” *Proceedings of the National Academy of Sciences* 105, no. 33 (August 19, 2008): 11945–50. doi:10.1073/pnas.0711697105.

Rao, Tara. “ANALYSIS OF THE MULTIPLE CHAPERONINS OF MYCOBACTERIUM SMEGMATIS,” n.d.

Rauhut, Reinhard, and Gabriele Klug. “mRNA Degradation in Bacteria.” *FEMS Microbiology Reviews* 23, no. 3 (June 1, 1999): 353–70. doi:10.1111/j.1574-6976.1999.tb00404.x.

Rittershaus, Emily S. C., Seung-Hun Baek, and Christopher M. Sassetti. “The Normalcy of Dormancy: Common Themes in Microbial Quiescence.” *Cell Host & Microbe* 13, no. 6 (June 12, 2013): 643–51. doi:10.1016/j.chom.2013.05.012.

Rivera, John de. “The Effects of Post-Transcriptional Processing on mRNA Stability in M. Smegmatis.” *WPI*, March 25, 2016.

Rustad, Tige R., Kyle J. Minch, William Brabant, Jessica K. Winkler, David J. Reiss, Nitin S. Baliga, and David R. Sherman. “Global Analysis of mRNA Stability in Mycobacterium Tuberculosis.” *Nucleic Acids Research*, November 3, 2012. doi:10.1093/nar/gks1019.

Sassetti, Christopher M., Dana H. Boyd, and Eric J. Rubin. “Genes Required for Mycobacterial Growth Defined by High Density Mutagenesis.” *Molecular Microbiology* 48, no. 1 (April 1, 2003): 77–84. doi:10.1046/j.1365-2958.2003.03425.x.

Sassetti, Christopher M., and Eric J. Rubin. “Genetic Requirements for Mycobacterial Survival during Infection.” *Proceedings of the National Academy of Sciences* 100, no. 22 (October 28, 2003): 12989–94. doi:10.1073/pnas.2134250100.

Shamputa, Isdore Chola, Leen Rigouts, Lovet Achale Eyongeta, Nabil Abdullah El Aila, Armand van Deun, Abdul Hamid Salim, Eve Willery, Camille Locht, Philip Supply, and Françoise Portaels. “Genotypic and Phenotypic Heterogeneity among Mycobacterium Tuberculosis Isolates from Pulmonary Tuberculosis Patients.” *Journal of Clinical Microbiology* 42, no. 12 (December 1, 2004): 5528–36. doi:10.1128/JCM.42.12.5528-5536.2004.

Shell, Scarlet S., Jing Wang, Pascal Lapierre, Mushtaq Mir, Michael R. Chase, Margaret M. Pyle, Richa Gawande, et al. “Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape.” *PLOS Genet* 11, no. 11 (November 4, 2015): e1005641. doi:10.1371/journal.pgen.1005641.

Shine, J., and L. Dalgarno. “Determinant of Cistron Specificity in Bacterial Ribosomes.” *Nature* 254, no. 5495 (March 6, 1975): 34–38.

Swain, Peter S., Michael B. Elowitz, and Eric D. Siggia. “Intrinsic and Extrinsic Contributions to Stochasticity in Gene Expression.” *Proceedings of the National Academy of Sciences* 99, no. 20 (October 1, 2002): 12795–800. doi:10.1073/pnas.162041399.

Waters, Lauren S., and Gisela Storz. “Regulatory RNAs in Bacteria.” *Cell* 136, no. 4 (February 20, 2009): 615–28. doi:10.1016/j.cell.2009.01.043.

“WHO | Tuberculosis.” *World Health Organization (WHO)*. Accessed October 6, 2016. http://www.who.int/mediacentre/factsheets/fs104/en/.