5-2023

Determining the Functional Roles of Short Open Reading Frames in Mycobacterium Smegmatis

Lauren Fox

Determining the Functional Roles of Short Open Reading Frames in *Mycobacterium smegmatis*

An honors thesis presented to the
Department of Biology,
University at Albany, State University of New York
in partial fulfillment of the requirements
for graduation with Honors in Biology
and
graduation from The Honors College

Lauren Rachelle Fox

Research Advisors: Keith Derbyshire, Ph.D & Todd Gray, Ph.D
Second Reader: Gabriele Fuchs, Ph.D

May 2023

Abstract

The study examines the roles of a subset of short open reading frames (sORFs) in *M. smegmatis* using targeted mutagenesis and the subsequent examination of phenotypes associated with sORF mutation or overexpression. sORFs are defined as stretches of nucleic acid encoding a protein with a maximum of 50 amino acids. Genome annotation pipelines overlook sORFs encoding small proteins (sproteins) and, thus, sproteins in bacteria have gone unnoticed, even those that are expressed at levels similar to larger, well-described proteins. As a result of recent advances in ribosome profiling and mass spectrometry, hundreds of previously unannotated sORFs have been identified, increasing the number of annotated genes in both the *M. tuberculosis* and *M. smegmatis* genomes. Previously studied sproteins have been found to carry out diverse functions within the cell, ranging from modulating enzymatic activity to stress response signaling. In the present study, annotated sORFs were inspected using a JBrowse genome viewer that displays RNA-seq and Ribo-seq data mapped to mycobacterial reference genomes to determine active gene boundaries. The sORFs chosen for study are thought to be physiologically relevant based on their high expression level and conservation within the *M. smegmatis* genome and across other mycobacterial species. Utilizing targeted mutagenesis protocols, mutant sORFs were created and then assayed for a reproducible phenotypic effect using a variety of molecular and physical assays that aim to accurately address the function of the synthesized sprotein itself, as well as examining the potential of the sprotein to participate in fundamental cellular processes. Studying sORFs and their encoded sproteins has the potential to provide insight into how they modulate biological functions, and to identify novel functions not previously considered due to their seemingly negligible size.

Keywords: *Bacterial small proteins, Mycobacterium smegmatis, short open reading frames, protein function*

Acknowledgements

I'm incredibly lucky to have had several sources of support throughout the duration of my academic endeavors. Their encouragement has made my accomplishments at UAlbany possible.

First, I would like to thank Dr. Keith Derbyshire and Dr. Todd Gray for allowing me the opportunity to conduct research in their lab. I was consistently challenged to think critically about the work I was completing and encouraged to approach problems in ways that allowed me to practice a wide variety of skills. Their patience, enthusiasm, and extensive knowledge has fostered my love for research. I would also like to thank the members of the Derbyshire-Gray Lab: Jill Canestrari, Emma Gordon, Neha Aurora, and Emma Weaver for their support and friendship throughout my time at the Wadsworth Center. I could not have asked for a more supportive and engaging introductory lab experience.

Second, I'd like to thank the UAlbany Biology Department, especially my academic advisor Dr. Robert Osuna, for his assistance in choosing a path of study that suited my interests and for encouraging me to pursue undergraduate research. In addition, I would like to thank Dr. Gabriele Fuchs for her mentorship throughout my undergraduate career, and for taking the time to provide feedback on my thesis. I'd also like to thank Dr. Pauline Carrico for her generosity and support throughout my time at UAlbany.

Third, the RNA Institute Summer Bioinformatics Program, especially Dr. Sweta Vangaveti, was crucial in developing my understanding of RNA-sequencing and bioinformatic pipelines, which allowed me to approach my thesis research from a new perspective.

Lastly, I'd like to thank my family and friends for their tireless support over the past four years, their words of encouragement and inspiration mean the world to me.

List of Tables

List of Figures

Table of Contents

Introduction

The genus *Mycobacterium* encompasses over 100 species of bacteria, including known pathogenic species. These species can cause varied infections across humans and animals alike, ranging from localized skin lesions to more widespread infections of the respiratory system. The most known and researched species include the *Mycobacterium tuberculosis* complex, which causes tuberculosis infections in humans, and *Mycobacterium leprae*, which causes Hansen's disease. These pathogens are especially dangerous in developing nations with poor health infrastructure and to patients with compromised immune systems due to HIV/AIDs or other underlying health issues. Globally, over ten million individuals develop active TB each year, and around ten percent of that population are those living with HIV (Centers for Disease Control, 2023). TB also poses a major public health concern due to the emergence of antibiotic-resistant *M. tuberculosis*, usually resulting from incomplete or mismanaged drug regimens. The main threat is contraction of multi-drug resistant TB, which is more difficult to treat successfully, especially in regions lacking strong public health infrastructure (Centers for Disease Control, 2023). In addition, another non-tuberculosis mycobacterium (NTM) such as *M. abscessus*, has become a prevalent health concern as it is a rapidly growing, innately drug resistant pathogen, and infection often contributes to rapid lung decline in cystic fibrosis patients (Degiacomi et al., 2019).

One *Mycobacterium* species, *Mycobacterium smegmatis*, is a nonpathogenic, GC-rich, bacillus-shaped bacteria that has a doubling time of roughly two hours. These factors make it favorable for genetic research and more readily utilized in a lower biosafety-level laboratory than is required for *M. tuberculosis*. Because of these advantages, *M. smegmatis* can serve as a model organism for research into mycobacterial species, as two-thirds of *M. tuberculosis* proteins share

greater than 50 percent sequence homology with *M. smegmatis* homologs (Sparks et al., 2023). Utilizing a model organism allows both genetic and biochemical experiments to be conducted to determine a gene's function, the relative location of a particular synthesized protein within the cell, and involvement in various other biochemical pathways and mechanisms. Identifying a model non-pathogenic organism that grows faster in a simple medium was vital in the effort to infer gene functions and test hypotheses with the aim of extrapolating data back to studies of both slow-growing mycobacterial pathogens and other non-tuberculosis mycobacterium, as thousands of genes found in *M. smegmatis* are conserved across other mycobacterial species.

An open reading frame (ORF) is defined as the portion of a DNA (and the transcribed mRNA) sequence that is translated into a protein. ORFs contain a series of nucleotide triplets that encode various amino acids, beginning with a start codon encoding methionine and ending with one of three stop codons to signal the end of protein synthesis. Most of the existing bioinformatic genome annotation assumes that ORFs encoding fewer than 50 amino acids (therefore consisting of fewer than 150 nucleotides) would be nonfunctional and not worthy of further study (Miravet-Verde et al., 2019). Thus, many short open reading frames (sORFs) were ignored. In addition, genome annotation algorithms typically rely on hallmark 5' UTR and Shine-Dalgarno sequences important for protein translation in prokaryotes. In mycobacteria, about 25% of mRNAs are leaderless (LL), lacking 5'UTR and Shine-Dalgarno sequences and are therefore poorly annotated by conventional annotation pipelines (Shell et al., 2015) (Smith et. al, 2022). The discovery of sORFs resulted from studying mutations in intergenic regions and unannotated regions of the chromosome, as well as the development of ribosome profiling strategies to determine which open reading frames are actively being translated (Hemm et al., 2010). Transcription profiling (RNA-seq) allows for genome-wide surveys of all actively

transcribed genes while translational profiling (Ribo-seq) identifies those mRNAs that are being translated by ribosomes, and Ribo-RET, which identifies ribosome initiation sites. These combined profiling strategies have identified hundreds of previously unannotated sORFs, greatly increasing the number of genes in both the *M. tuberculosis* and *M. smegmatis* genomes. To determine the role of a subset of sORFs in *M. smegmatis*, both previously annotated sORFs, and sORFs identified by profiling were mutated to determine whether their absence or overexpression results in observable characteristics of the cell. By deleting a gene or making it otherwise nonfunctional through mutation, the resulting phenotype can give insight into the gene's function.

The products of sORF translation, small proteins (sproteins), are less likely to function on their own due to their small size (Andrews & Rothnagel, 2014). Therefore, determining their function is far more difficult than studying larger proteins. Previously studied sproteins have been found to carry out diverse functions within the cell including modulating transport and enzymatic activities (Duval et al., 2017), with many acting near the cell membrane due to their hydrophobic nature (Orr et al., 2020). Therefore, a multi-faceted approach was developed through targeted mutagenesis and the associated phenotype in response to both gene absence and overexpression were studied.

In this thesis, the functions of seven small proteins were investigated by studying the phenotypic effects of knockout, overexpression and fusion with the green fluorescent protein Dendra. While many of the small proteins did not exhibit any observable or reproducible phenotype while conducting these preliminary assays; when mutating one small protein, *msmeg1999*, an altered growth phenotype became apparent in liquid culture. Several assays were developed to better understand how mutations in *msmeg1999* affect the cell's ability to

reproduce within liquid culture, as well as investigations into the transcriptional profile of *msmeg1999* mutants.

Materials and Methods

Parameters for sORF Identification

RNA profiling information from the JBrowse database was analyzed to determine both annotated genes and open reading frames within the *M. smegmatis* genome. The RNA profiling data on JBrowse includes both Ribo-seq data and RNA-seq data to determine gene boundaries and potentially translated sequences. These data can also be utilized to identify possible open reading frames in unannotated sections of the genome, as RNA profiling data is agnostic to genome annotations. Importantly for this study, empirical determination by RNA and Ribo-seq allows identification of previously unannotated genes. An additional criterion that was considered when searching for sORF candidates was gene conservation in other mycobacterial species. This suggests functional importance of the role that a synthesized protein could have within the cell that has been conserved among several bacterial species. For mutagenesis by gene knockout, singleton sORFs were chosen rather than those found within an existing operon, because singleton sORFs can be deleted without disrupting genes nearby, which could complicate interpretations of an observed phenotype.

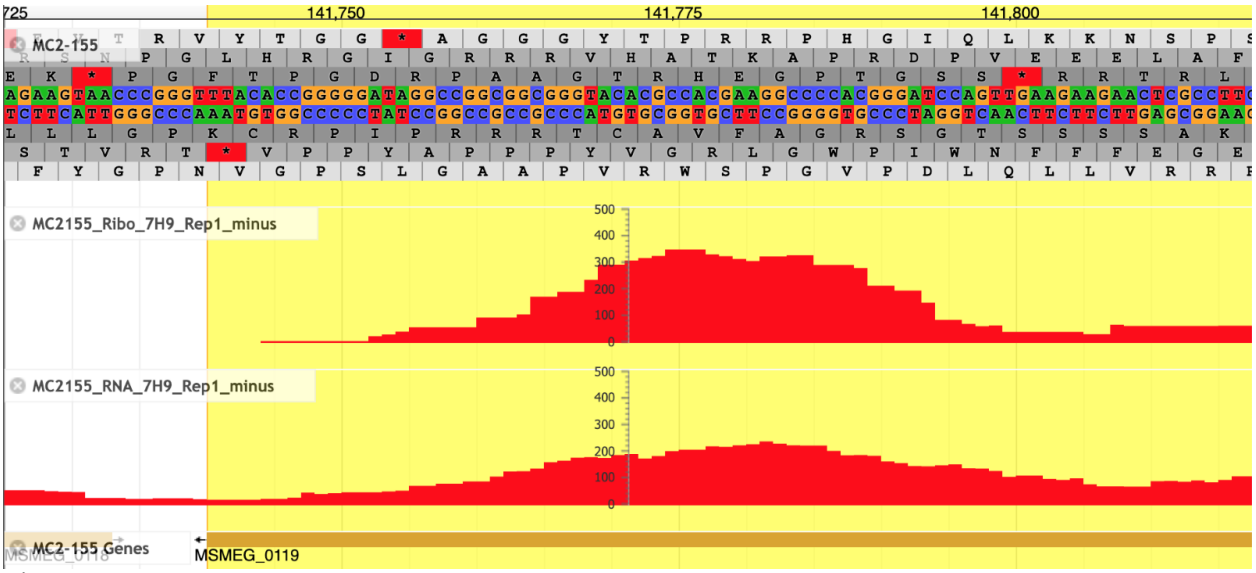

Figure 1. **RNA and Ribosome profiling information for *M. smegmatis* gene 119** (*msmeg0119*), obtained from the JBrowse database (https://www.wadsworth.org /research/scientific-resources/interactive-genomics) to include Ribo-Seq and RNA-Seq data for the determination of the transcribed gene's boundaries, highlighted in yellow. The gene highlighted in this window is expressed in the antisense direction, denoted by the black arrow pointing left, as a leaderless transcript. To describe the structure of the window, starting from the top and working down, the numbers in the gray bar denote the relative position of the gene in relation to the entire sequenced genome. Just below the position number, the associated amino acid sequence is given, followed by the individual nucleotide sequence with each stop codon (*) highlighted in red. The transcript shown in the window begins with a start codon, methionine, at position number 141850 (not shown) and ends at position 141740. The first plot shows the translational (Ribo-seq) profiles for the expressed gene and the second panel towards the bottom of the window depicts the transcriptional (RNA-seq) profiles for the expressed gene. By examining both data sets, it is possible to map the transcriptional and translational start site for a gene based off the relative number of reads, comparing each to the Y-axis scale which is set at a maximum of 500 mapped reads at the given position within genome.

Targeted Mutagenesis

To determine the function of the selected sORFs within the *M. smegmatis* genome, the sORF was either deleted or overexpressed. The objective of mutating the gene of interest is to study the phenotypic effect of cells produced in the absence, or otherwise nonfunctional state of the synthesized protein, or in a state where the sprotein is overexpressed. Gene knockout provides relevant information as to how the cell functions in the absence of the gene of interest's wild-type expression, and any observable phenotypic effects can be studied to attribute a

function to this gene. Overexpressing the gene of interest and studying the mutant phenotype can provide relevant information regarding the protein's function within the cell and examine the functions of their target proteins. Stoichiometry of expression is thought to be important for biological processes, and incremental changes in copy number can cause mutant phenotypes, as the level of each protein is highly optimized to maximize cellular functionality. Therefore, intentional manipulation by overexpression or knockouts of individual genes can provide a useful tool for connecting genes to biological pathways in the cell (Prelich 2012).

Gene Knockout

This approach to targeted mutagenesis is ideal for singleton genes not transcribed as part of an established operon. Gene knockout designs were made for seven sORFs: *msmeg0119, msmeg5154, msmeg1999, msmeg4499, msmeg6401a, msmeg5128a,* and *msmeg3439.* To achieve a knockout of each of these genes, flanking primers were designed and used to amplify the regions upstream and downstream of the gene of interest by PCR. An antibiotic resistance cassette was then introduced in place of the gene of interest by SOEing PCR. The resulting substrate was electroporated into electrocompetent, recombineering-proficient mc^2155 cells and cells were selected by plating on TSA + zeocin at a concentration of 50 µg/mL. Once mutants had been cured of the recombineering plasmid, mutants were subject to assays to determine the phenotypic effects of sORF mutation.

CRISPRi-Mediated Gene Knockout

For loss-of-function mutational assays, the CRISPRi system allows for variable degrees of repression of gene expression and can be used to repress expression entirely, in order to study the cell in the absence of the gene of interest (Rock et al., 2017). CRISPRi allows for sequence-specific repression, and therefore is especially useful for sORFs that are transcribed as an operon.

One gene, *msmeg5128a*, overlaps with downstream genes and is thought to be transcribed as part of an operon. To establish a loss-of-function mutant strain, customized sgRNAs were designed for transcriptional repression of the gene of interest (Larson et al., 2013). In addition, if phenotypic effects are observed, a copy of the gene can be used to restore wild type expression as a comparison to verify the phenotypic effect is a result of the nonfunctional sORF only. The resulting cells were subjected to assays in the same manner as the traditional knockout strains.

Gene Overexpression

To test the phenotypic effect on the cell of overexpressing the sprotein product, each sORF was cloned into the pMSR3 vector backbone (Figure 2) by in-fusion reaction between the HindIII and AseI sites. The pMSR3 vector contains a strong, constitutively expressed mycobacterial promoter (Pmyc) to drive expression of the sORF. Once the sORF has been cloned into the vector, the plasmid was transformed into competent *M. smegmatis* cells, which were subjected to various assays to determine the impact of overexpressing each sORF.

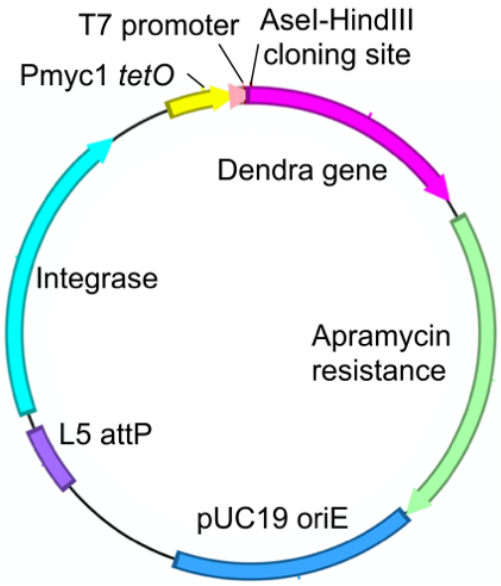

Figure 2. **Plasmid map for sORF clone insertion into pMSR3**, used to drive overexpression of the sORF within a mycobacterial cell.

Dendra Fusion Protein Construct

Each sORF was cloned to create a fusion protein with the fluorescent protein Dendra. Dendra cloning can be utilized in two ways: one that uses live fluorescence microscopy of the expressing bacterium and another that relies on the relatively large protein fusion to disrupt interactions of multiprotein complexes. In order to create this gene fusion, the modified pMSR3 plasmid (Figure 2) was used to clone the sORFs in frame with Dendra. Each cloned sORF was sequence verified, and subsequently electroporated into electrocompetent *M. smegmatis* cells.

Sequence-based and Structure-based Analysis

The goal of this bioinformatic analysis was to use sequence-based and structure-based tools to help inform function of sproteins by identifying known or more well-described proteins that are structurally similar or have significant sequence homology. The dataset used was comprised of five annotated small proteins: msmeg0119, msmeg5154, msmeg1999, msmeg4499, and msmeg0945 and the corresponding nucleotide and amino acid sequences.

Sequence-Based Analysis

The first sequence-based analysis that was performed was a multiple sequence alignment in which amino acid sequences were input into the NCBI BLASTP suite (National Center for Biotechnology Information, 2023). To ensure a well-rounded alignment, non-redundant sequences were selected among the advanced database choices. Next, sequence homology, conservation of protein domains, mutation events, and indels between species were determined for each of the five annotated small proteins. MSA outputs, gene IDs or UniProt IDs were not available to input into a gene ontology analysis, therefore structure-based analysis was utilized to obtain the GO terms for potential molecular functions.

Structure-based Analysis

For structure-based analysis tools, two web-based platforms were utilized: AlphaFold and Iterative Threading Assembly Refinement (I-TASSER). AlphaFold (https://alphafold.ebi.ac.uk) is a computational method that predicts the three-dimensional coordinates of atoms in a given protein utilizing the primary amino acid sequence and aligned sequences of homologues as inputs (Jumper et al., 2021). AlphaFold was chosen due to its ability to predict protein structure when no known or similar structure has been resolved, which is especially relevant when analyzing novel small proteins (Jumper et al., 2021). I-TASSER (http://zhang.bioinformatics.ku.edu/I-TASSER) uses replica exchange Monte Carlo simulations, which can account for the probability of different outcomes when random variables are present and, therefore, can provide function predictions from gene ontology by matching the structural models of the target protein to known proteins in the Protein Database (PDB). This is a meta-server threading approach, in which each threading program generates many different template alignments, but ultimately uses the templates of the highest significance (Yang et al., 2015). From I-TASSER, GO terms and consensus predictions for three ontologies were able to be obtained: biological processes, cellular components, and molecular function, and the latter was chosen to focus on, as the goal was to determine the potential function of the small protein.

To assess structure prediction accuracy, several metrics were considered. Confidence for AlphaFold predictions is based upon the predicted local distance difference test (pLDDT) to evaluate differences in the distances of all atoms in a model, which reliably predicts the Cα local-distance difference test (lDDT-Cα) and has been established to determine inaccuracy of the corresponding prediction (Jumper et al., 2021, Varadi et al., 2022). I-TASSER provides metrics consisting of a confidence score, which evaluates global and local similarity between the input

and template proteins. The range for the confidence score is -5 to 2, where higher values indicate more confident predictions. In addition, there is a TM score, which measures structural similarity between the input and template protein.

Finally, the molecular visualization tool PyMOL, which is capable of multi-structure alignment, was used to compare the generated structures in a qualitative manner. PDB files were loaded and aligned in the same orientation, so any discrepancies between the predicted structures from AlphaFold and I-TASSER would be visible.

Assays to Determine sORF Function

Inducing mutations in sORFs encoding sproteins may not result in obvious phenotypes. Commonly used biochemical assays used to identify a large protein's function employ methods that may not accurately report on a sprotein's function within the cell. In this study, assays were conducted to determine the cell's ability to carry out certain functions in the respective sprotein's absence or mutated form and subsequently compared to each respective wild type *M. smegmatis* phenotype. The assays were designed as follows.

Competitive Growth Curve Construction

To determine if the absence or mutation of each sORF would affect cell growth and proliferation in liquid co-culture, a competitive growth curve was constructed. A measurement of cell growth and proliferation in liquid culture was measured against a reference strain of *M. smegmatis* (MKD10), to inform whether the mutated sORF confers a growth advantage or growth disadvantage. This assay was conducted by starting overnight cultures of each knockout clone and a reference mycobacterial strain, normalized to a mutual starting optical density at 600 nm (measured as OD_{600}), and a co-culture of MKD10 and each knockout strain. The cultures were incubated with shaking for the duration of the eight-day assay, and each co-culture was

plated on selective media at regular intervals to determine the relative time at which each stage of growth occurred, whether it be lag, exponential, stationary, or cell death phase. The co-culture was re-expanded daily and was diluted and plated four times over the eight-day period, every two days. This assay can inform if the reference mycobacterial strain MKD10 outcompeted the knockout strains in terms of growth in liquid co-culture, or vice-versa.

Next-Generation Sequencing Barcoding

Based on experimental results from previous competitive co-culture growth, specific mutant strains containing a knockout of *msmeg1999* and overexpression of *msmeg1999* were found to have growth delays in liquid culture. To more accurately address these strain-specific growth differences, a pooled growth competition assay was developed to distinguish specific mutants under varying conditions. A total of eight mutant knockout strains (*msmeg0119, msmeg5154, msmeg1999, msmeg5128a, msmeg1916, msmeg0945, msmeg4499,* and *msmeg6401a)* were grown overnight and normalized to a mutual starting OD_{600} of 0.1 and combined in liquid co-culture. The pilot assay was designed to determine changes in the co-culture composition over time, therefore specific timepoints were identified in early and late growth, to measure the OD_{600} and isolate a co-culture sample. From the co-culture samples, genomic DNA was then isolated and used as a template in a multi-step PCR to amplify mutant-specific barcodes unique to each strain. Once these sequences were amplified, they are able to be sequenced together using Next Generation sequencing and the barcode used to determine changes in the amount of each strain in culture as co-culture growth progressed.

Fluorescent Microscopy of Dendra Mutants

sORFs were expressed as fusion protein constructs with Dendra and were visualized using fluorescent microscopy. To visualize single cells, wild-type and mutant derivatives were

cultured overnight in M9 minimal media and covered with low-temperature gelling agar to immobilize the cells for microscopy. These cells were then analyzed using a Nikon Eclipse Ti-E Inverted microscope to determine the potential localization of each protein within the cell and to examine any cell morphology changes due to the interference of the large protein fusion with cell growth and development.

Drug Susceptibility

Drug susceptibility assays were conducted to determine whether the sprotein affected basic cellular mechanisms. Pathways or functions can be determined by selecting antibiotics or reagents that disrupt specific cellular processes and measuring the effect each treatment has on each mutant clone compared with a wild-type control. Antibiotics can be used to target various cellular structures, pathways, and protein functions and thereby reveal mutations that compromise cell wall integrity or permeability or ribosome function. A top agar layer of the desired knockout cells or wild-type mc^2155 cells, from a prepared overnight culture, was poured over a tryptic soy agar plate containing and allowed to set. 5 µL of the antibiotic or reagent to be tested was dispensed onto a sterile filter paper disk and placed onto both the experimental and control plates in pre-measured sections. These plates were then incubated at 37°C for three days and the ring of inhibition for each treatment was measured and compared between each knockout and the wild-type control.

Antibiotics were grouped by their mechanism of action, falling into three broad categories: ribosome-targeting, cell wall biosynthesis, and DNA replication. 20 µg/mL chloramphenicol was chosen to inhibit the 50S ribosomal subunit, which prevents the transfer of amino acids to the growing peptide chains and blocks peptide bond formation. 200 µg/mL streptomycin was chosen to interfere with the initiation complex assembly between the mRNA

and the ribosome, which effectively inhibits the initiation of protein synthesis. 50 μg/mL capreomycin was utilized to inhibit protein synthesis by binding to the 70S ribosomal unit. To target cell wall biosynthesis and cell membrane permeability, 100 μg/mL vancomycin, and 50 μg/mL cycloserine were utilized. Finally, 0.2 μg/mL ciprofloxacin was chosen to inhibit DNA replication by blocking the action of both DNA gyrase and DNA topoisomerase.

Nonantibiotic reagents were also chosen to test if there was an inhibitory effect on cell growth. First, 0.5 M ethylenediaminetetraacetic acid (EDTA) was utilized as it sequesters divalent cations. 0.1% sodium dodecyl sulfate was used as an anionic detergent used to denature proteins, and 1% Triton X-100, a non-ionic surfactant, was chosen due to its ability to solubilize proteins. 0.1 M copper sulfate was utilized as it has been shown to have cell lysing capabilities, and 1% hydrogen peroxide was utilized as it has been shown to damage cell walls and induce cell death. Finally, 10 μg/mL ethidium bromide was tested, as it has been shown to intercalate DNA and effectively disrupt replication.

Biofilm Formation

Biofilm assays were conducted to determine if deleting or overexpressing each sORF affected the cell's ability to produce a pellicle biofilm, a characteristic of bacterial cells. Mycobacterial pellicle biofilms form at the interface between air and the liquid culture and require many functioning genes and involve changes in gene expression and community interactions. Therefore, if a biofilm could not be adequately produced or was altered phenotypically, it could indicate that the sprotein was involved in an aspect of biofilm formation. Previous studies have shown that sproteins can alter biofilm folding, a phenotypic result of modified protein interactions when a sORF is mutated (Judd et al., 2021). Under normal conditions, biofilms are formed within seven to ten days. This assay was conducted by seeding

complete biofilm media with cells from an overnight culture. The original optical density at 600 nm was 0.1 before being diluted 1:100, 1:1,000, and 1:10,000 in complete biofilm media. The dilutions were added to a six-well cell tissue culture plate and incubated at 30°C and the progression of biofilm formation was monitored on days 5, 7, 10, and 14. The cells containing each sORF mutation were compared to that of the wild type mc^2155 cells to determine any phenotypic differences.

RNA-sequencing

RNA-sequencing was chosen due to the capability to determine all the genome-wide transcriptional changes occurring in the cell as a result of either a knockout or overexpression of the sORF. For this reason, RNA-seq libraries were constructed for samples of the knockout *msmeg1999* strain and msmeg*1999* overexpression strain. Cell extracts were prepared in parallel to those used for qRT-PCR. A phenol-chloroform extraction and isopropanol precipitation were performed to extract RNA from total cell lysate. Ribosomal RNA was removed using the rRNA depletion kit (NEB). RNA fragmentation and library preparation were performed using the Ultra II Library Prep Kit (NEB) and prepared libraries were sent for Next-Generation Sequencing (Illumina). Results were analyzed to determine gene expression changes due to the absence or overexpression of gene *msmeg1999*.

qRT-PCR

To accurately quantify the relative amounts of *msmeg1999* transcript being produced in the various mutant constructs, qRT-PCR was performed on both the knockout *msmeg1999* strain and *msmeg1999* overexpression strain, as well as an independent control strain expressing *msmeg1999* at wild type levels. Primer pairs were designed to target the *msmeg1999* open reading frame and an independent gene in the genome to serve as a control. The relative

expression of *msmeg1999* in each of the samples was calculated based on the use of reference gene, rpoB, as an endogenous control.

Results

Summary of sORF Negative Phenotypes

Throughout the various assays conducted, including biofilm formation, drug susceptibility studies, fluorescent microscopy of dendra clones, and competition growth assays, many of the selected sORF mutant strains did not produce a reproducible phenotype that varied from the wild-type control. *Msmeg0119, msmeg5154, msmeg4499, msmeg6401a, msmeg5128a,* and *msmeg3439* all produced phenotypes identical to the wild type controls utilized in each of the assays conducted to determine a function associated with the mutant sORF. Images taken using fluorescent microscopy of each of the sORF:Dendra protein fusions did not identify any distinct foci or localization patterns for the fusion proteins. Instead, the signal was diffuse and cytoplasmic, similar to the control expressing just Dendra. For drug susceptibility assays conducted with these sORF mutants, the rings of inhibition that formed after incubation did not vary from the rings of inhibition exhibited by the wild-type control; and successive platings during the competition growth assay had the same number of colonies on the control strain plate as each of the sORF mutant strains.

However, while conducting these assays, both the knockout *msmeg1999* strain and the overexpression *msmeg1999* strain exhibited altered growth phenotypes. Growth in liquid culture was often delayed when compared to both wild-type *M. smegmatis* strains and other knockout and overexpression constructs. To better characterize this phenotype, several assays were developed and performed repeatedly, as described below.

Structural Analysis

Table 1. **Summary of molecular function GO terms for the top five structural homologs to the target used in template-based modeling.**

Gene Annotation	**PDB IDs of Top 5 Homologous GO Templates**	**Associated Molecular Function GO Terms**
msmeg0119	3ljbA, 3mliC, 2a5yA, 2xggB, 1aj8A	Protein binding, transferase activity
msmeg5154	2oz6A, 1zvrA, 3f41A, 2vhdA, 3isrB	Nucleotide binding, phosphatase activity, peptidase activity
msmeg1999	1dowA, 1u3cA, 2i0mA, 1yuxB, 1c02A	Structural molecule activity, DNA photolyase activity, phosphorelay signal transduction system
msmeg4499	1c0wD, 1c0wB, 2a9fA, 3kkcA, 2xmxA	DNA binding, metal ion binding
msmeg0945	3c9iB, 3hpcX, 3ojaA, 2gyqA, 3bj4B	Protein binding, metal ion binding

As seen in Table 1, the associated molecular function gene ontology (GO) terms are listed for each of the five gene annotation inputs (*msmeg0119, msmeg5154, msmeg1999, msmeg4499,* and *msmeg0945*). The molecular function GO terms were determined from the top five identified structural homologs used in template-based molecular modeling. From these associated GO terms, the gene product of most inputs were predicted to be involved in binding activity, whether protein, nucleotide, metal binding, or DNA binding. Secondary GO terms were in the general class of transport activity.

Characterization of msmeg1999

Msmeg1999 is a sORF that encodes a sprotein of 47 amino acids and is conserved in other mycobacterial species. A knockout of *msmeg1999* grew normally on solid medium, however it exhibited an altered growth phenotype, often taking longer for the liquid culture to grow to a comparable optical density of a wild-type strain. This effect was also seen in the strain

that overexpressed *msmeg1999*. Based on preliminary data that demonstrated this phenotypic effect during standard growth in liquid media, assays were developed to further understand the effects of altered levels of *msmeg1999* in the cell. Additionally, *msmeg1999* was chosen for study due to results examined with the Dendra fusion protein, including altered cell morphology resulting in bent and angled cells and an associated slow-growth phenotype.

To better characterize the delayed growth phenotype in relation to other mycobacterial strains, a competition growth assay was developed over an eight-day period with daily re-expansions and culture platings every two days (Figure 3). Colony forming units (CFUs) were counted for both the reference strain MKD10 (plated on TSA plates containing kanamycin) and a knockout of *msmeg1999* (plated on TSA plates containing zeocin) after identical dilutions of the original culture. The reference mycobacterial strain MKD10 outcompeted the *msmeg1999,* mutant evident in the higher colony numbers recorded for MKD10 throughout the course of the assay, with the most notable difference at day 6. The mutant strain appears to struggle initially in growth and proliferation in liquid culture, as the colony number is very low for the first two days in comparison to the reference strain. After these initial two days of very low growth, the mutant strain then starts to grow at the same rate as wild-type reference strain, apparent with similar growth rates between days 4 and 6, only with the mutant strain having lower total colony counts. After this point, any of the liquid-culture that was plated out and incubated did not form colonies, and it was likely that after this time point that most of both the knockout strain and wild-type cells had died.

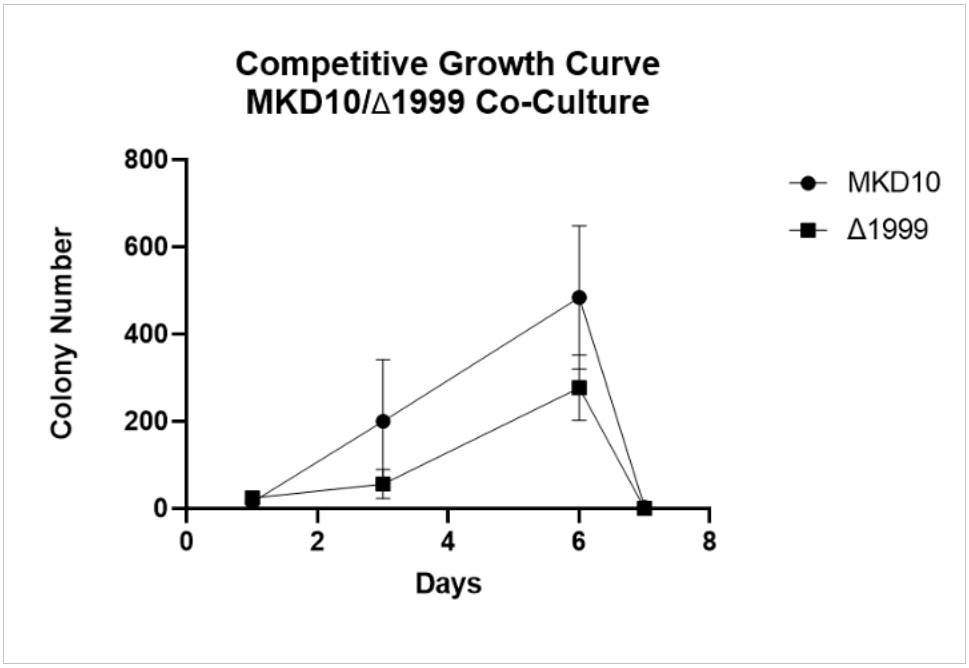

Figure 3. **Competitive growth co-culture data for a mycobacterial strain containing a gene knockout of *msmeg1999*** against a wild-type reference mycobacterial strain, MKD10, over a span of seven days.

To study the altered growth phenotype of the overexpression strain present in liquid culture, monoculture growth in liquid culture was repeated to construct growth curves over a 25-hour period. After 18 hours of growth, the strain overexpressing *msmeg1999* did not grow to the same cell density as the wild-type strain (Figure 4). Over the course of the assay, the overexpression strain continued to grow at a slower rate and did not reach stationary phase during this observed time period.

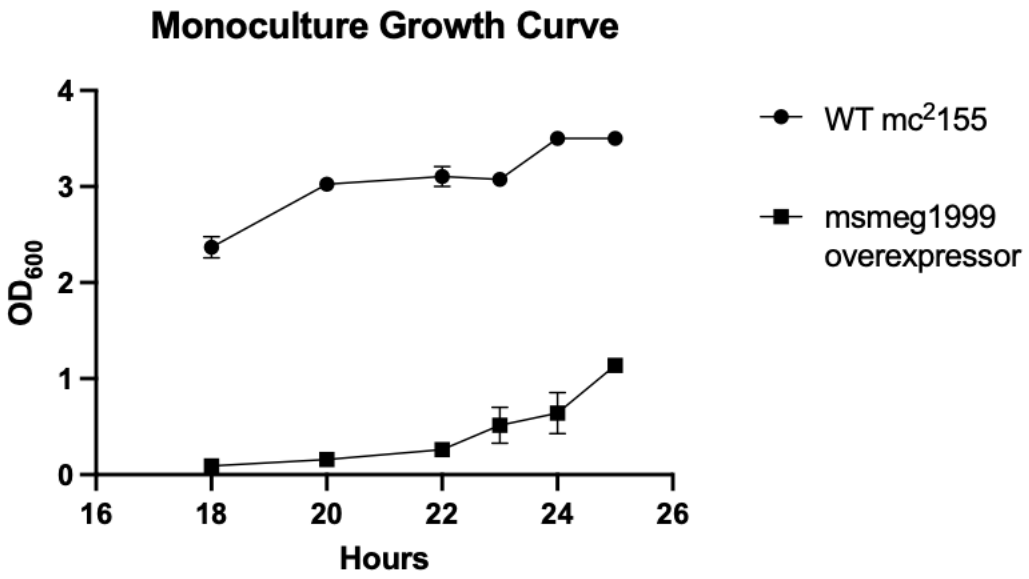

Figure 4. **Monoculture growth curves of wild-type mc²155 and *msmeg1999* overexpression strain over the course of 25 hours**. The growth of both samples, measured by optical density (OD_{600}) in liquid culture (TSB + 0.05% tween80) were sampled on an hourly basis to determine differences in doubling time.

To better understand the phenotypic growth effect from a transcriptional point of view and determine differential gene expression in knockout and overexpression *msmeg1999* strains, RNA-Seq libraries were constructed. Analysis of these data were expressed as measurements of log2fold change in gene expression, when gene transcripts produced in the knockout strain were compared to gene transcripts produced in the overexpression strain. From these data, genes with a log2fold change of greater than 5 were considered significant for further study, as many genes had little to no measured change in expression. Seen in Figure 5, 18 genes were upregulated by greater than a log2fold change of five when *msmeg1999* expression was mutated. The gene with the most significant expression change was *msmeg4442,* which has been characterized as a zinc-binding oxidoreductase. When expression of *msmeg1999* was mutated, 13 genes were found to have a significant downregulation, the most significant of which was *msmeg5336*, which has

been identified as an amidate substrates transporter protein, which falls under the general class of proteins involved with transmembrane substrate transportation.

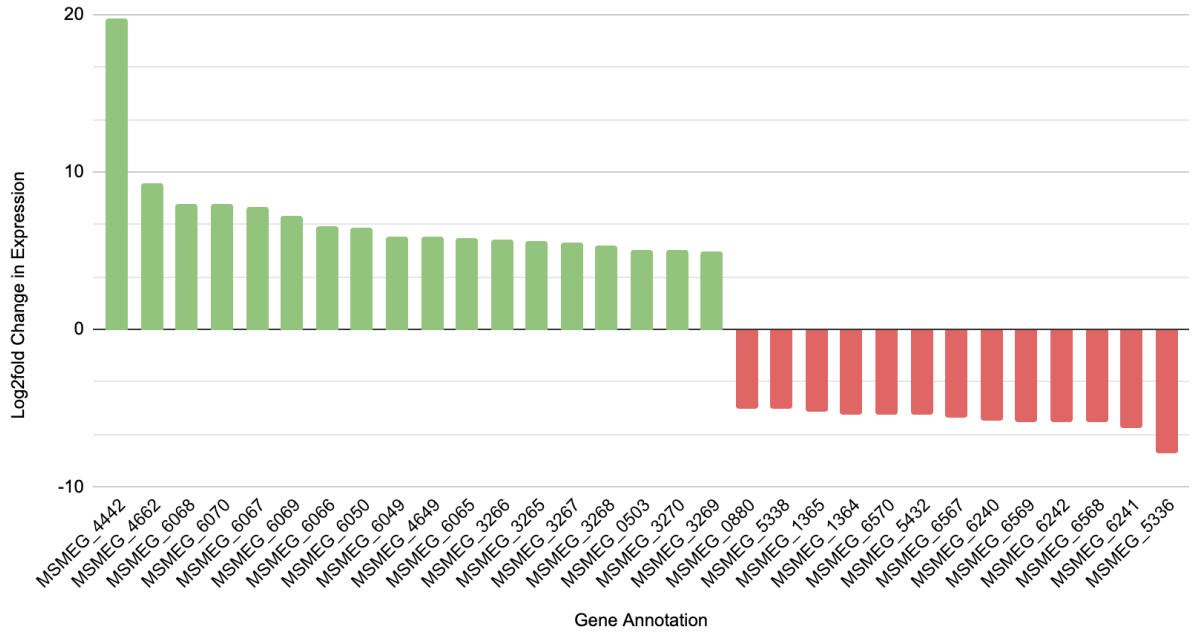

Figure 5. **Log2fold change in expression of 31 differentially regulated genes identified by RNA-Seq.** Upregulation of genes, relative to wild-type expression, is depicted by green bars, where subsequent downregulation of genes relative to wild-type expression is denoted by a red bar. Log2fold change was determined using the transcript per million counts for each gene in both the knockout and overexpressed *msmeg1999* constructs.

To quantify expression of *msmeg1999* in the strains for which RNA-Seq libraries were constructed, qRT-PCR was performed on the same RNA purified for the RNA-seq library construction. Seen in Figure 6, the control strain showed an average delta-delta Ct of 1, which was expected as the expression of *msmeg1999* was wild type. The measured mRNA transcripts expression of *msmeg1999* in both of the mutant strains was found to be lower than the wild-type control, with average delta-delta Ct values of 6.71 and 12.81, respectively. As expected, mRNA for *msmeg1999* was undetectable for the deletion strain, which lacks a functional copy of the gene. Surprisingly, 1999 mRNA for the overexpression strain was also considerably less than the wild-type strain.

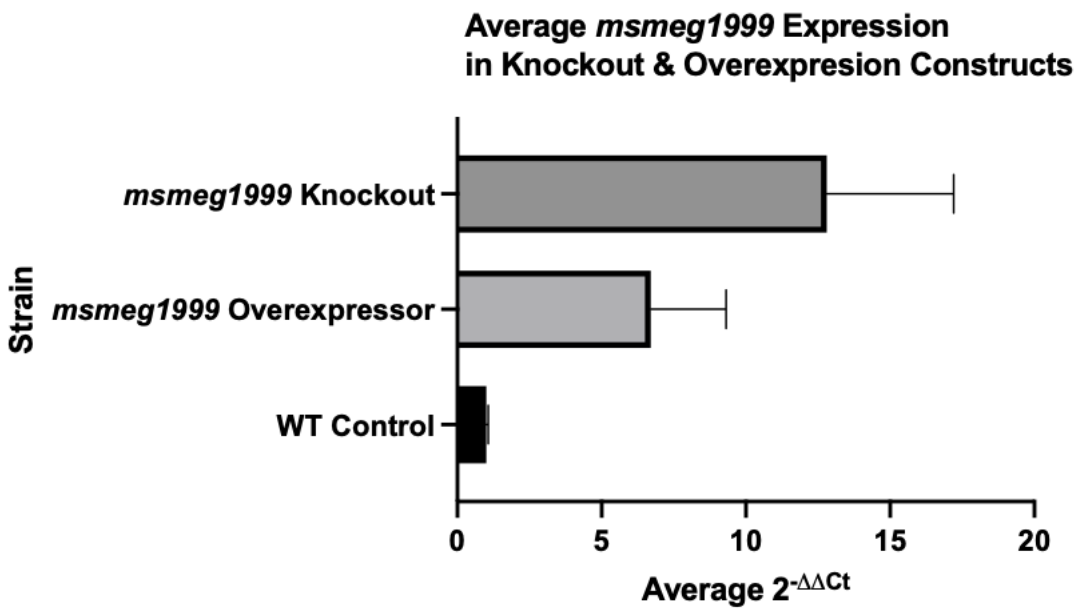

Figure 6. **Average expression of *msmeg1999* in knockout and overexpression strains measured by qRT-PCR**. Expression levels were analyzed by calculated the average delta-delta Ct, in which Ct levels are inversely proportional to the amount of target nucleic acid in the sample, indicating that lower Ct values correspond to greater amounts of the gene of interest measured in the sample.

Figure 7 shows that expression of the msmeg1999:Dendra fusion caused notable changes in cell morphology. Some cells appeared to be bent and somewhat angular as opposed to straight rod-shaped cells, characteristic of *M. smegmatis*. Not all cells pictured appeared to be bent or misshapen, some remained straight and rod-shaped similar to the wild type control.

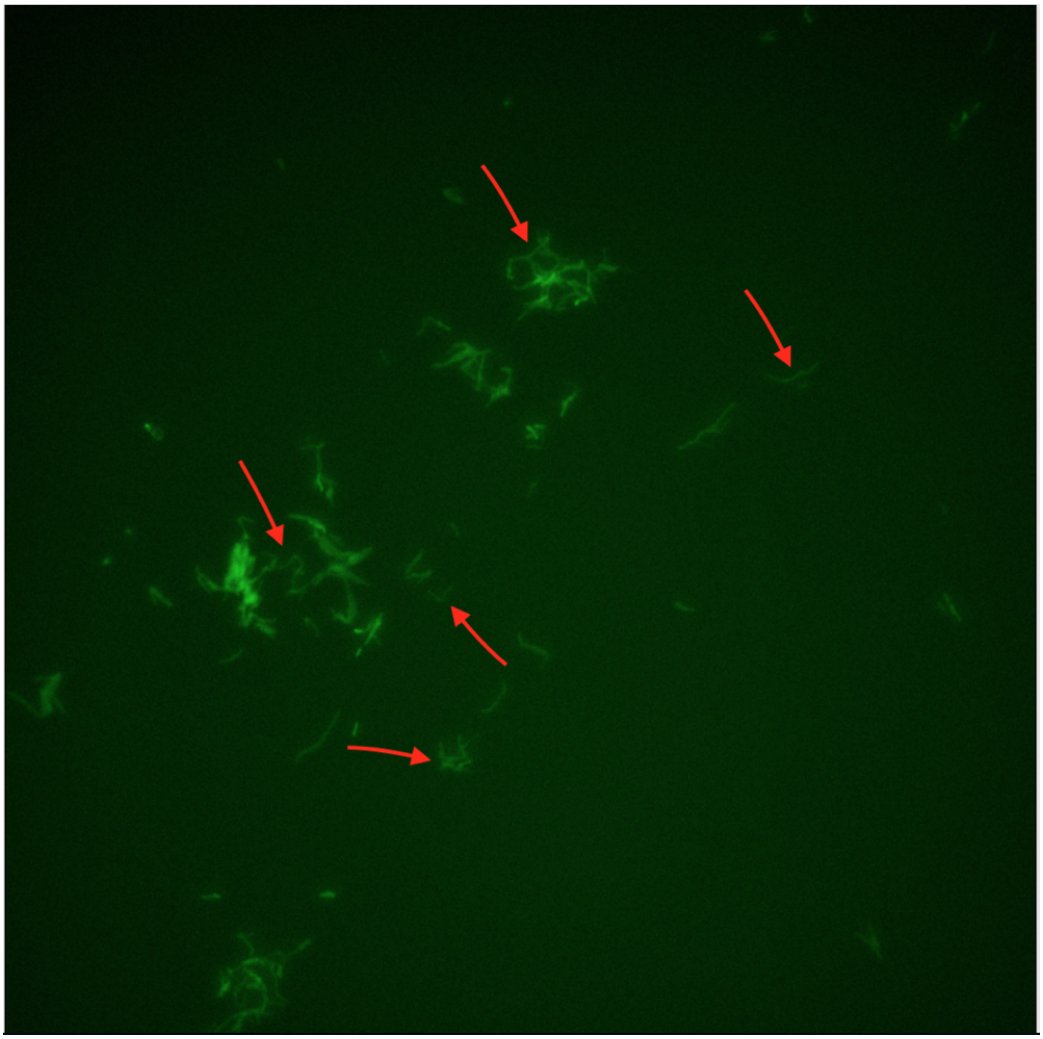

Figure 7. **Fluorescence microscopy images of *M. smegmatis* expressing a msmeg1999:Dendra fusion protein.** The red arrows denote specific cells for which a distinct change in cell shape is evident.

Discussion

After developing a series of *M. smegmatis* strains with various sORF mutations and constructing Dendra gene fusions to each sORF, assays were developed to identify novel phenotypes associated with sORF deletion or overexpression. The aim of developing such assays was to associate sORF function with an observable phenotype. The preliminary results and summary of negative phenotypes demonstrate that the current assays should be further optimized

to determine if mutant sORFs cause phenotypic results that are reproducible. The completed biofilm assays showed no distinct difference between wild-type mc^2155 and any of the selected gene-knockouts. This suggests that the absence of any of these functioning genes does not affect biofilm growth under these conditions. Drug susceptibility assays that were conducted with each of the constructed gene knockouts did not display any notable differences to the wild-type control. Various reagents and antibiotic treatments did have an inhibitory effect on *M. smegmatis* cell growth, however, the mutations in the sORFs did not alter the sensitivity, suggesting that the sproteins do not alter any of the pathways that these reagents and antibiotics target. In addition, knockout strains of *msmeg0119, msmeg5154, msmeg4499, msmeg6401a, msmeg5128a,* and *msmeg3439* were not outcompeted in liquid co-culture with the reference mycobacterial strain MKD10, which suggests that functionally mutating the production of these sproteins does not confer a growth disadvantage to the cell. Not all sORFs and synthesized sproteins are expected to have a phenotype, however further development of these and other functional assays could assist in elucidating their function, especially investigating various growth conditions that could potentially induce stress on the cell.

The goal of the sequence-based and structure-based analysis was to predict the function of the small proteins within the cell and gain insight into the relationship that exists between structure and function. The molecular function ontology was thought to be the most relevant in determining a potential phenotype for the synthesized sproteins, but it would be worthwhile to examine both the cellular and biological processes ontologies to further characterize these sproteins. Molecular function terms describe activities that occur at the molecular level, but do not specify where, when, or in what context the action takes place. Generally, molecular functions correspond to activities that can be performed by individual gene products but some of

these identified activities could be performed by molecular complexes composed of multiple gene products. An important conclusion from these preliminary studies of the sprotein structure-function relationship is that several of the molecular terms were repeated. This repetition of terms could support the notion that sproteins as a general class of proteins usually participate in these activities. Protein binding and metal ion binding were terms for several of the templates, indicating that these sproteins may participate in cellular processes as components of a larger protein complex. For those with metal ion binding identified as a GO term, further studies should be identified to determine potential sites within the protein that could serve as binding sites.

One mutant that consistently showed a phenotype was *msmeg1999*. As seen in Figure 3, the proliferation of MKD10 outpaced the *msmeg1999* knockout strain growth in liquid co-culture lacking antibiotics, with the largest difference in growth rate evident at day 6. The two individual cultures used to start the liquid co-culture began at identical optical densities at 600 nm, however the *msmeg1999* knockout strain had consistently lower density readings from overnight cultures and cell counts. A similar delay was also observed in monoculture (Figure 4). This growth delay both in monoculture and in co-culture with another strain of mycobacteria suggests that the absence of this functioning sORF or altered levels of Msmeg1999 in the cell confer a growth disadvantage in liquid culture, which can be attributed to several potential functions of the sORF within the cell. Furthermore, examination of the slow-growth phenotype of cells overexpressing *msmeg1999*, further suggests that levels of Msmeg1999 in the cell must be maintained within a narrow range to not inhibit growth and proliferation. When msmeg1999 was overexpressed, overnight cultures often failed to reach an OD600 of 0.1 and therefore could not be used for additional assays. When the msmeg1999 overexpressor was included in a competitive growth

curve with MKD10, the growth of MKD10 far outpaced the rate of the msmeg1999 overexpression strain.

While conducting several of the assays, the sORF knockout and overexpression clone of *msmeg1999* produced notable growth differences within liquid culture. Cultures of the *msmeg1999* knockout strain exhibited an extended lag phase of growth compared to the wild-type strain, and a longer delay before optical density measurement started to increase when cultures were started at the same cell density. In addition to the slow-growth phenotype in the absence of *msmeg1999* in competitive coculture (Figure 3) there was a noticeable growth delay in mutant strains overexpressing *msmeg1999,* seen in Figure 4. When considered together, these results suggest a role for msmeg1999 within the mycobacterial cell for proliferation in liquid culture, as both mutated derivatives have a growth lag when compared to wild-type strains. When this slow-growth phenotype was further investigated by developing a competitive growth assay under varying conditions such as elevated temperature and more minimal liquid medium composition, these factors did not contribute to any changes in growth delay, rather the growth delay was identical to that occurring at standard incubation temperature (37°C) and standard media (TSB + 0.05% Tween80). To address the growth delay, a construct with an inducible tetracycline repressor was designed to allow conditional overexpression of the protein using anhydrous tetracycline to induce expression. Assays with this inducible overexpression system were relatively inconsistent, likely because the exact concentration of anhydrous tetracycline required to reproducibly induce de-repression of the system was not effectively established. Once concentrations required for robust induction have been determined, additional assays will be conducted to determine and characterize the effects of msmeg1999 overexpression in a controlled fashion. Previous data suggests that the growth delay is most evident at low cell

densities early on in liquid culture growth (Figure 4), which will need to be taken into account when developing these assays.

To better determine transcriptional changes due to *msmeg1999* mutation, RNA-seq libraries were prepared for the knockout and overexpression strains. RNA-seq data provides a broad view of the global expression changes that result from a mutated form of the synthesized small protein in the cell. Subsequent bioinformatic analysis of the RNA-seq data provides a transcriptional profile of strains lacking a functional Msmeg1999 protein or overexpressing this protein. From these analyses genes that are significantly upregulated or downregulated in response to *msmeg1999* absence or overexpression are identified and can be used to inform on the role of 1999 in the cell. By understanding the transcriptional profile of these strains, more targeted assays can be developed to understand the function of the sORFs within the cell based on what genes might be upregulated and downregulated when expression is altered. For example, Figure 5 shows that *msmeg4442* is highly upregulated in response to altered *msmeg1999* expression, and this gene has been identified as a zinc-binding oxidoreductase. Using the information that this gene is highly upregulated, assays including growth curves and biofilm formation assays in the presence and absence of zinc could provide a greater understanding of the role of *msmeg1999* in the cell.

RNA-sequencing data helped to clarify transcriptional changes occurring in the cell as a result of mutated *msmeg1999* expression. The most significantly upregulated gene was determined to be *msmeg4442*, which has been identified as a zinc-binding oxidoreductase. This gene participates in catalyzing the transfer of electrons during reactions involving zinc ions, which are required for DNA repair, various enzymatic reactions, and in responses to oxidative stress in bacteria. Significant upregulation of this gene in particular suggests that msmeg1999

could be involved in one of these fundamental processes. Several of the upregulated genes, including *msmeg6068, msmeg6070, msmeg6069* and *msmeg6066*, are identified as ribosomal proteins. These gene products comprise the mycobacterial ribosome, and upregulation of these genes indicates that mis-expression altered bio availability of msmeg1999 could have an impact on ribosomal functions in the cell. In mycobacteria, zinc depletion has been studied in relation to ribosome hibernation, (Li et al., 2018) and some zinc-free ribosomal proteins are induced in low zinc growth conditions. In terms of the downregulation of genes in response to altered *msmeg1999* expression, many of the identified genes were not differentially expressed at levels as high as the identified upregulated genes. These include identified regulatory proteins or membrane proteins, including transmembrane substrate transporter proteins. An interesting correlation, and potential validation of the analyzed data, is the differential expression of entire operons in the same manner, for example the ribosomal genes mentioned earlier. Both significantly upregulated and downregulated genes identified by RNA-seq analysis contained several genes within that were transcribed as a single operon. This serves as a potential validation of the data, as regulation of the expression of these genes is controlled by a single promoter.

To quantify the changes in *msmeg1999* expression in each of the strains analyzed by RNA-Seq, qRT-PCR data was analyzed. Interestingly, repeated experiments demonstrated that the expression of *msmeg1999* in the overexpression construct was lower than those in the wild-type control (Figure 6). This unexpected result could stem from errors in RNA extraction or cDNA preparation during the assay. Another explanation for this result could be that the cell represses msmeg1999 overexpression. To understand this potential need for the cell to tightly regulate and maintain wild-type expression of *msmeg1999*, the plasmid driving overexpression

of *msmeg1999* was introduced into the *msmeg1999* knockout strain for complementation studies. Using this strain, studies of growth and proliferation can take place to determine if restoring *msmeg1999* expression in a knockout background allows for wild-type growth.

In preliminary studies, no defined localization patterns were detected with the sORF Dendra fusion proteins. The lack of any localized fluorescence during microscopy can be for several reasons such as the sprotein is localized in the cytoplasm or the fusion is not stable or poorly expressed which would make the detection of fluorescence foci difficult. In addition, the larger dendra protein may have rendered the sprotein non-functional and, therefore, not appropriately localized. This may have been the case for the Msmeg1999:Dendra fusion, seen in Figure 6, as the cells appeared to be somewhat bent and angular. This cell morphology change could indicate that the significantly larger Dendra protein fusion, which exceeds 500 amino acids, hindered the ability of the sprotein to carry out its function within the cell, causing the cells to become distorted. The change in cell shape could have resulted from a lack of protein localization or functional interference causing a change in cell division or shape, as this strain did grow slightly slower and with more clumping in liquid culture. Considering the change in cell shape, this result indicates that *msmeg1999* may play a role in cell wall biosynthesis or cell division, hence the delayed growth, or general structure, and the lack of this functional sprotein impacts the normal development and/or division of the characteristic bacillus shape.

Conclusion & Future Direction

In conclusion, many of the sORFs chosen for mutation did not exhibit an observable or reproducible phenotype based on these preliminary assays. To further validate and expand on the altered growth phenotype observed in previously completed assays, independently repeated studies using biological replicates of both the *msmeg1999* knockout and overexpression

construct are necessary. By repeating assays using independently created clones, the results will validate existing data and create a higher level of confidence when determining the function of *msmeg1999* in *M. smegmatis*.

To better associate sORF mutations or overexpression with a reproducible phenotype, future assays should be optimized based on the preliminary results explained here. Future experiments should develop a functional protocol for the barcoding growth curves for liquid co-culture, which functions to accurately address these strain-specific growth differences in a pooled growth competition assay and to determine changes in the co-culture composition over time. From the co-culture samples, genomic DNA can be isolated and used as a template in a multi-step PCR to amplify mutant-specific barcodes unique to each strain and once sequenced together, the barcode used to determine changes in the amount of each strain in culture as co-culture growth progressed. Therefore, a fully developed barcoding assay can then be used to determine if the mutation or deletion of each sORF impacts cell proliferation in liquid culture. This approach would allow for a higher throughput, and more accurate and quantifiable assessment of differences observed in co-culture. By quantifying these changes using NGS reads, a change in the amounts of each strain in co-culture can be accurately quantified, as opposed to counting plated colonies, which requires differing antibiotic resistance conferred to each strain. The data obtained from this assay would also allow for expansion to include many strains as each strain has a defined and distinguishing barcode, its deletion.

Based upon the preliminary RNA-Seq results, examining growth rates and biofilm development in the presence and absence of zinc could be used to explore the potential relationship between altered production of *msmeg1999* and upregulation of genes involved with zinc regulation in the cell. Previous studies have suggested zinc plays an important role in

ribosome hibernation and remodeling processes, which may be relevant in examining the differential gene expression of zinc-associated genes seen here, especially those associated with the bacterial ribosome (Li et al., 2018). In addition, examining the morphology of the msemg1999 overexpression strain using higher resolution microscopy could provide insight into changes in cell wall structure or shape, and allow further characterization of the cell morphotype observed (Figure 7). However, it might be useful to switch to a smaller tag as it cannot be ruled out that the rather large Dendra tag distorts the cells and prevents sORFs from functioning. In relation to the structure-function interpretation, it would be worthwhile to perform gene ontology enrichment for associated terms from the I-TASSER predictions, as a means to create a network analysis from these results. In addition, I-TASSER provides predictions for active binding sites, which may be worthwhile to explore when designing enzymatic assays. The vast availability of information resulting from bioinformatic analysis of mutant sORF strains has the potential to inform targeted assays that more accurately address sprotein function.

Potential sORFs continue to be identified in the same manner as the sORFs previously described, which may result in the identification of additional, functional sORFs in less studied bacteria. The nature of sORF identification and the extended period of time in which they were excluded from study resulted in a lack of appreciation for the roles sproteins play within bacterial cells and eukaryotes. With increased recognition and study, functionally characterizing sproteins can reveal novel regulation mechanisms, components of the cell membrane, and identify new mechanisms of action for therapeutic targets (Gray et al., 2022). However, altered approaches to determine sORF functions may be required, including expression of tagged sORFs followed by purification and mass spectrometry analysis to determine interacting proteins would identify larger complexes that could subsequently be probed to determine sprotein function.

References

Tuberculosis (TB). Centers for Disease Control and Prevention (2023). Retrieved from https://www.cdc.gov/tb/topic/research.htm.

Degiacomi, G., Sammartino, J. C., Chiarelli, L. R., Riabova, O., Makarov, V., & Pasca, M. R. (2019). Mycobacterium abscessus, an Emerging and Worrisome Pathogen among Cystic Fibrosis Patients. International journal of molecular sciences, 20(23), 5868. https://doi.org/10.3390/ijms20235868

Sparks, I. L., Derbyshire, K. M., Jacobs, W. R., Jr, & Morita, Y. S. (2023). Mycobacterium smegmatis: The Vanguard of Mycobacterial Research. *Journal of bacteriology*, *205*(1), e0033722. https://doi.org/10.1128/jb.00337-22

Miravet-Verde, S., Ferrar, T., Espadas-García, G., Mazzolini, R., Gharrab, A., Sabido, E., Serrano, L., & Lluch-Senar, M. (2019). Unraveling the Hidden Universe of Small Proteins in Bacterial Genomes. *Molecular Systems Biology*, 15(2), e8290. https://doi.org/10.15252/msb.20188290

Hemm, M. R., Paul, B. J., Miranda-Ríos, J., Zhang, A., Soltanzad, N., & Storz, G. (2010). Small stress response proteins in Escherichia coli: proteins missed by classical proteomic studies. *Journal of bacteriology*, 192(1), 46–58. https://doi.org/10.1128/JB.00872-09

Smith, C., Canestrari, J. G., Wang, A. J., Champion, M. M., Derbyshire, K. M., Gray, T. A., & Wade, J. T. (2022). Pervasive translation in Mycobacterium tuberculosis. *eLife*, 11, e73980. Advance online publication. https://doi.org/10.7554/eLife.73980

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

Andrews, S. J., & Rothnagel, J. A. (2014). Emerging Evidence for Functional Peptides Encoded by Short Open Reading Frames. *Nature Reviews Genetics*, 15(3), 193–204. https://doi.org/10.1038/nrg3520

Duval, M., & Cossart, P. (2017). Small bacterial and phagic proteins: an updated view on a rapidly moving field. *Current opinion in microbiology*, 39, 81–88. https://doi.org/10.1016/j.mib.2017.09.010

Orr, M. W., Mao, Y., Storz, G., & Qian, S. B. (2020). Alternative ORFs and Small ORFs: Shedding Light on the Dark Proteome. *Nucleic Acids Research,* 48(3), 1029–1042. https://doi.org/10.1093/nar/gkz734

Prelich G. (2012). Gene overexpression: uses, mechanisms, and interpretation. Genetics, 190(3), 841–854. https://doi.org/10.1534/genetics.111.136911

Rock, J. M., Hopkins, F. F., Chavez, A., Diallo, M., Chase, M. R., Gerrick, E. R., Pritchard, J. R., Church, G. M., Rubin, E. J., Sassetti, C. M., Schnappinger, D., & Fortune, S. M. (2017). Programmable transcriptional repression in mycobacteria using an orthogonal CRISPR interference platform. *Nature microbiology*, *2*, 16274. https://doi.org/10.1038/nmicrobiol.2016.274

Larson, M. H., Gilbert, L. A., Wang, X., Lim, W. A., Weissman, J. S., & Qi, L. S. (2013). CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature protocols*, *8*(11), 2180–2196. https://doi.org/10.1038/nprot.2013.132

Basic Local Alignment Search Tool [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2023 March 1]. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Yang, J., & Zhang, Y. (2015). I-TASSER server: new development for protein structure and function predictions. *Nucleic acids research*, *43*(W1), W174–W181. https://doi.org/10.1093/nar/gkv342

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., … Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, *50*(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061

Judd, J. A., Canestrari, J., Clark, R., Joseph, A., Lapierre, P., Lasek-Nesselquist, E., Mir, M., Palumbo, M., Smith, C., Stone, M., Upadhyay, A., Wirth, S. E., Dedrick, R. M., Meier, C. G., Russell, D. A., Dills, A., Dove, E., Kester, J., Wolf, I. D., Zhu, J., Derbyshire, K. M. (2021). A Mycobacterial Systems Resource for the Research Community. *mBio*, *12*(2), e02401-20. https://doi.org/10.1128/mBio.02401-20

Li, Y., Sharma, M. R., Koripella, R. K., Yang, Y., Kaushal, P. S., Lin, Q., Wade, J. T., Gray, T.

A., Derbyshire, K. M., Agrawal, R. K., & Ojha, A. K. (2018). Zinc depletion induces ribosome hibernation in mycobacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(32), 8191–8196. https://doi.org/10.1073/pnas.1804555115

Gray, T., Storz, G., & Papenfort, K. (2022). Small Proteins; Big Questions. *Journal of bacteriology*, *204*(1), e0034121. https://doi.org/10.1128/JB.00341-21