# The Effects of mRNA Abundance on Degradation Rate

A Major Qualifying Project

Submitted to the Faculty of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Bachelor of Science In Biology & Biotechnology

and

Degree of Bachelor of Science In Bioinformatics & Computational Biology


Authored by:

Adrian Orszulak

and

Van Le


April 28th, 2022

Approved by:

Scarlet S. Shell, PhD

Lane T. Harrison, PhD

# Part I

Assessing the Effects of mRNA Abundance on mRNA Degradation Rates in *Mycolicibacterium smegmatis*

# Assessing the Effects of mRNA Abundance on mRNA Degradation Rates in *Mycolicibacterium smegmatis*

A Major Qualifying Project

Submitted to the Faculty of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Bachelor of Science

In

Biology & Biotechnology

Authored by:

Adrian Orszulak

and

Van Le

April 28th, 2022

Approved by:

Scarlet S. Shell, PhD

*This report represents the work of one or more WPI undergraduate students submitted to the faculty as evidence for completion of a degree requirement. WPI routinely publishes these reports on the web without editorial or peer review.*

## Abstract

*Mycobacterium tuberculosis*, the etiological agent of tuberculosis, is a difficult pathogen to treat, requiring a lengthy treatment course with numerous antibiotics. It is believed that a robust regulation of gene expression contributes to high tolerance to antibiotics and other stressors. mRNA concentration is one physical factor that may impact mRNA half-life, a contributing factor to overall gene expression. Previous work in *M. tuberculosis* and other bacteria indicates a lack of consensus regarding whether mRNA abundance and mRNA half-life show a strong, negative correlation or a weak, positive correlation. Additionally, mRNA abundance may impact protein abundance in a non-linear fashion. However, there is a lack of consensus regarding the relationship between mRNA abundance and protein abundance. By understanding the impact of mRNA abundance on regulating gene expression, we sought to gain a greater understanding of how *M. tuberculosis* is able to effectively respond to stress. Using a tetracycline-inducible gene expression system in the model organism *Mycolicibacterium smegmatis*, we tested various combinations of concentrations of anhydrotetracycline (aTc) and induction times to determine conditions that would provide the widest range of mRNA and protein expression levels. We established that a range of aTc concentrations from 0 ng/mL to 50 ng/mL at a 4-hour induction time provided a wide range of *gfpmut3* expression. The degradation data were too noisy to determine half-life and make meaningful conclusions regarding the relationship between mRNA abundance and mRNA half-life. Additionally, our system could not be used to investigate the relationship between mRNA abundance and protein abundance due to a loss of inducer-based expression for undetermined reasons.

## Introduction

Tuberculosis (TB) is an upper respiratory infection that is a major cause of ill health globally (World Health Organization, 2020). The etiological agent for this disease is the pathogenic bacteria *Mycobacterium tuberculosis* (World Health Organization, 2020). In 2019 alone, 1.4 million deaths were attributed to *M. tuberculosis*, making it one of the leading causes of death by an infectious disease in the world (World Health Organization, 2020). While the number of cases in high-income countries has decreased significantly since the 1940s, *M. tuberculosis* is still a major problem in medium and low-income nations (CDC, 1990; World Health Organization, 2020). Therapies for TB are combinatorial and lengthy, incurring a physical, mental, and economic toll on the individual patient (World Health Organization, 2020). One reason for the lengthy course of treatment is that *M. tuberculosis* inside granulomas is tolerant to antibiotics and a number of stresses, such as hypoxia, nutrient starvation, low pH, and reactive oxygen species (ROS). The mechanisms of tolerance in *M. tuberculosis* stem from a rigorous and well-evolved regulation of gene expression that allows them to survive in such conditions (Reviewed in: Connolly et al., 2007; Reviewed in: Prax, M., & Bertram, R, 2014; Reviewed in: Boldrin et al., 2020).

Within stressful environments, *M. tuberculosis* is able to regulate its gene expression to adapt and persist. Regulation of the gene expression profile in *M. tuberculosis*, like any other bacteria, can occur at a select number of points: transcription of a gene, degradation of mRNA, translation of mRNA into protein, and degradation of the protein (Hausser et al., 2019). The degradation of mRNA is of key interest in studying the regulation of gene expression given the unstable nature of mRNA and the high energy cost of mRNA and protein synthesis within the cell (Pato et al., 1973; Reviewed in: Russel & Cook, 1995; Stouthamer, 1979; Dressaire et al., 2013). Additionally, studies examining the impacts of bacterial transcriptional and posttranscriptional regulatory mechanisms on the half-lives of mRNA in *M. tuberculosis* and *Mycolicibacterium smegmatis*, a model organism for *M. tuberculosis*, show that mRNA half-lives are extended in response to stress (Rustad et al., 2013; Vargas-Blanco et al., 2019). Understanding the mechanisms through which mRNA levels and degradation rates are regulated in response to resource and energy stress will yield a greater knowledge base of information regarding how *M. tuberculosis* is able to effectively respond to stress.

A number of features of mRNAs have been assessed to understand the mechanisms for regulating mRNA degradation in bacteria. These features include stem-loops (Emory & Belasco, 1992), leadered or leaderless gene transcripts (Chen et al., 1991; Nouaille et al., 2017; Nguyen et al., 2020), interaction with regulatory proteins and sRNAs (Arnvig & Young, 2012; Chen et al., 2015; Sinha et al., 2018), RNA-binding proteins (e.g., CsrA, Hfq) (Liu et al., 2010; Timmermans et al., 2010), poly-A tails (O'Hara et al., 1995), and codon content (Lenz, et al., 2011; Boël et al., 2016). In addition to these structural and mechanistic features, many studies have investigated the association between mRNA half-life and mRNA concentration. Bernstein et al. (2002) used DNA microarrays to identify an inverse relationship between mRNA half-lives and mRNA abundance in *E. coli*. Another study in *E.coli showed* a negative association between mRNA

half-lives and mRNA concentration (Esquerré et al., 2015). Using reporter systems with either arabinose-inducible or nisin-inducible promoters in log-phase, Nouaille et al. (2017) observed an inverse correlation between the half-life and concentration of *lacZ* mRNA in *E. coli* as well as *lacLM* mRNA in *Lactococcus lactis*. An inverse correlation between mRNA half-life and mRNA concentration in log-phase *M. tuberculosis* was shown by Rustad et al. (2013). A much weaker inverse correlation between mRNA half-life and mRNA concentration was observed in *M. smegmatis* (Sun et al., manuscript in preparation).

In contrast to studies that showed negative relationships, there were two studies that indicated a weakly positive correlation between mRNA half-life and mRNA abundance in *B. cereus* and *E. coli* through transcriptome-wide measurement of mRNA half-lives using RNA-seq (Kristoffersen et al., 2012; Chen et al., 2015). In a study conducted by Redon et al. (2005), *L. lactis* experiencing carbon starvation showed a positive correlation between mRNA half-life and mRNA concentration. There is a lack of consensus regarding the direction and extent of the correlation between mRNA half-life and mRNA concentration. Furthermore, the causality of this correlation is not well characterized, as concluded by Nouaille et al. (2017). Formulating a complete understanding of the relationship between mRNA half-life and mRNA concentration in mycobacteria would fill a key gap in our understanding of how mRNA degradation occurs and how it is regulated.

In addition, the relationship between mRNA abundance and protein abundance and the relationship's impacts on protein synthesis rate are not well characterized in bacteria. A number of studies have identified factors that affect translation rate and contribute to translational regulation in bacteria. These factors include codon adaptation index (cAI) (Tuller et al., 2010; Riba et al., 2019), tRNA adaptation index (tAI) (Lenz et al., 2010; Riba et al., 2019), 5' UTR (Chen et al., 1991; Kozak et al., 2005; Review: Ren et al., 2017), and Shine-Dalgarno affinity strength (Li et al., 2012; Saito et al., 2020; Tarai & Asai, 2020). The steady-state relationship between mRNA abundance and protein abundance has been studied in *Pseudomonas aeruginosa*, *E. coli*, and *Saccharomyces cerevisiae*. Kwon et al. (2014) found a strong positive correlation between protein abundance and mRNA abundance within two closely related *Pseudomonas aeruginosa* strains in log-phase using DNA microarrays and LC-MS/MS proteomics. Comparing the protein to mRNA ratios between these two strains also produced a positive correlation, indicating that the protein to mRNA ratios are evolutionarily conserved between closely related strains (Kwon et al., 2014). In de Sousa Abreu et al. (2009), a meta-analysis of protein and mRNA abundance in both *E. coli* and *S. cerevisiae* established a weakly positive correlation between protein abundance and mRNA abundance in the two species. Taniguichi et al. (2010) reported no direct correlation between steady-state mRNA concentration and protein concentration using a yellow fluorescent protein translationally fused to the C-terminus of proteins of interest in their native positions within *E. coli*. Similar conclusions were drawn in another study of *E. coli* examining mRNA concentration and protein concentration of a subset of native genes (Lee et al., 2003). Another study that evaluated the mean translation rate per mRNA in single living cells using fluorescence correlation spectrometry (FCS) showed a positive, linear

correlation between the mRNA concentration and protein concentration for dsRED in *E. coli* (Guet et al., 2008). There is a lack of consensus regarding the exact nature of the steady-state relationship of mRNA and protein concentration. These differences could be the result of different methods being used to assess and analyze the relationship between mRNA abundance and protein abundance. For example, most studies examined large numbers of different transcripts and proteins expressed at their native levels, while some examined a single transcript and protein expressed at different levels. Understanding this relationship further would address whether the efficiency of mRNA translation is affected by the mRNA concentration. Additionally, this relationship has not been assessed in *M. smegmatis*. Addressing this goal in *M. smegmatis* could yield information regarding the relationship between mRNA and protein concentration in the context of *M. tuberculosis*.

Guided by previous research, this study sought to answer questions related to understanding the impact of mRNA concentration on gene expression. First, we sought to further investigate the nature of the negative correlation between mRNA half-life and mRNA concentration by examining the impact of transcription rate on mRNA stability in *M. smegmatis*. Second, we sought to examine the relationship between the protein concentration and mRNA abundance in *M. smegmati*s. To accomplish these goals, a set of *M. smegmatis* strains were constructed containing a tet-ON inducible system for temporal regulation of fluorescent proteins. Due to high variability and noise between biological replicates observed in the degradation data, we were unable to determine a meaningful relationship between mRNA abundance and mRNA half-life. In addition, our tetracycline-inducible gene expression system could not be used to investigate the relationship between mRNA abundance and protein abundance as a result of a decrease in inducer-based protein expression for undetermined reasons.

## Materials and Methods

### Strains and culture conditions

*Mycolicibacterium smegmatis* $mc^2155$ strain and all constructed strains were grown in Difco™ Middlebrook 7H9 medium with albumin dextrose catalase (ADC; final concentrations: 5 g/L bovine serum albumin fraction V (BSA), 2 g/L dextrose, 0.85 g/L NaCl, and 3 mg/L catalase), 0.2% glycerol, and 0.05% Tween 80. Cultures were shaken at 200 rpm and 37°C to an optical density at 600 nm ($OD_{600}$) between 0.5 to 0.8 at the time of harvest. Cultures grown and induced with anhydrotetracycline (aTc) were wrapped in aluminum foil to protect the photosensitive inducer.

### Plasmid construction

Plasmid pSS303 (Nguyen et al., 2020) was used as a vector backbone and the *yfp* coding sequence in that plasmid was replaced with either a *gfpmut3* or *mCherry* fluorescent reporter gene to create pSS470 and pSS471 (Table 1). The plasmid contained a strong constitutive $P_{myc1}$ *2X tetO* promoter , which was repressed in the presence of tet repressor protein (TetR) and initiated with the addition of aTc, an antibiotic derivative of tetracycline, with an associated $P_{myc1}$ 5' UTR (Blokpoel et al., 2005; Carroll et al., 2005; Ehrt et al., 2005). Two synthetic terminators were present in the vector; *tsynA* (Czyz et al., 2014) was upstream of the reporter gene, while *ttsbiB* (Huff et al., 2010) was downstream of the reporter gene. A 6×Histidine tag was added at the C terminus for the GFPmut3 protein (complete amino acid sequence:MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKL PVPWPTLVTTFGYGVQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAE VKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNI EDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGIT HGMDELYKCHHHHHH) and the mCherry protein (complete amino acid sequence:MAIIKEFMRFKVHMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFA WDILSPQFMYGSKAYVKHPADIPDYLKLSFPEGFKWERVMNFEDGGVVTVTQDSSLQD GEFIYKVKLRGTNFPSDGPVMQKKTMGWEASSERMYPEDGALKGEIKQRLKLKDGGH YDAEVKTTYKAKKPVQLPGAYNVNIKLDITSHNEDYTIVEQYERAEGRHSTGGMDELY KHHHHHH). The plasmids were integrated into the Giles phage site using an integrase promoter and protein sequence. Additionally, a *tet repressor* gene along with its strong constitutive $P_{myc1}$ promoter and associated $P_{myc1}$ 5' UTR were used in conjunction with the pSS221 vector backbone to create pSS555. The plasmid was integrated into the L5 phage site using an integrase promoter and protein sequence.

All constructs were built using NEBuilder HiFi DNA assembly master mix (catalogue E2621) (Table 1). To create *M. smegmatis* $mc^2155$ strain with both *gfpmut3* and *tetR* genes, pSS470 was transformed using electroporation (Bio-Rad, catalogue 1652100) into *M. smegmatis* $mc^2155$ strain and integrated at the Giles phage site and selected with 250 μg/mL Hygromycin B. pSS555 was later transformed using electroporation and integrated in the same strain at the L5 phage site and selected with 40 μg/mL nourseothricin (Table 1). These steps were repeated to

create *M. smegmatis* $mc^2$_155 strain with both *mCherry* and *tetR* genes (Table 1). Additionally, the previously mentioned steps were used to create *M. smegmatis* $mc^2$155 strains with only pSS470, pSS471, or pSS555 (Table 1). To confirm successful plasmid integrations at Giles and L5 phage sites, several different primers were used (Table 2 and Table 3).

**Table 1.** A list of the *Mycolicibacterium smegmatis* strains and plasmids. *HygR* refers to a gene that produces a protein which confers resistance to Hygromycin B. *NAT* refers to a gene that produces nourseothricin N-acetyl transferase which confers resistance to nourseothricin.

| Strain | Plasmid | Plasmid Description |
|---|---|---|
| SS-M_0836; SS-M_1097; SS-M_1098; SS-M_1099 | pSS470 | $P_{myc1}$ *2X tetO* promoter + $P_{myc1}$ 5′ UTR+ *gfpmut3*-6xHis + *HygR* |
| SS-M_0837; SS-M_0838 | pSS471 | $P_{myc1}$ *2X tetO* promoter + $P_{myc1}$ 5′ UTR+ *mCherry*-6xHis + *HygR* |
| SS-M_1062; SS-M_1063 | pSS470 + pSS555 | $P_{myc1}$ *2X tetO* promoter + $P_{myc1}$ 5′ UTR+ *gfpmut3*-6xHis + *HygR* (Giles site) and $P_{myc1}$ promoter + $P_{myc1}$ 5′ UTR+ *tetR* + *NAT* (L5 site) |
| SS-M_1073; SS-M_1074 | pSS471 + pSS555 | $P_{myc1}$ *2X tetO* promoter + $P_{myc1}$ 5′ UTR+ *mCherry*-6xHis + *HygR* (Giles site) and $P_{myc1}$ promoter + $P_{myc1}$ 5′ UTR+ *tetR* + *NAT* (L5 site) |
| SS-M_1071; SS-M_1072 | pSS555 | $P_{myc1}$ promoter + $P_{myc1}$ 5′ UTR+ *tetR* + *NAT* |

**Table 2.** Primers used for plasmid construction, colony-checking PCR, and verifying plasmid integration into the Giles and L5 sites. *HygR* refers to a gene that produces a protein which confers resistance to Hygromycin B. *NAT* refers to a gene that produces nourseothricin N-acetyl transferase, which confers resistance to nourseothricin. Forward primers are denoted with an 'F' character whereas reverse primers are denoted with an 'R' character.

| Plasmid | Primers for Amplification and Plasmid Creation | Primers for Colony-Checking PCR for fidelity | Primers for verifying plasmid integration |
|---|---|---|---|
| pSS470 | Insert Amplification: SSS1992F,SSS1993R<br><br>Vector Amplification: SSS1989F,SSS1800R | SSS1171F, SSS1851R | Giles Left integration: SSS1172F, SSS1174R<br><br>Giles Right integration: SSS1173F, SSS1175R |
| pSS471 | Insert Amplification: SSS1990F,SSS1991R<br><br>Vector Amplification: SSS1989F,SSS1800R | SSS1171F, SSS1851R | Giles Left integration: SSS1172F, SSS1174R<br><br>Giles Right integration: SSS1173F, SSS1175R |
| pSS555 | Insert Amplification: SSS2159F,SSS2205R<br><br>Vector Amplification SSS1488R,SSS1519F | SSS2315F, SSS1641R | L5 Left integration: SSS1103F, SSS142R<br><br>L5 Right integration: SSS1104F, SSS144R |

**Table 3.** Primer sequences. Listed in the table below are the sequences of primers listed in Table 2.

| Primer | Description | Sequence (5' → 3') |
|---|---|---|
| SSS1992 | Forward primer to amplify *gfpmut3* in pMV762 | TTAAGAAGGAGATATACATCATGAGTAAAGGAGAAGAAC |
| SSS1993 | Reverse primer to amplify *gfpmut3* in pMV762 | GTGATGGTGATGGTGATGACATTTGTATAGTTCATCCATGC |
| SSS1990 | Forward primer to amplify *mCherry* in pSS374 | TTAAGAAGGAGATATACATCATGGCCATCATCAAGGAGTTC |
| SSS1991 | Reverse primer to amplify *mCherry* in pSS374 | TGATGGTGATGGTGATGACACTTGTACAGCTCGTCCATGC |
| SSS1989 | Forward primer to amplify pSS303 | TGTCATCACCATCACCAT |
| SSS1800 | Reverse primer to amplify pSS303 | GATGTATATCTCCTTCTTAAT |
| SSS2159 | Forward primer to amplify $P_{myc1}$ promoter + $P_{myc1}$ 5′ UTR+ *tetR* in pSS221 | CAAACTCTTCCTGTCGTCATATAGAAATATTGGATCGTCGG |
| SSS2205 | Reverse primer to amplify $P_{myc1}$ promoter + $P_{myc1}$ 5′ UTR+ *tetR* in pSS221 | GTTAACTACGTCGACATCGATATTAAGACCCACTTTCACATTTAAG |
| SSS1519 | Forward primer to amplify pJEB402 | TATCGATGTCGACGTAGTTAAC |
| SSS1488 | Reverse primer to amplify pJEB402 | ATATGACGACAGGAAGAGTT |
| SSS1171 | Forward primer to amplify *gfpmut3* or *mCherry* for colony-checking PCR | GGAAAAGAGGTCATCCAGGA |
| SSS1851 | Reverse primer to amplify *gfpmut3* or *mCherry* for colony-checking PCR | GGCAACGCCCTAGTGATGGTGATGGTGATGAC |

| | | |
|---|---|---|
| SSS2315 | Forward primer to amplify $P_{myc1}$ promoter + $P_{myc1}$ *5′* UTR+ partial *tetR* for colony-checking PCR | CGGTGAACGCTCTCCTG |
| SSS1641 | Reverse primer to amplify $P_{myc1}$ promoter + $P_{myc1}$ *5′* UTR+ partial *tetR* for colony-checking PCR | TAGGCTGCTCTACACCAAGC |
| SSS1172 | Forward primer to check left junction in Giles site in *M. smegmatis* | CTCCGAACTCCTCCGAAACC |
| SSS1173 | Forward primer to check right junction in Giles site in *M. smegmatis* | ACATATCTGTCGAAGCGCCC |
| SSS1174 | Reverse primer to check left junction in Giles site in *M. smegmatis* | TGACGATCAACTCCGCGGGGCCGGGCCA |
| SSS1175 | Reverse primer to check right junction in Giles site in *M. smegmatis* | CGGTGGATCCGCGCAACCTG |
| SSS1103 | Forward primer to check left junction in L5 site in *M. smegmatis* | TGGATTTGGTTTCAGCTCCC |
| SSS142 | Reverse primer to check left junction in L5 site in *M. smegmatis* | TAGAGCCGTGAACGACAGG |
| SSS1104 | Forward primer to check right junction in L5 site in *M. smegmatis* | GACCTTGGTGCAGAAATCGC |
| SSS144 | Reverse primer to check right junction in L5 site in *M. smegmatis* | TCGATGAGCCGCTTCTCGC |

## Polymerase Chain Reaction (PCR) and DNA Recovery

Polymerase Chain Reaction (PCR) was performed in a 25 µL sample reaction volume. When using Q5, high-fidelity polymerase, the following volumes of necessary components were added to each reaction: 5.0 µL of 5X Q5 polymerase buffer (New England Biolabs, catalogue B9027S), 5.0 µL of 5X GC enhancer (New England Biolabs, catalogue B9208A), 0.50 µL of 10 µM forward primer, 0.50 µL of 10 µM reverse primer, 0.5 µL of 10 µM dNTPs, 0.25 µL of Q5 polymerase (New England Biolabs, catalogue M0491L), and 1.0 µL of the DNA to be amplified. UltraPure™ DNase/RNase-free distilled water was added for the remaining volume. When using Taq polymerase, the following volumes of necessary components were added to each reaction: 2.5 µL of 10X Taq polymerase buffer (New England Biolabs, catalogue B9014S), 1.0 µL of DMSO, 0.5 µL of 10 µM forward primer, 0.5 µL of 10 µM reverse primer, 0.5 µL of 10 µM dNTPs, 0.125 µL of Taq polymerase (New England Biolabs, catalogue M0273L), and approximately 5.0 µL of colony for colony-checking PCR. UltraPure™ DNase/RNase-free distilled water was added for the remaining volume. The forward and reverse primers for each strain can be found in Table 2. PCR was carried out at an initial denaturation step at 95℃ for 2 minutes, (i) a denaturation step of 95℃ for 20 seconds, (ii) an annealing step at an appropriate primer annealing temperature (℃) for 40 seconds, (iii) an elongation step at 72℃ for 2 minutes, and a final elongation step at 72℃ for 5 minutes. The steps labeled i, ii, and iii were repeated 35 times. Annealing temperatures were optimized for each primer set using the New England Biolabs Tm calculator. Additionally, the elongation time was based on the size of the PCR product following the precedent of 1 minute/kilobase.

All products were analyzed by gel electrophoresis using a 1.0% agarose gel with 0.2-0.5 µg/mL of ethidium bromide (EtBr) (depending on the mass of the gel) prepared in 1X Tris-acetate-EDTA (TAE) buffer. Gels were visualized using UV-light (Bio-Rad) and Quantity One software. Bands of interest were cut from the gel and purified using the Zymoclean™ Gel DNA Recovery Kit (Zymo Research, catalogue D4002) following the manufacturer's instructions. A NanoDrop One (Thermo Scientific) was used to measure sample concentrations of DNA. Any DNA sample intended for sequencing was sent to QuintaraBio following the company's instructions.

## Flow cytometry

Cultures of *M. smegmatis* were grown in duplicate at a variety of induction times (1 hour, 4 hours, 24 hours, 26 hours, 28 hours, and 30 hours) and aTc induction concentrations (0 ng/mL, 1 ng/mL, 2 ng/mL, 5 ng/mL, 10 ng/mL, 20 ng/mL, 50 ng/mL, 100 ng/mL, and 200 ng/mL) at $OD_{600}$ between 0.5 and 1.3. Harvested cultures were placed on ice, and diluted to form two, 1.0 mL cultures, which served as duplicates, with an $OD_{600}$ of 0.025 freshly filtered 7H9 media and then filtered with a 5 µm filter needle to remove clumps . A CytoFlex flow cytometer was used to measure 5000 events of each culture with one universal gate drawn to encompass the densest region of cells on a forward scatter (FSC) vs violet side scatter (SSC) plot. The gain values were

500 for FSC and 50 for violet SSC. The thresholds were 100,000 for violet SSC-H and 40,000 for FSC-H. FlowJov10.8.0 was utilized to draw gates and analyze fluorescence data.

### RNA extraction

RNA extractions were performed on triplicate cultures. Cell cultures were induced with one of the following aTc concentrations for 4 hours: 0 ng/mL, 2.5 ng/mL, 5 ng/mL, 10 ng/mL, 20 ng/mL, or 50 ng/mL. The cultures were then spun down to form a cell pellet. Any pellets that were not used immediately for extraction were stored at -80 ℃ and were thawed on ice before extraction. Cells were resuspended in 1.0 mL of TRIzol Reagent (VWR, catalogue MSPP-TR118), and pipetted into a 2.0 mL beating tube (OPS Diagnostics; 100-µm zirconium lysing matrix, molecular grade). The cells were then lysed using a FastPrep-24 5G instrument (MP Biomedical) (3 cycles of 7 m/s for 30 s, with 2 min on ice between cycles). Samples were treated with 300 µL of chloroform before being centrifuged for 15 minutes at 15,000 rpm and 4℃. The resulting aqueous layer was recovered from the sample, and extraction was completed using Direct-Zol™ RNA miniprep (Zymo Research, catalogue R2052) following the manufacturer's instructions with an in-column DNase treatment. Sample concentrations and absorbance ratios were measured using a NanoDrop One (Thermo Scientific) before being stored at -80℃. RNA quality was assessed by gel electrophoresis. A volume containing 300 ng of extracted RNA was mixed with 2X RNA loading dye (New England Biolabs, catalogue 50-427-9) and heated at 65℃ for 5 minutes using a heat block. Heated samples were loaded onto a 1.0% agarose gel with 0.2-0.5 µg/mL of EtBr (depending on the mass of the gel), and run with 1X Tris/Borate/EDTA (TBE) buffer. Gels were visualized using UV-light (Bio-Rad) and Quantity One software.

### cDNA synthesis and cleanup

Each solution of extracted RNA was diluted into two, separate 5.25 µL, samples containing 600 ng of RNA using UltraPure™ DNase/RNase-free distilled water, one used as a negative control without reverse transcriptase (no RT) and one with reverse transcriptase (RT). A volume of 1.0 µL mix, containing 0.83 µL of 100 mM Tris, pH 7.5, and 0.17 µL of 3.0 mg/mL of random primers (Invitrogen: Catalogue No. 48190011), was pipetted to each of the diluted RNA samples. Samples were incubated at 70°C for 10 minutes before being snap-frozen in an ice water bath for 5 minutes and then transferred onto ice. A 3.75 µL mix was pipetted to the RT samples containing the following components: 2 µL of ProtoScript II RT Reaction Buffer (NEB: Catalogue No. B0368S), 0.5 µL of 10 mM each dNTPs, 0.5 µL of 100 nM DTT (NEB: Catalogue No. B1034A), 0.25 µL of RNase Inhibitor, Murine, New England Biolabs (40,000 U/mL), and 0.5 µL of ProtoScript® II Reverse Transcriptase, New England Biolabs (200,000 U/mL). A 3.75 µL mix was pipetted to the no RT samples containing the following components: 2 µL of ProtoScript II RT Reaction Buffer, 0.5 µL of 10 mM each dNTPs, 0.5 µL of 100 nM DTT, 0.25 µL of RNase Inhibitor, Murine, New England Biolabs (40,000 U/mL), and 0.5 µL of UltraPure™ DNase/RNase-free distilled water. Samples were incubated at 25°C for 10 minutes followed by 42°C for 2 hours. Samples were treated with 10 µL of master mix containing 5 µL of

0.5 M of EDTA and 5 µL of 1N NaOH to degrade any remaining RNA. Samples were then incubated at 65°C for 15 minutes before 12.5 µL of Tris HCl, pH 7.5, was added to neutralize the pH. All cDNA samples were cleaned up using the MinElute PCR Purification Kit (NEB #T1030L) following the manufacturer's instructions.

### Quantitative PCR (qPCR)

All qPCR reactions were performed in a BSL-2 approved biosafety cabinet. qPCR was performed on 3 biological replicates. All samples of cDNA were obtained from RNA extraction and cDNA synthesis and cleanup as described above. Samples of cDNA were diluted firstly to 1.0 ng/µL from their original concentration in pure water. Those samples were then diluted to the desired concentration of 200 pg/µL. A 2.5 µM primer mix for *sigA* was created with a 1.12 sample error. The final primer mix contained 2.5 µM of JR273 and 2.5 µM of JR274 (Table 4). A similar 2.5 µM primer mix was created for *gfpmut3* resulting in a final primer mix with 2.5 µM of SSS306 and 2.5 µM of SSS308 (Table 4). For each of the targets, a mastermix was created using 1 µL of the appropriate primer mix, 2 µL of ultra-pure water, and 5 µL of iTaq Universal SYBR Green Supermix (Bio-Rad: Catalogue No. 172-5124). Mastermixes were kept on ice until use. Within a 96-well PCR microplate (Axygen™: Catalogue No. 14-222-334), 2 µL of cDNA was pipetted into each well. Once all wells were filled with cDNA, 8 µL of the appropriate mastermix for *sigA* or *gfpmut3* respectively was added to each well and mixed carefully using a micropipette. Water controls were made with the specific mastermixes but replacing the 2 µL of cDNA with 2 µL of ultra-pure water. The plate was then covered with a sealing film (Axygen: Catalogue No. UC-500). qPCR was run in a QuantStudio 6 Pro Real-Time PCR System (QuantStudio: Catalogue No. A43180). Samples in the microplate were incubated at (i) 50°C for 2 minutes, (ii) a 95°C for 1 minute, (iii) 95°C hold for 15 seconds, and (iv) 61°C hold for 2 minutes. In the final step (iv), the fluorescence of SYBR Green was recorded. Steps (iii) and (iv) were cycled through 40 times. To obtain information regarding transcript abundance and relative expression, the *gfpmut3* gene was normalized to the *sigA* housekeeping gene. For each sample, the number of cycles ($C_t$) of a gene of interest required to pass a threshold ($C_T$) of 0.1 was compared to *sigA*. The difference was used to calculate the $\Delta C_t$ for each sample. Relative expression was calculated as $2^{-\Delta Ct}$. Analysis and figure creation was completed using GraphPad Prism 9.

**Table 4.** Primers used for qPCR.

| Primer | Description | Sequence (5' → 3') |
|---|---|---|
| JR273 | Forward primer to amplify *sigA* cDNA in *M. smegmatis* | GACTACACCAAGGGCTACAAG |
| JR274 | Reverse primer to amplify *sigA* cDNA in *M. smegmatis* | TTGATCACCTCGACCATGTG |
| SSS306 | Forward primer to amplify *gfpmut3* cDNA | GAAGGTGATGCAACATACGG |
| SSS308 | Reverse primer to amplify *gfpmut3* cDNA | TCCTGTACATAACCTTCGGG |

## Rifampicin Experiment for mRNA Half-Life Determination

Thirty-two mL cultures of *M. smegmatis* were grown to an $OD_{600}$ of 0.5 and 0.7 in 250 mL Erlenmeyer flasks for each concentration of aTc for each biological replicate. Four hours before the step of adding rifampicin, aTc was added to achieve final concentrations of 0, 2.5, 5, 10, 20, 50 ng/mL. Three and a half hours later, 5 mL of bacterial culture was aliquoted from each flask into 5, 15 mL conical tubes, for five time points (0, 0.5 minute, 1 minute, 2 minutes, and 4 minutes). The conical tubes were placed on the tissue culture rotator and spun for 30 minutes. Cultures were then treated with rifampicin at a final concentration of 150 ug/mL to halt transcription and snap-frozen in liquid nitrogen after 0, 0.5 minute, 1 minute, 2 minutes, and 4 minutes. The cultures were then stored at -80 degrees Celsius for future RNA extractions, which was followed by cDNA synthesis and qPCR.

Transcript abundance of *sigA* and *gfpmut3* were used to determine mRNA half-lives. For each gene, the $C_t$ was made negative, which represents transcript abundance on a $\log_2$ scale, and linear regression was performed on a plot of the negative $C_t$ versus time using GraphPad Prism 9. Half-life was defined as the negative reciprocal of the best-fit slope (Equation 1).

$$\text{Half-life} = -\frac{1}{slope} \quad (1)$$

As seen previously in the context of mycobacteria, plotting $\log_2$ abundance over time produces a biphasic decay curve with a period of faster exponential decay followed by a period of much slower or undetectable exponential decay (Nguyen et al., 2020). Other studies in *E. coli* have observed similar biphasic curves for a variety of different genes (Blum et al., 1999; Brescia et al., 2004; Chen et al., 2015; Sinha et al., 2018).

## Results & Discussion

### Constructing *Mycolicibacterium smegmatis* strains with aTc-inducible GFPmut3 and mCherry

To determine the degree of correlation between mRNA half-life and concentration and whether transcription is causal in that relationship, we needed to measure mRNA half-life at a variety of mRNA concentrations for a single gene in *Mycolicibacterium smegmatis*. An anhydrotetracycline (aTc) induced expression system was selected to create a set of *M. smegmatis* strains given its strict temporal regulation of gene expression (Ehrt et al., 2005). We constructed three plasmids: one containing the two *tet-operator* $P_{myc1}$ promoter and the associated $P_{myc1}$ 5' UTR linked to a *gfpmut3* gene, one containing the two *tet-operator* $P_{myc1}$ promoter with the associated $P_{myc1}$ 5' UTR linked to an *mCherry* gene and another containing the same promoter and 5' UTR without the tet operators, linked to a *tet repressor* gene (Figure 1A, Figure 1B, and Figure 1C). We conducted fluorescence microscopy to verify the expected protein expression in the experimental strains containing *gfpmut3 + tetR*, *mCherry + tetR* and control strains with *gfpmut3* only and *tetR* only to validate that the system performed as expected in the presence and absence of aTc. Strains were induced with 0 ng/mL and 200 ng/mL of aTc and incubated for 24 hours for the *gfpmut3* only strain, the *tetR* only strain, and the *gfpmut3 + tetR* strain. Strains were induced with 0 ng/mL and 200 ng/mL of aTc and incubated for 4 hours for the *mCherry* only strain, the *tetR* only strain, and the *mCherry + tetR* strain. The *gfpmut3* and *mCherry* only strains fluoresced brightly in the presence and absence of aTc as expected (Figure 1C and Figure 2C). The *tetR* only strain did not fluoresce in either the presence and absence of aTc (Figure 1D and Figure 2D). The experimental strains containing *gfpmut3 + tetR* or *mCherry + tetR* only showed fluorescence in the presence of aTc (Figure 1E and Figure 2E). From these results, the performance of the tetracycline-inducible system was validated.

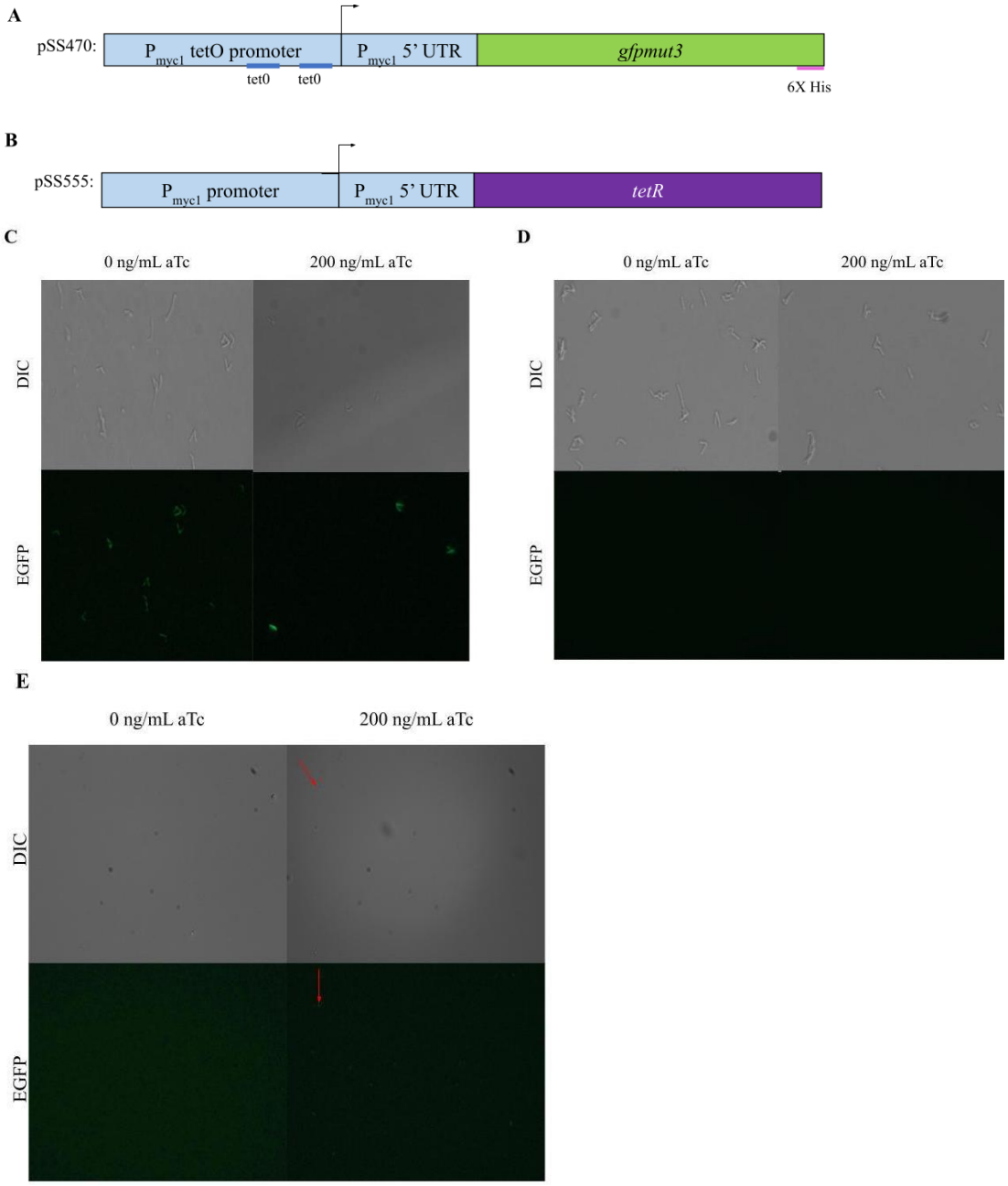



**Figure 1.** Constructing the aTc inducible system for GFPmut3 protein expression. **(A)** Schematic of the plasmid with the $P_{myc1}$ promoter with associated $P_{myc1}$ 5' UTR and two *tet-operator* regions and the *gfpmut3* gene inserted into the *M. smegmatis* Giles site **(B)** Schematic for the plasmid with the $P_{myc1}$ promoter with associated $P_{myc1}$ 5' UTR *tetR* gene inserted into the *M. smegmatis* L5 site **(C)** Fluorescence microscopy of *M. smegmatis* strains that contained a $P_{myc1}$ promoter with associated $P_{myc1}$ 5' UTR and two *tet-operator* regions linked *gfpmut3* gene inserted into the *M. smegmatis* Giles site **(D)** Fluorescence microscopy of *M. smegmatis* strains that contained a $P_{myc1}$ promoter with associated $P_{myc1}$ 5' UTR linked *tetR* gene **(E)** Fluorescence microscopy of *M. smegmatis* strains that contained a $P_{myc1}$ promoter with associated $P_{myc1}$ 5' UTR and two *tet-operator* regions linked *gfpmut3* gene inserted into the *M. smegmatis* Giles site and a $P_{myc1}$ promoter with associated $P_{myc1}$ 5' UTR linked *tetR* gene inserted into the *M. smegmatis* L5 site. Red arrows point out fluorescing cells.

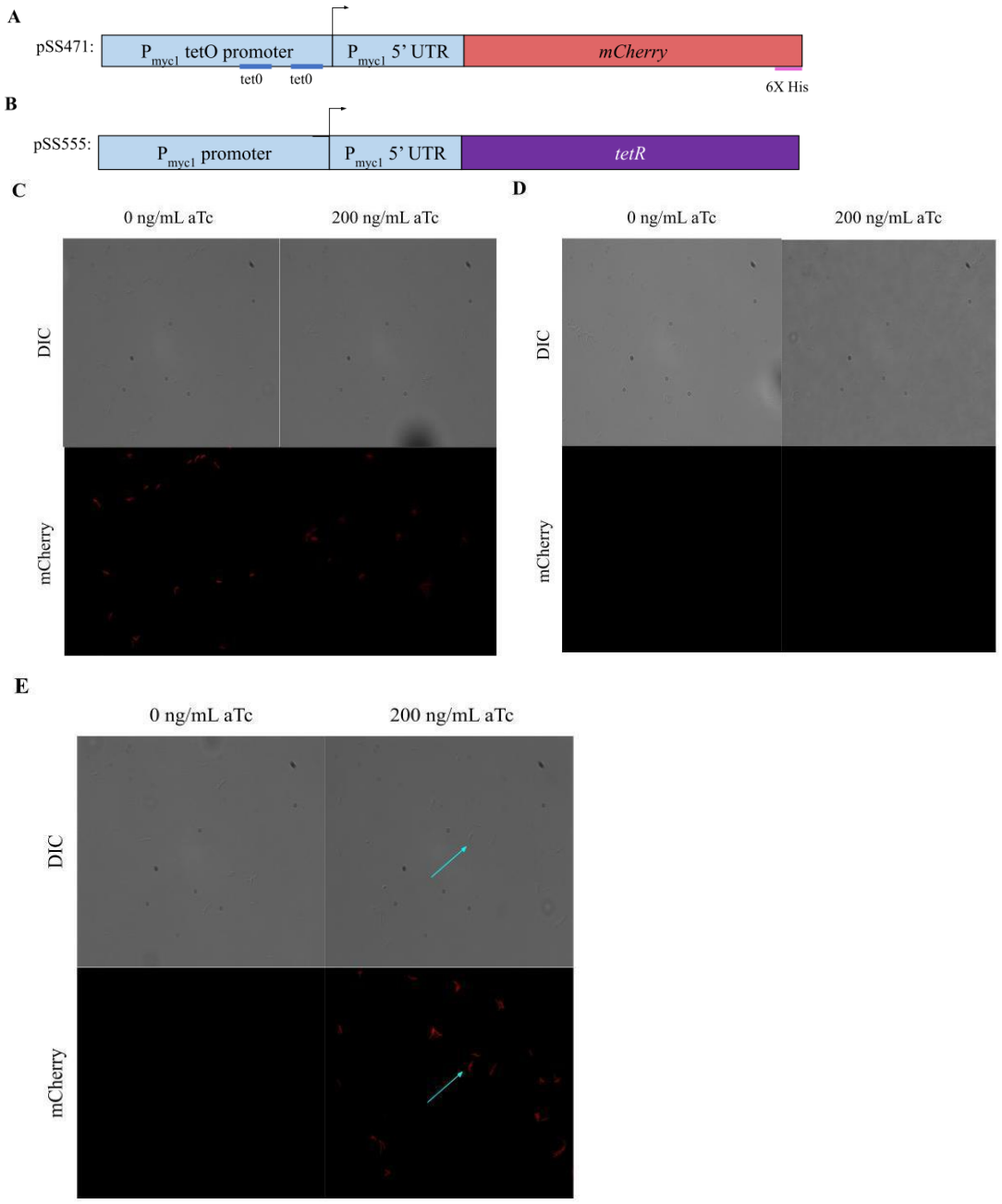



**Figure 2.** Constructing the aTc inducible system for mCherry protein expression. **(A)** Schematic of the plasmid with the $P_{myc1}$ promoter with associated $P_{myc1}$ 5' UTR and two *tet-operator* regions linked *mCherry* gene inserted into the *M. smegmatis* Giles site **(B)** Schematic of the plasmid with the $P_{myc1}$ promoter with associated $P_{myc1}$ 5' UTR *tetR* gene inserted into the *M. smegmatis* L5 site **(C)** Fluorescence microscopy of *M. smegmatis* strains that contained a $P_{myc1}$ promoter with associated $P_{myc1}$ 5' UTR and two *tet-operator* regions linked *mCherry* gene inserted into the *M. smegmatis* Giles site **(D)** Fluorescence microscopy of *M. smegmatis* strains that contained a $P_{myc1}$ promoter with associated $P_{myc1}$ 5' UTR linked *tetR* gene **(E)** Fluorescence microscopy of *M. smegmatis* strains that contained a $P_{myc1}$ promoter with associated $P_{myc1}$ 5' UTR and two *tet-operator* regions linked *mCherry* gene inserted into the *M. smegmatis* Giles site and a $P_{myc1}$ promoter with associated $P_{myc1}$ 5' UTR linked *tetR* gene inserted into the *M. smegmatis* L5 site. Blue arrows point out fluorescing cells.

**Four hours induction generates the widest range of GFPmut3 expression for many aTc induction concentrations**

Studies have reported mixed data regarding the relationship between steady-state mRNA and protein concentration (Lee et al., 2003; Guet et al., 2008; de Sousa Abreu et al., 2009; Taniguchi et al., 2011; Kwon et al., 2014). This relationship is not well-studied in the context of *M. smegmatis*. Using this tetracycline-inducible gene expression system, it was a goal of this study to understand the steady-state relationship between mRNA abundance and protein abundance in *M. smegmatis***.** Guided by previous work in *M. smegmatis* and *Mycobacterium tuberculosis*, we found that the aTc concentrations of 0, 1, 5, 10, 20, 50, 100, 200 ng/mL were the most used and applicable to generate our desired wide range of mRNA levels (Sinha et al., 2007; Raghavan et al., 2008; Korch et al., 2009; Goyal et al., 2011; Minch et al., 2012). To determine the appropriate induction time with those aTc concentrations that would provide us with a wide range of GFPmut3 protein expression, we conducted flow cytometry after three different induction times: 1 hour, 4 hours, and 24 hours. After one hour or 24 hours of induction, samples exposed to higher aTc concentrations were fluorescent but those exposed to lower aTc concentrations did not consistently have above-background fluorescence (Figure 3A and 3C). In contrast, the 4 hour induction time point showed a greater range of GFPmut3 fluorescence (Figure 3B). Strains were also induced with the noted aTc concentrations and incubated for 24 hours, 26 hours, 28 hours, and 30 hours, to examine the potential of increasing the spread of GFPmut3 expression levels. Median fluorescence levels of GFPmut3 decreased as the incubation time increased, indicating a decrease in protein expression levels (Figure 4). From these results, we determined that 4 hours induction was the best time point that would allow us to establish the correlation between mRNA abundance and mRNA half-life.

Due to loss of aTc induced expression over time, we were not able to study the steady-state relationship between mRNA abundance and protein abundance in *M. smegmatis* in this paper. Instead, our results shed light on the use of a tetracycline-inducible gene expression system for studying the expression of fluorescent proteins and mRNA. When examining induction times greater than 24 hours, GFPmut3 expression decreased as induction time increased (Figure 4). This decrease in expression is most likely due to a loss of aTc. The inducer aTc has been shown to be temperature sensitive and especially photosensitive in L-broth (LB) and M9 media (Ehrt, et al., 2005; Politi et al., 2014; Baumschlager, et al., 2020). One study that worked to understand aTc degradation in the context of *E. coli* grown in LB and M9 through a linear model attributed nearly 42.5% of degradation to residual error unexplained with their linear model while 41.9% of degradation was due to temperature (Politi et al., 2014). Long-term exposure to higher temperatures, such as those used to grow bacterial cultures, appears to be a significant factor that accounts for aTc degradation (Politi et al., 2014). To better improve and use aTc in future *M. tuberculosis* and *M. smegmatis* experiments, a similar study as Politi et al. (2014) could be conducted in 7H9 media and at 37°C. Politi et al. (2014) also assessed the half-lives of other inducers beside aTc including IPTG and HSL and found that IPTG has the highest half-life out of three inducers and is stable over 32 hours. IPTG could be considered as a

potential inducer for future experiments, especially regarding longer induction times. Long GFPmut3 protein half-lives and loss of aTc induced expression (Figure 4) indicates that steady-state *gfpmut3* mRNA expression (see below) does not coincide with steady-state GFPmut3 protein expression.

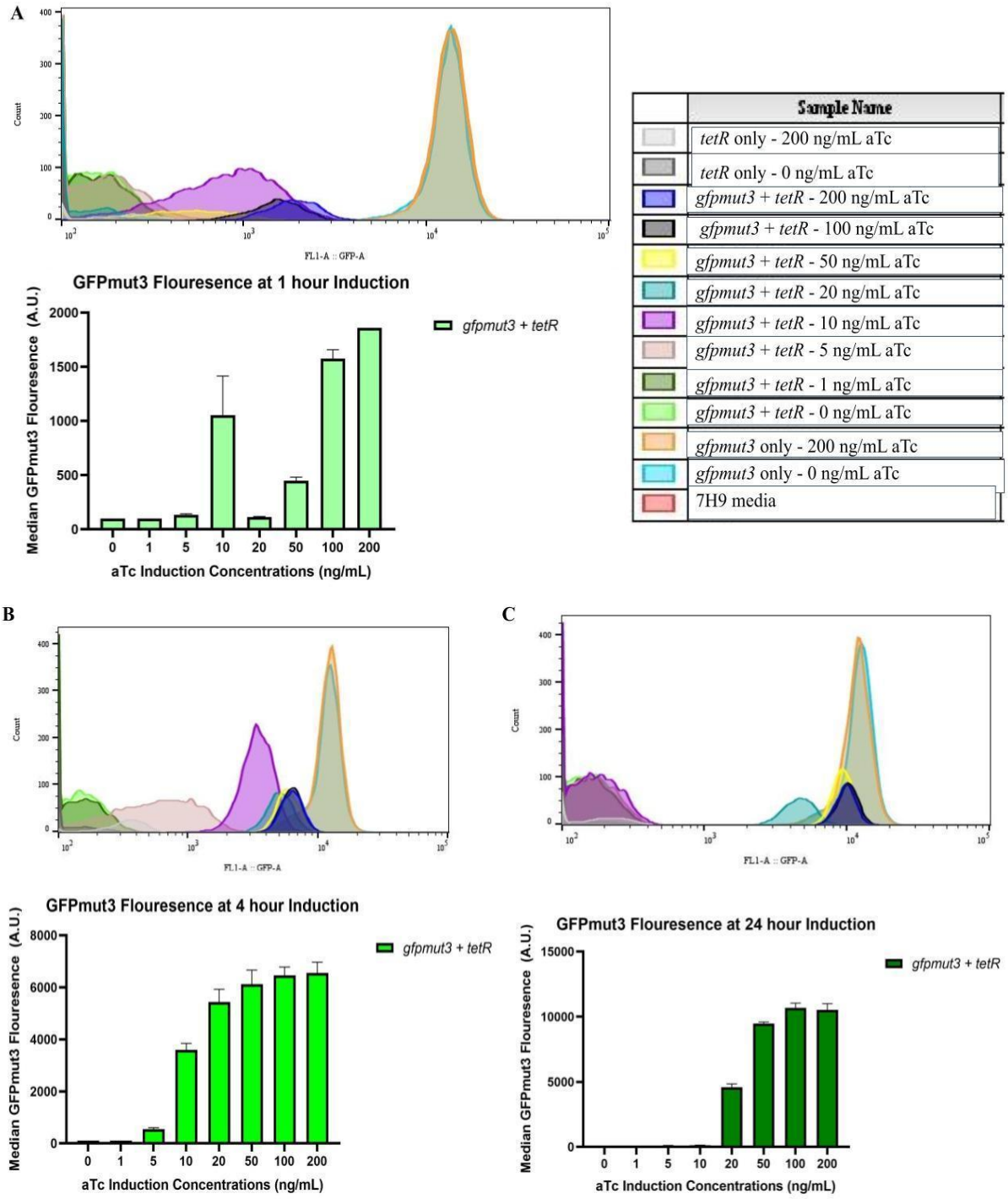



**Figure 3.** Validating the use of a 4 hour aTc induction time point for greatest spread of protein expression levels. Samples of the strains containing *gfpmut3 + tetR* were incubated at 0 ng/mL, 1 ng/mL, 5 ng/mL, 10 ng/mL, 20 ng/mL, 50 ng/mL, 100 ng/mL, or 200 ng/mL of aTc for different induction times. Samples containing *gfpmut3* only, and *tetR* only were incubated at 0 ng/mL and 200 ng/mL of aTc at different induction times. The fluorescence histogram from one duplicate is shown. The mean of median fluorescence of GFPmut3 expression of the duplicate samples and standard deviation were quantified using FlowJo. **(A)** 1-hour induction **(B)** 4-hours induction **(C)** 24-hours induction.

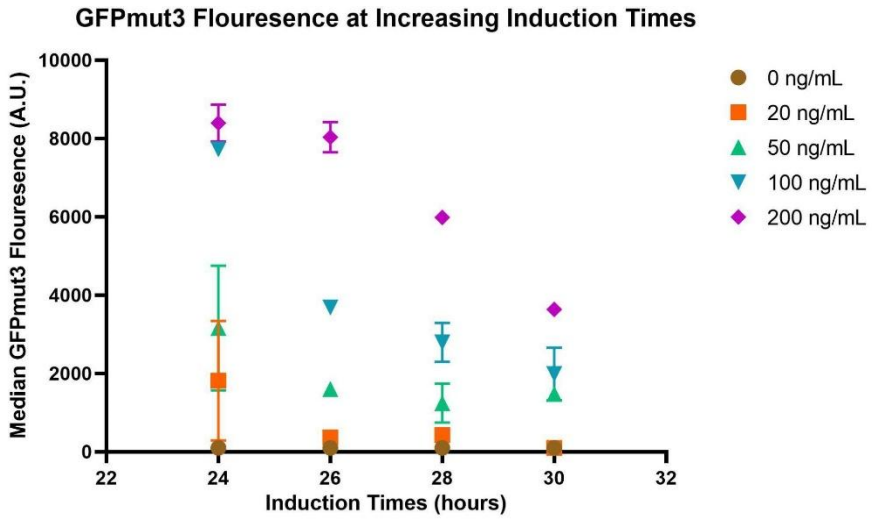



**Figure 4.** Induction times above 24 hours led to a decrease in GFPmut3 fluorescence. Samples of the strains containing *gfpmut3 + tetR* were incubated at 0ng/mL, 20 ng/mL, 50 ng/mL, 100 ng/mL, or 200 ng/mL of aTc for 24 hours, 26 hours, 28 hours, and 30 hours. Fluorescence was quantified by flow cytometry. The mean of median fluorescence of GFPmut3 expression of the duplicate samples and standard deviation were quantified using FlowJo.

**TetR-controlled mCherry acts like an on-off inducible switch system for the variety of aTc concentrations**

The initial goal of this research was to engineer plasmids with similar constructs but with two different genes, either *gfpmut3* or *mCherry*, to see if our findings held true for different genes. Based on the results of flow cytometry with the *gfpmut3* strains (Figure 3), we chose to test the *mCherry* strains after 4 and 24 hours of induction using the aTc concentrations of 0, 5, 10, 20, 50, 100, and 200 ng/mL. The 1-hour induction was excluded from testing the *mCherry* strains because we found that 1-hour induction was not long enough to give a wide range of GFPmut3 protein levels in the *gfpmut3 + tetR* strains (Figure 3A). In contrast to the *gfpmut3 + tetR* results, 4-hours induction was not long enough to allow strains containing *mCherry + tetR* with low concentrations from 5 ng/mL to 50 ng/mL of aTc to produce expression greater than the negative control strains (Figure 5A). In addition, there was a large difference between the fluorescence of samples induced with 100 ng/mL and 200 ng/mL of aTc compared to samples induced with 50 ng/mL of aTc. The *mCherry + tetR* induction system therefore seemed to behave as an on-off inducible switch system instead of a titratable system as seen in Figure 3B where a wide range of GFPmut3 fluorescence was observed. The 24 hours induction compared to 4 hours induction showed a sharp decrease in the median fluorescence levels of strains containing *mCherry + tetR* with 100 ng/mL and 200 ng/mL of aTc, which brought them closer to the median fluorescence levels of negative control strains (Figure 5B). Because of these observations, we decided that the *mCherry* strains were not able to establish a wide enough range of protein expression by using these aTc concentrations and time points of induction. As a result, these strains were excluded from the rest of this study.

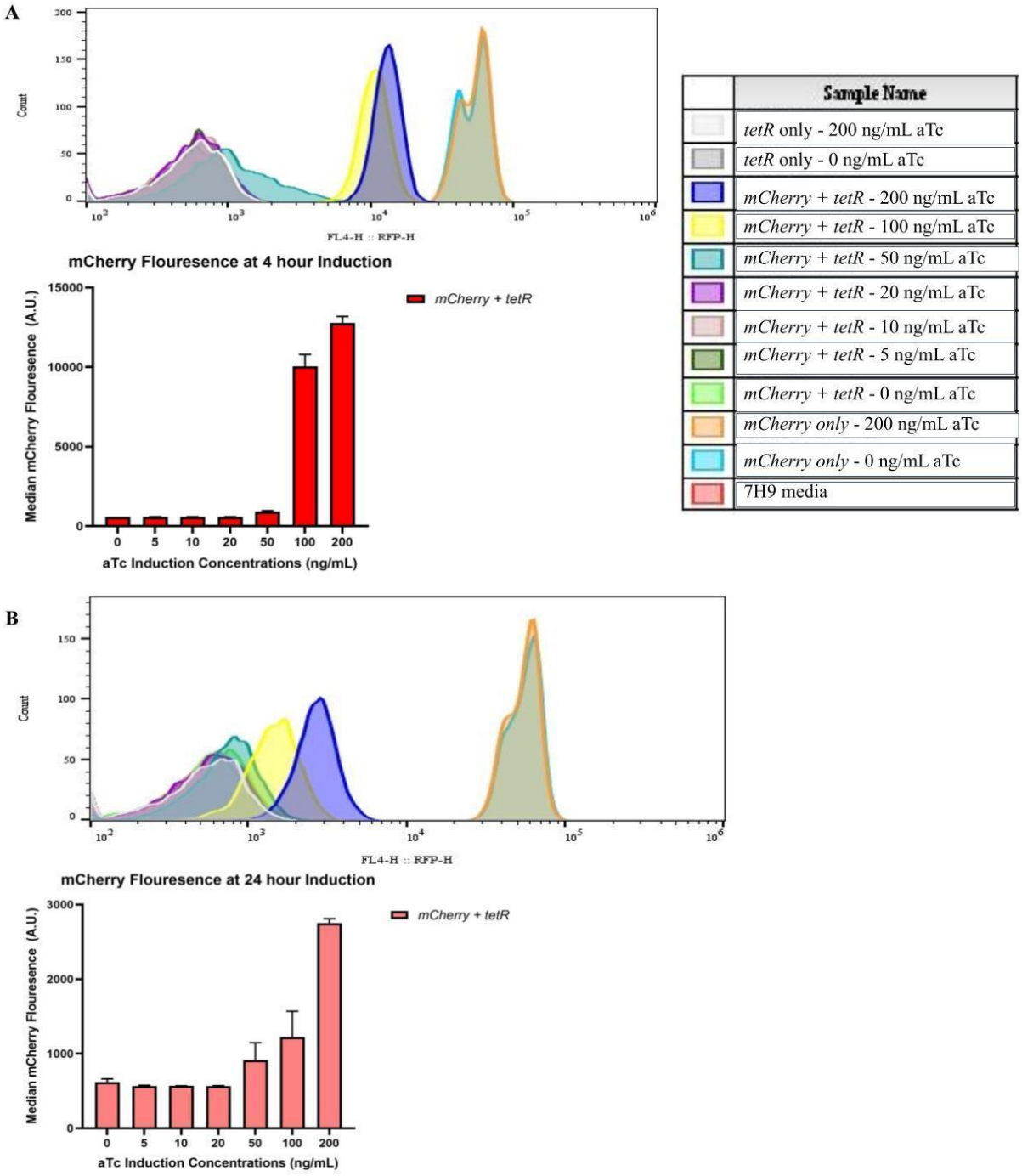



**Figure 5.** The tested range of aTc concentrations and induction times produced two mCherry fluorescence states: on and off. Samples of the strains containing *mCherry + tetR* were incubated at 0 ng/mL, 5 ng/mL, 10 ng/mL, 20 ng/mL, 50 ng/mL, 100 ng/mL, or 200 ng/mL of aTc for different induction times. Samples containing *mCherry* only, and *tetR* only were incubated at 0 ng/mL and 200 ng/mL of aTc at different induction times. The fluorescence histogram from one duplicate is shown. The mean of median fluorescence of GFPmut3 expression of the duplicate samples and standard deviation were quantified using FlowJo. **(A)** 1 hour induction **(B)** 4 hours induction **(C)** 24 hours induction.

**A four-hours induction time point generates wide range of *gfpmut3* mRNA abundance for a range of aTc induction concentrations**

qPCR was used to measure *gfpmut3* mRNA abundance four hours after induction with aTc. From previous flow cytometry results (Figure 3C), since there was a significant difference in GFPmut3 protein expression levels between 1 ng/mL and 5 ng/mL aTc (Figure 3B), we added an intermediate concentration of 2.5 ng/mL aTc in an attempt to increase the number of distinct expression levels in our study. We found that the differences in mRNA abundance in the 0-10 ng/mL aTc range were small (Figure 6). Furthermore, mRNA abundance at 50 ng/mL and 100 ng/mL aTc showed no discernible difference in expression from the 20 ng/mL of aTc (Figure 6) despite the differences in protein levels (Figure 3B). Therefore, we decided to exclude 1 ng/mL and 100 ng/mL of aTc from half-life experiments. Moving forward, we decided to induce *gfpmut3* + *tetR* strains with 0, 2.5, 5, 10, 20, 50 ng/mL of aTc for 4 hours to determine mRNA half-lives.

Although we did not detect above-background fluorescence in the absence of aTc (Figure 3B), there was mRNA present at levels similar to the housekeeping gene *sigA* (Figure 6). The ability to tightly regulate expression by the *tet* repressor systems is not as clear or straightforward as may seem. Our data support the notion that there is some leakiness associated with the $P_{myc1}$ 2X *tetO* promoter as there was some expression of *gfpmut3* mRNA in the absence of the aTc inducer (Figure 6). Future studies using this system should note this limitation.

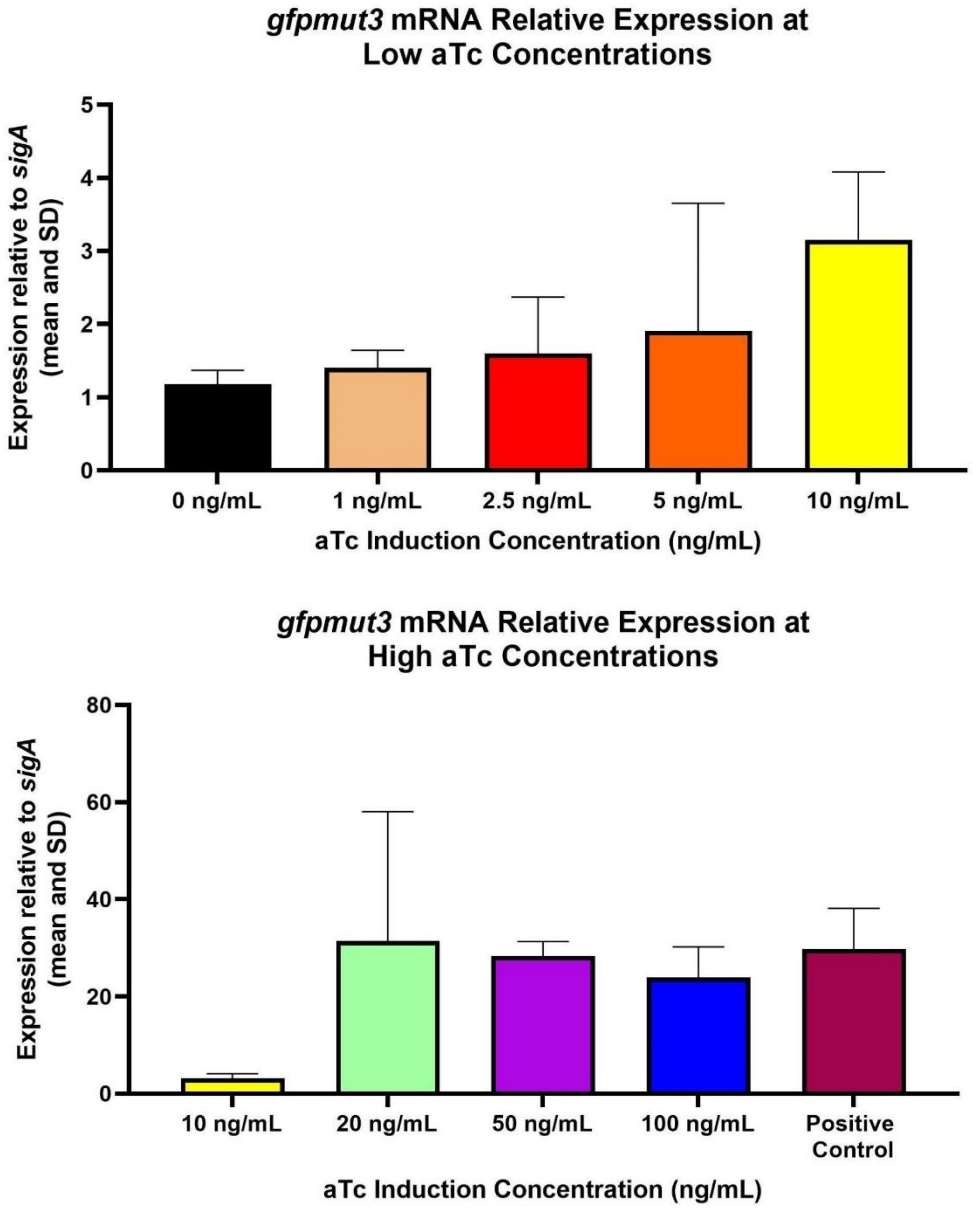



**Figure 6.** qPCR to measure expression levels of *gfpmut3* mRNA relative to *sigA*. (Top) *gfpmut3* + *tetR* strains samples were incubated for 4 hours with the lower aTc concentrations: 0 ng/mL, 1 ng/mL, 2.5 ng/mL, 5 ng/mL, and 10 ng/mL. (Bottom) *gfpmut3* + *tetR* strains samples were incubated for 4 hours with the higher aTc concentrations: 20 ng/mL, 50 ng/mL, and 100 ng/mL. The 10 ng/mL concentration was included as a reference for comparison between the mRNA expression levels of the lower and higher aTc concentrations. The positive control of *gfpmut3* only was incubated for 4 hours in the absence of aTc.

**mRNA half-lives were indeterminable due to high noise in the mRNA degradation curves**

There is a lack of consensus regarding the direction and extent of the correlation between mRNA abundance and mRNA half-life. While some studies note the presence of a weak, positive correlation in non-growing conditions (Redon et al., 2005) and log-phase (Kristoffersen et al., 2012; Chen et al., 2015), others note an inverse correlation in log-phase growth (Bernstein et al., 2002; Redon et al., 2005; Rustad et al., 2013; Esquerré et al., 2015; Nouaille et al., 2017; Sun et al., manuscript in progress). Additionally, the causality for this correlation is unknown as noted by Nouaille et al. (2017). Using a tetracycline-inducible gene expression system of *gfpmut3*, we were interested in determining if mRNA abundance affected the rate of mRNA degradation and transcriptional causality within this relationship in *M. smegmatis*. Hence, we intended to determine the half-lives of *gfpmut3* mRNA at different induction levels. We prepared four biological replicate cultures (batches A-D) that were induced with 0, 2.5, 5, 10, 20, and 50 ng/mL of aTc for 4 hours. We then treated cultures with a high dose of rifampicin to block transcription and harvested RNA after different time points: 0, 0.5, 1, 2, and 4 minutes (Figure 7 and Figure 8). In all the degradation curves obtained for *gfpmut3* mRNA and *sigA* mRNA, there was high noise and high variability between the four batches across all aTc concentrations (Figure 7 and Figure 8). We found that there was little degradation of *gfpmut3* mRNA even after 4 minutes of treatment, implying that *gfpmut3* mRNA may have a longer half-life than expected (Figure 7). Unexpectedly, there appeared to be increased mRNA abundance following rifampicin treatment for some batches. In contrast to the *gfpmut3* degradation curves (Figure 7), there was degradation of *sigA* mRNA across all aTc concentrations (Figure 8). The overall rate of *sigA* degradation over time was slower than expected (Nguyen et al., 2020). Interestingly, we noticed that for all aTc concentrations, the trend of *sigA* degradation between the four batches was not consistent (Figure 8). Batches A and C were more similar to each other, with a faster rate of *sigA* mRNA degradation, whereas batches B and D were most similar to each other and had a slower rate of *sigA* mRNA degradation (Figure 8). Due to the high noise and variability in these mRNA degradation curves, we were not able to calculate *gfpmut3* mRNA half-lives with confidence and establish the correlation between mRNA abundance and mRNA half-lives.

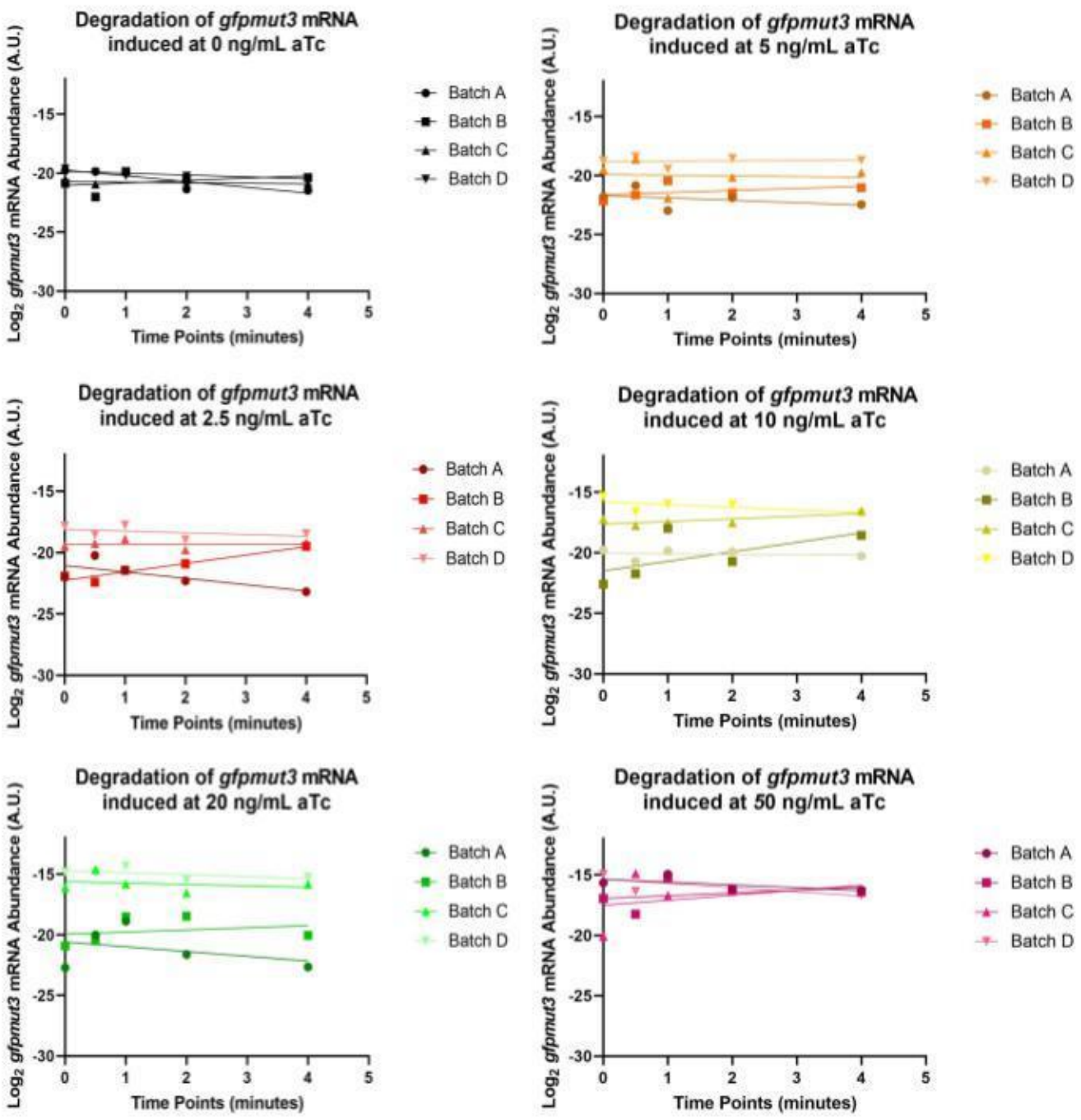



**Figure 7.** Degradation curves of *gfpmut3* mRNA induced for 4 hours at various concentrations of aTc. Four biological replicates, denoted as different batches, of *gfpmut3* + *tetR* strains were induced with 0, 2.5, 5, 10, 20, and 50 ng/mL of aTc for 4 hours and treated with rifampicin for 0, 0.5, 1, 2, and 4 minutes. mRNA abundance was measured by qPCR and $-C_t$ plotted on the y-axis as a unit of $\log_2$ abundance. Linear regression was performed on each batch using GraphPad Prism 9.

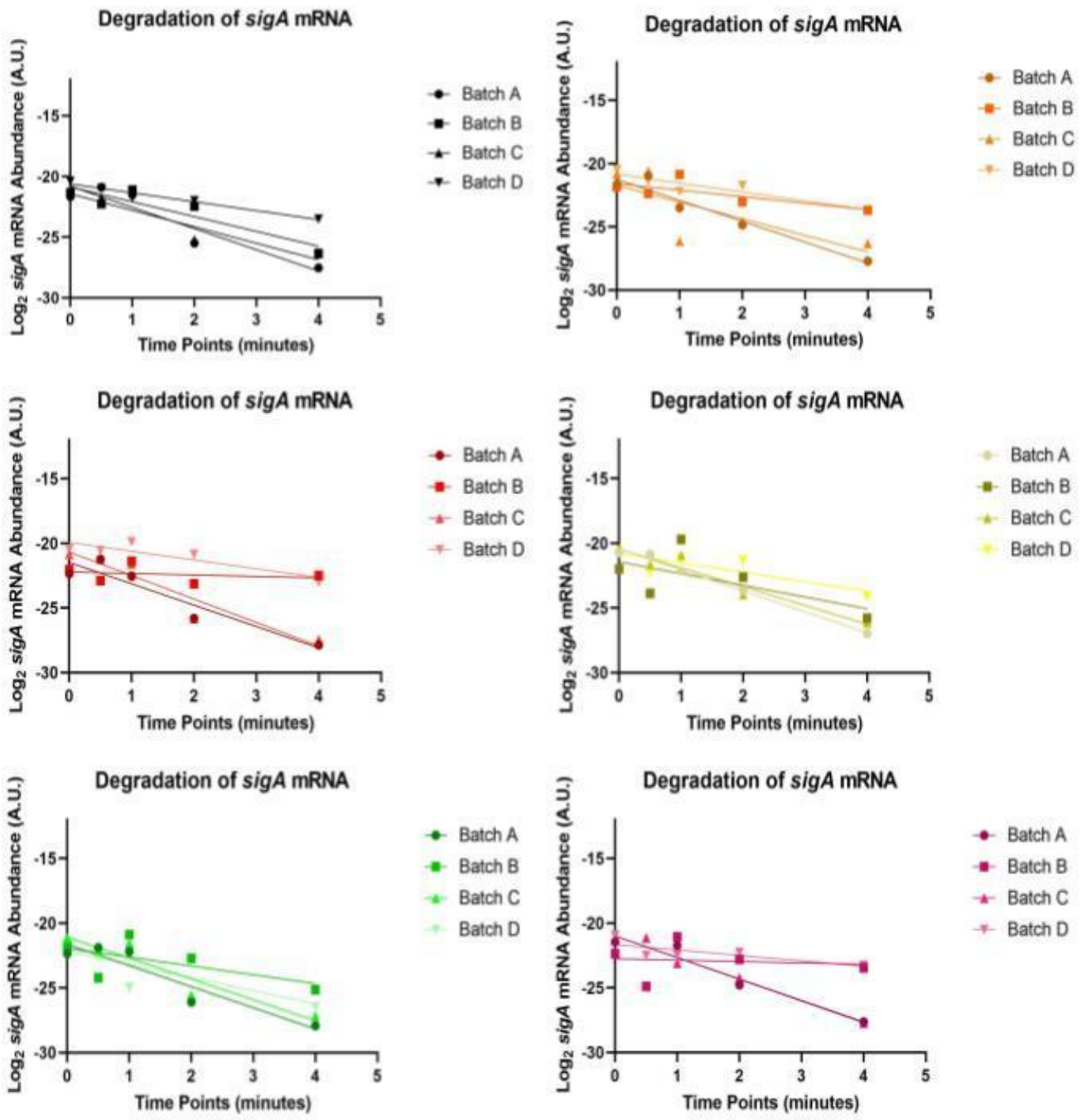



**Figure 8.** Degradation curves of *sigA* mRNA induced for 4 hours at various concentrations of aTc. Four biological replicates, denoted as different batches, of *gfpmut3 + tetR* strains were induced with 0, 2.5, 5, 10, 20, and 50 ng/mL of aTc for 4 hours and treated with rifampicin for 0, 0.5, 1, 2, and 4 minutes. mRNA abundance was measured by qPCR and $-C_t$ plotted on the y-axis as a unit of $\log_2$ abundance. Linear regression was performed on each batch using GraphPad Prism 9.

**The high variability and noise in degradation curves came from the RNA samples rather than qPCR error**

To determine a probable cause of the high noise in the mRNA degradation curves among four biological replicates (Figure 7 and Figure 8), we repeated the qPCR for all samples induced with 20 ng/mL of aTc. Prior to repeating the qPCR, the concentrations of the cDNA samples were re-measured. The amplification plots, cDNA concentrations, and $C_t$ values were compared between the two trials. Beginning with the amplification plots, we saw that Batch A and Batch B replicates had shallow curves in the first trial (Figure 9A). In the second trial, all curves were parallel to each other as expected (Figure 9B). Therefore, any technical errors in the qPCR were resolved. When looking at the amplification plot for the second trial, there appeared to be no significant difference between the *gfpmut3* curves at different time points (Figure 9B). Therefore, qPCR technique and technical errors were likely not the cause for the variability seen in the degradation curves (Figure 7 and Figure 8).

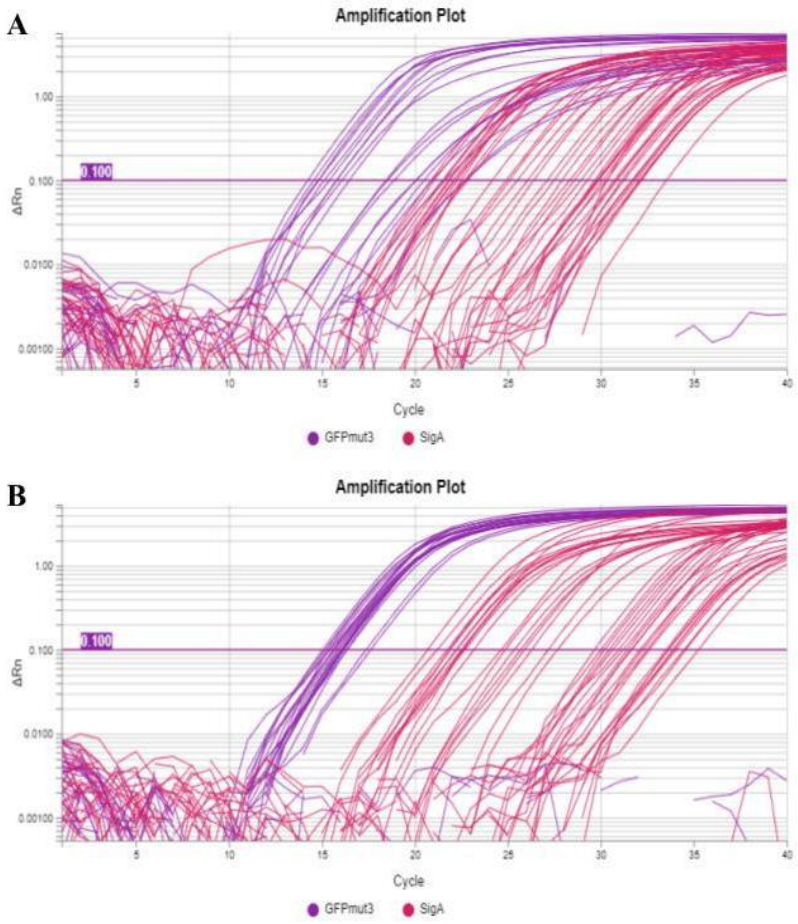



**Figure 9**. qPCR amplification plots of *gfpmut3* (purple) and *sigA* (red) with the $C_t$ of 0.1 for all four replicates induced with 20 ng/mL of aTc for 4 hours. **(A)** The amplification plot from the first qPCR trial. **(B)** The amplification plot from the second qPCR trial.

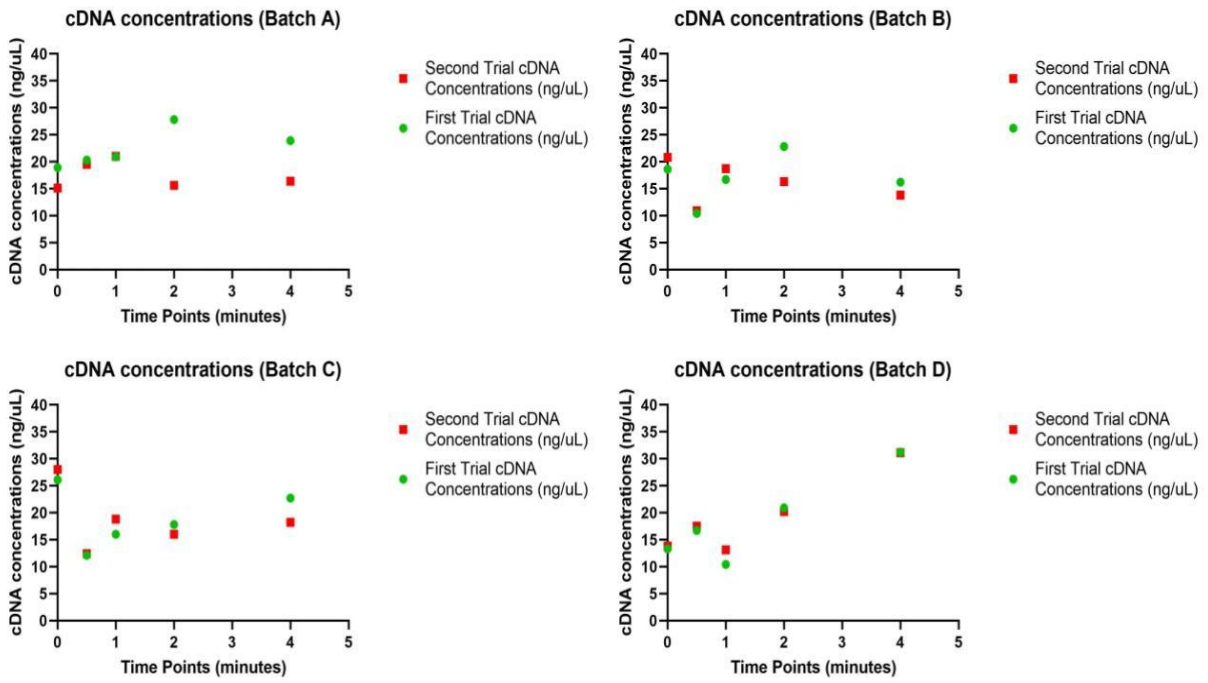



**Figure 10.** Comparison of the first trial cDNA concentrations (ng/uL) (green circle) and the second trial cDNA concentrations (ng/uL) (red square) at the different time points of rifampicin exposure for all replicates induced with 20 ng/mL of aTc for 4 hours. Each of the replicates was represented in their respective graphs: replicate batch A (top-left), replicate batch B (top-right), replicate batch C (bottom-left), replicate batch D (bottom-right).

To determine whether the cDNA samples were the cause of the variability in the degradation curves, we compared the cDNA concentrations and $C_t$ values between the two qPCR trials for the 20 ng/mL aTc induction samples. To assess cDNA concentration quantification error, we compared the first trial cDNA measurements to the second trial cDNA measurements (Figure 10). Aside from two samples in replicate batch A that were significantly different, cDNA concentrations between the first and second qPCR trials did not vary substantially (Figure 10). Therefore, we believe the noise did not originate from quantifying cDNA.

To assess potential error in cDNA dilution or the qPCR, we plotted the first trial $C_t$ values for *sigA* against the second trial $C_t$ values for *sigA* (Figure 11). Our analysis was performed using *sigA* since other lab members have measured the half-life of *sigA* previously so we had an expectation for how the results should look. When comparing the first and second trial $C_t$ values at each rifampicin exposure time point, there was no significant difference between them for all four replicates (Figure 11A and B). Therefore, the cDNA dilution and qPCR were not the cause of the high noise and variability we saw in the *sigA* degradation curves (Figure 8). When comparing the $C_t$ values of the first trial against the $C_t$ values of the second trial for *sigA*, we saw the different rifampicin exposure time points were out of expected order (Figure 11C). One would expect that as rifampicin exposure increased, the mRNA abundance for *sigA* would decrease. Thus, higher exposure time points should have higher $C_t$ values. In our experiment,

however, all four replicates showed unexpected orders in the exposure time points (Figure 11C). Thus, the source of these cDNA samples, the RNA collected from the rifampicin experiment, are the likely source of the noise and variability in the *sigA* degradation curves (Figure 8).

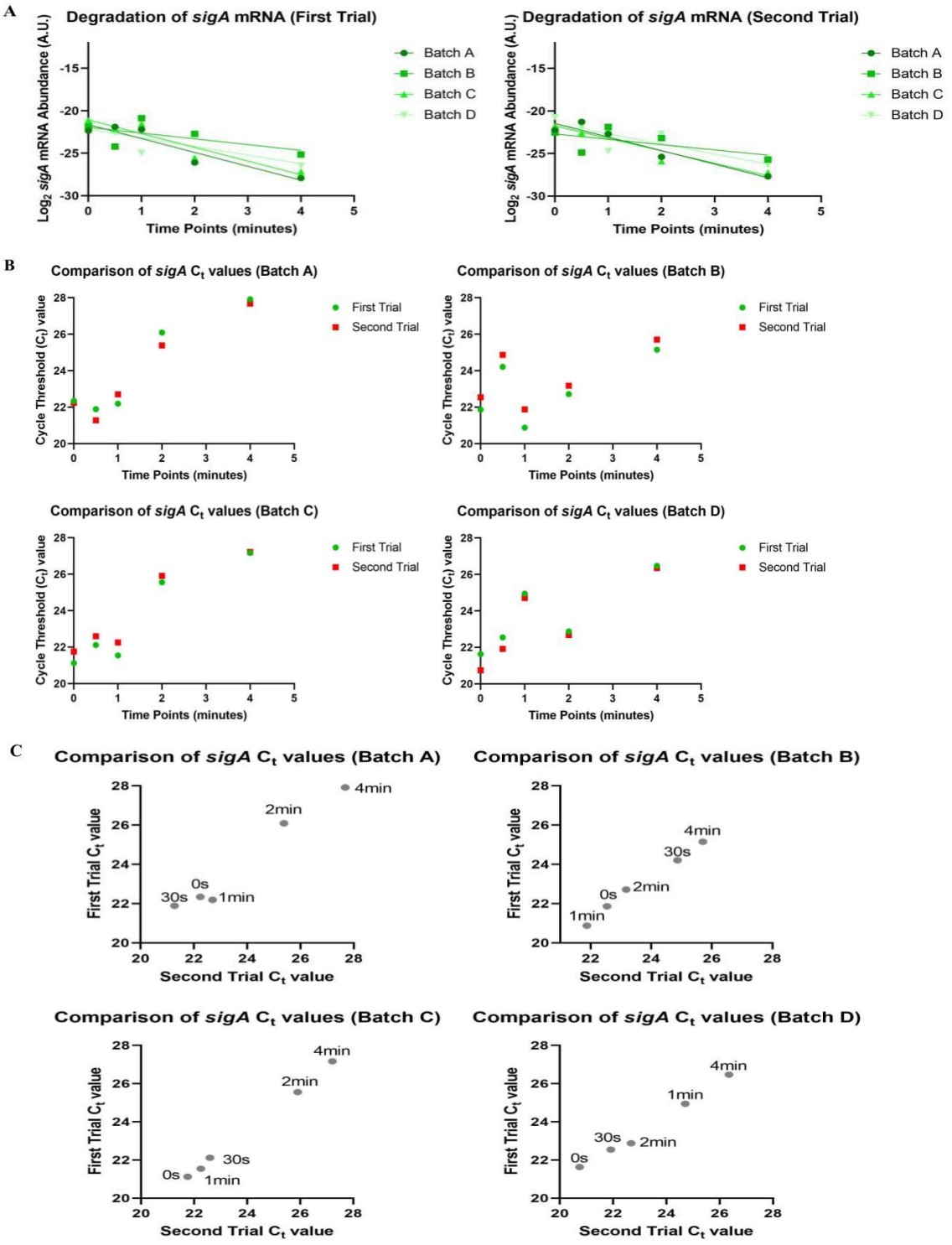



**Figure 11.** Comparing *sigA* $C_t$ values between two qPCR trials done on the same set of cDNA samples. The data represent four replicates induced at 20 ng/mL of aTc for 4 hours. **(A)** The degradation curves of *sigA* in the first trial (left) and second trial (right) **(B)** Comparison of $C_t$ values between the first (green circle) and second (red square) trials at different time points of rifampicin exposure **(C)** Comparison of $C_t$ values of the first trial against the second trial. Labels at each data point denote the rifampicin exposure time point. The four batches of replicates were graphed separately.

The RNA samples are likewise the most likely cause of the noise seen in the *gfpmut3* mRNA degradation curves as well. The degradation curves for *gfpmut3* were more variable in the first trial than in the second trial (Figure 12A). However, both trials do not show any sign of mRNA degradation (Figure 12A). When comparing the $C_t$ values between the first and second trials, there was substantial variability in replicates batch A and batch B (Figure 12B), which is consistent with the shallow curves observed in Figure 9A. We therefore focused on batch C and batch D replicates. When examining these two replicates, there was no significant difference in the $C_t$ values between the first and second trials at the different rifampicin exposure time points. We observed that the time points were all clustered and in an unexpected order (Figure 12C). The unexpected order indicates that, as with *sigA*, the RNA samples are a cause for the noise in the degradation curves of *gfpmut3* (Figure 7). The clustering indicates that *gfpmut3* mRNA did not degrade as quickly as expected (Figure 12C). The time points for rifampicin exposure were selected based on a previous study using *yfp* in *M. smegmatis* (Nguyen et al., 2020). Our data potentially indicate that the half-life of *gfpmut3* mRNA is longer than *yfp* mRNA. Hence, we suggest performing a similar half-life experiment with longer rifampicin exposure time points such as 0, 1, 2, 4, 8, 10, and 12 minutes for each aTc concentration.

To summarize, the degradation curves were too noisy to determine half-lives of *gfpmut3* mRNA. Thus, we were unable to establish a relationship between mRNA abundance and mRNA half-life. We could not conclude whether transcription is causal in this relationship either. Additionally, loss of aTc-induced GFPmut3 protein expression over time prevented us from investigating the steady-state relationship between mRNA abundance and protein abundance. Future work should consider the recommendations for studying half-life and limitations of our gene expression system in *M. smegmatis*.

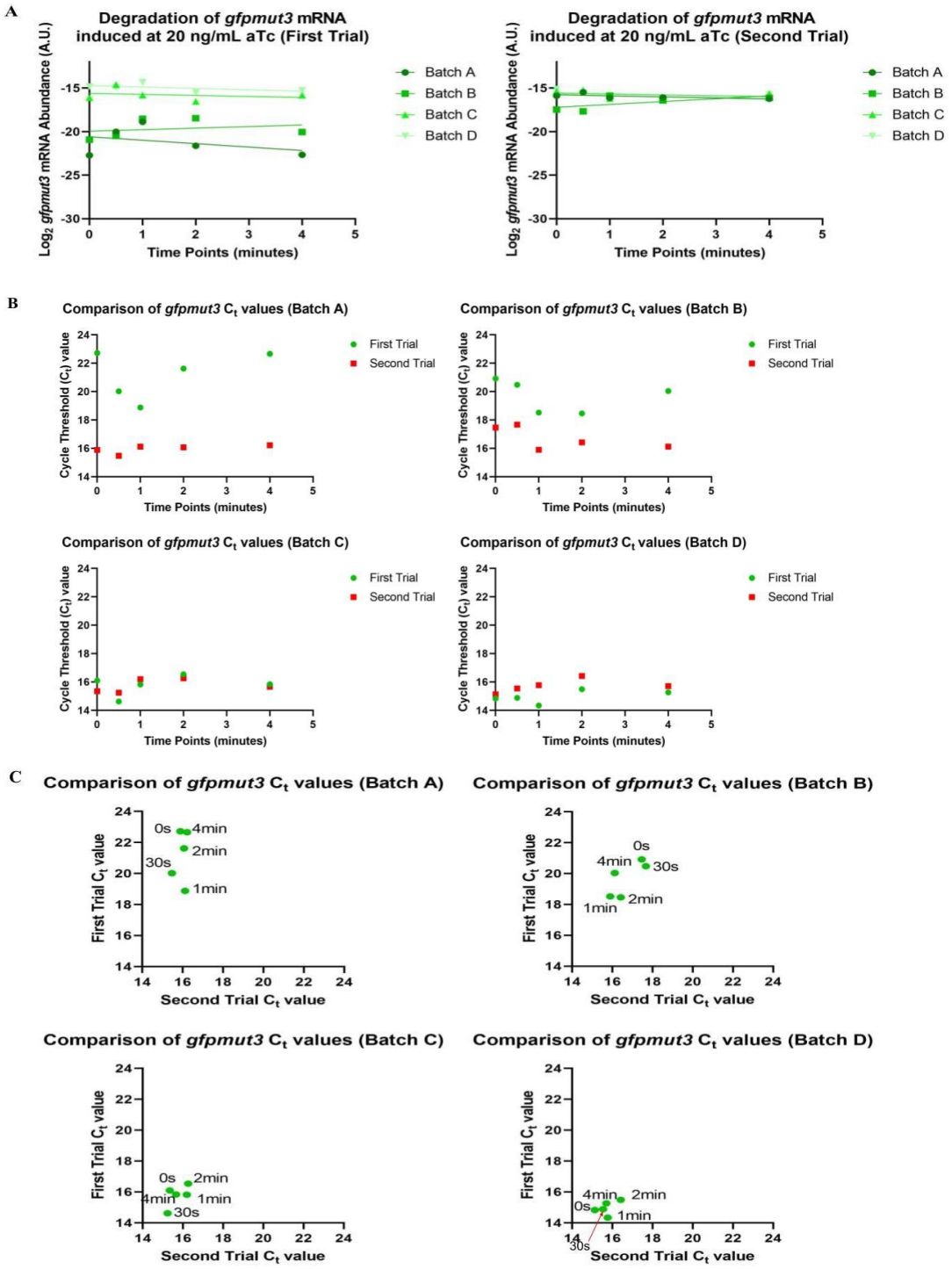



**Figure 12.** Comparing *gfpmut3* $C_t$ values between the first and second qPCR trials done on the same set of cDNA samples. The data represent four replicates induced at 20 ng/mL of aTc for 4 hours. **(A)** The degradation curves of *gfpmut3* in the first trial (left) and second trial (right) **(B)** Comparison of $C_t$ values between the first (green circle) and second (red square) trials at different time points of rifampicin exposure **(C)** Comparison of $C_t$ values of the first trial against the second trial. Labels at each data point denote the rifampicin exposure time point. The four batches of replicates were graphed separately. In batch D, the red arrow denotes the 30 second rifampicin exposure time point.

## Acknowledgements

First and foremost, we would like to thank our Biology and Biotechnology advisor Dr. Scarlet Shell. Her invaluable guidance, continuous support, and profound knowledge were invaluable to us.

We would also like to express our gratitude to Diego Vargas-Blanco, Ying Zhou, María Carla Martini, Julia Ryan, Mara Natalia Alonso, and all the members of the Shell Lab. Their help, guidance, and positivity within the lab was always helpful and appreciated.

## References


Arnvig, K. & Young, D. (2012) Non-coding RNA and its potential role in *Mycobacterium tuberculosis* pathogenesis. *RNA Biology*, (9)4, 427-436. DOI: 10.4161/rna.20105

Baumschlager, A., Rullan, M., & Khammash, M. (2020) Exploiting natural chemical photosensitivity of anhydrotetracycline and tetracycline for dynamic and setpoint chemo-optogenetic control. *Nat Communications*, 11, 3834. https://doi.org/10.1038/s41467-020-17677-5

Bernstein, J. A., Khodursky, A. B., Lin, P. H., Lin-Chao, S., & Cohen, S. N. (2002). Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proceedings of the National Academy of Sciences*, *99*(15), 9697-9702. https://doi.org/10.1073/pnas.112318199

Blokpoel, M. C., Murphy, H. N., O'Toole, R., Wiles, S., Runn, E. S., Stewart, G. R., Young, D. B., & Robertson, B. D. (2005). Tetracycline-inducible gene regulation in mycobacteria. *Nucleic acids research*, *33*(2), e22-e22.

Blum, E., Carpousis, A. J., & Higgins, C. F. (1999). Polyadenylation promotes degradation of 3′-structured RNA by the *Escherichia coli* mRNA degradosome in vitro. *Journal of Biological Chemistry*, 274(7), 4009-4016.

Boël, G., Letso, R., Neely, H., Price, W. N., Wong, K. H., Su, M., Luff, D. J., Valecha, M., Everett, K. J., Acton, B. T., Xiao, R., Montelione, T. G., Aalberts, P. D., & Hunt, J. F. (2016). Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature*, *529*(7586), 358-363. https://doi.org/10.1038/nature16509

Boldrin, F., Provvedi, R., Cioetto Mazzabò, L., Segafreddo, G., & Manganelli, R. (2020). Tolerance and persistence to drugs: a main challenge in the fight against *Mycobacterium tuberculosis*. *Frontiers in Microbiology*, *11*, 1924.

Brescia, C. C., Kaw, M. K., & Sledjeski, D. D. (2004). The DNA binding protein H-NS binds to and alters the stability of RNA in vitro and in vivo. *Journal of molecular biology*, *339*(3), 505-514.

Carroll, P., Muttucumaru, D. N., & Parish, T. (2005). Use of a tetracycline-inducible system for conditional expression in *Mycobacterium tuberculosis* and *Mycobacterium smegmatis*. *Applied and environmental microbiology*, *71*(6), 3077-3084.

Centers for Disease Control and Prevention (CDC). (1990, August 24). *Tuberculosis in developing countries*. Centers for Disease Control and Prevention. https://www.cdc.gov/mmwr/preview/mmwrhtml/00001729.htm.

Chen, L. H., Emory, S. A., Bricker, A. L., Bouvet, & Belasco, J. G. (1991). Structure and function of a bacterial mRNA stabilizer: analysis of the 5'untranslated region of ompA mRNA. *Journal of bacteriology*, *173*(15), 4578-4586. https://doi.org/10.1128/jb.173.15.4578-4586.1991

Chen, H., Shiroguchi, K., Ge, H., & Xie, X. S. (2015). Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*. *Molecular systems biology*, *11*(1), 781. https://doi.org/10.15252/msb.20145794

Connolly, L. E., Edelstein, P. H., & Ramakrishnan, L. (2007). Why is long-term therapy required to cure tuberculosis?. *PLoS medicine*, 4(3), e120. https://doi.org/10.1371/journal.pmed.0040120

Czyz, A., Mooney, R. A., Iaconi, A., & Landick, R. (2014). Mycobacterial RNA polymerase requires a U-tract at intrinsic terminators and is aided by NusG at suboptimal terminators. *MBio*, *5*(2), e00931-14.

de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M., & Vogel, C. (2009). Global signatures of protein and mRNA expression levels. *Molecular BioSystems*, *5*(12), 1512-1526

Dressaire, C., Picard, F., Redon, E., Loubière, P., Queinnec, I., Girbal, L., & Cocaign-Bousquet, M. (2013). Role of mRNA stability during bacterial adaptation. *PloS one*, *8*(3), e59059. https://doi.org/10.1371/journal.pone.0059059

Ehrt, S., Guo, X. V., Hickey, C. M., Ryou, M., Monteleone, M., Riley, L. W., & Schnappinger, D. (2005). Controlling gene expression in mycobacteria with anhydrotetracycline and Tet repressor. *Nucleic acids research*, *33*(2), e21-e21.

Emory, S. A., Bouvet, P., & Belasco, J. G. (1992). A 5'-terminal stem-loop structure can stabilize mRNA in *Escherichia coli*. *Genes & development*, 6(1), 135-148. https://doi.org/10.1101/gad.6.1.135

Esquerré, T., Moisan, A., Chiapello, H., Arike, L., Vilu, R., Gaspin, C., Cocaign-Bousquet, M., & Girbal, L. (2015). Genome-wide investigation of mRNA lifetime determinants in *Escherichia coli* cells cultured at different growth rates. *BMC genomics*, *16*(1), 1-13. https://doi.org/10.1186/s12864-015-1482-8

Goyal, R., Das, A. K., Singh, R., Singh, P. K., Korpole, S., & Sarkar, D. (2011). Phosphorylation of PhoP protein plays direct regulatory role in lipid biosynthesis of *Mycobacterium tuberculosis*. *Journal of Biological Chemistry*, *286*(52), 45197-45208.

Guet, C. C., Bruneaux, L., Min, T. L., Siegal-Gaskins, D., Figueroa, I., Emonet, T., & Cluzel, P. (2008). Minimally invasive determination of mRNA concentration in single living bacteria. *Nucleic acids research*, 36(12), e73-e73. https://doi.org/10.1093/nar/gkn329.

Hausser, J., Mayo, A., Keren, L., & Alon, U. (2019). Central dogma rates and the trade-off between precision and economy in gene expression. *Nature communications*, *10*(1), 1-15. https://doi.org/10.1038/s41467-018-07391-8

Huff, J., Czyz, A., Landick, R., & Niederweis, M. (2010). Taking phage integration to the next level as a genetic tool for mycobacteria. *Gene*, *468*(1-2), 8-19.

Korch, S. B., Contreras, H., & Clark-Curtiss, J. E. (2009). Three *Mycobacterium tuberculosis* Rel toxin-antitoxin modules inhibit mycobacterial growth and are expressed in infected human macrophages. *Journal of bacteriology*, *191*(5), 1618-1630.

Kozak, M. (2005). Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, 361, 13-37. https://doi.org/10.1016/j.gene.2005.06.037

Kristoffersen, S. M., Haase, C., Weil, M. R., Passalacqua, K. D., Niazi, F., Hutchison, S. K., Disney, B., Kolstø, A., Tourasse, N. J., Read, T. D., & Økstad, O. A. (2012). Global mRNA decay analysis at single nucleotide resolution reveals segmental and positional degradation patterns in a Gram-positive bacterium. *Genome biology*, *13*(4), 1-21. https://doi.org/10.1186/gb-2012-13-4-r30

Kwon, T., Huse, H. K., Vogel, C., Whiteley, M., & Marcotte, E. M. (2014). Protein-to-mRNA ratios are conserved between *Pseudomonas aeruginosa* strains. *Journal of proteome research*, *13*(5), 2370-2380.

Lee, P. S., Shaw, L. B., Choe, L. H., Mehra, A., Hatzimanikatis, V., & Lee, K. H. (2003) Insights Into the Relation Between mRNA and Protein Expression Patterns: II. Experimental Observations in *Escherichia coli*. *Wiley Periodicals Inc.*, 1, 2-9. DOI: 10.1002/bit.10841

Lenz, G., Doron-Faigenboim, A., Ron, E.Z., Tuller, T., & Gophna, U. (2011) Sequence Features of *E. coli* mRNAs Affect Their Degradation. *PLoS ONE*, 6(12): e28544. https://doi.org/10.1371/journal.pone.0028544

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

Liu, Y., Wu, N., Dong, J., Gao, Y., Zhang, X., Mu, C., Ningsheng, S.,& Yang, G. (2010). Hfq is a global regulator that controls the pathogenicity of Staphylococcus aureus. *PLoS One*, *5*(9), e13069.

Minch, K., Rustad, T., & Sherman, D. R. (2012). *Mycobacterium tuberculosis* growth following aerobic expression of the DosR regulon. *PLoS One, 7*(4), e35935.

Nguyen, T. G., Vargas-Blanco, D. A., Roberts, L. A., & Shell, S. S. (2020). The impact of leadered and leaderless gene structures on translation efficiency, transcript stability, and predicted transcription rates in *Mycobacterium smegmatis*. *Journal of bacteriology*, 202(9), e00746-19. https://doi.org/10.1128/JB.00746-19

Nouaille, S., Mondeil, S., Finoux, A. L., Moulis, C., Girbal, L., & Cocaign-Bousquet, M. (2017). The stability of an mRNA is influenced by its concentration: a potential physical mechanism to regulate gene expression. *Nucleic acids research*, 45(20), 11711-11724. https://doi.org/10.1093/nar/gkx781

O'Hara, E. B., Chekanova, J. A., Ingle, C. A., Kushner, Z. R., Peters, E., & Kushner, S. R. (1995). Polyadenylylation helps regulate mRNA decay in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 92(6), 1807–1811. https://doi.org/10.1073/pnas.92.6.1807

Pato, M. L., Bennett, P.M., & Von Meyenburg, K. (1973) Messenger Ribonucleic Acid Synthesis and Degradation in *Escherichia coli* During Inhibition of Translation. *Journal of Bacteriology*, 116(2), 710-718. https://journals.asm.org/doi/epdf/10.1128/jb.116.2.710-718.1973.

Politi, N., Pasotti, L., Zucca, S., Casanova, M., Micoli, G., Cusella De Angelis, M. G., & Magni, P. (2014). Half-life measurements of chemical inducers for recombinant gene expression. *Journal of biological engineering*, 8(1), 1-10.

Prax, M., & Bertram, R. (2014). Review: Metabolic aspects of bacterial persisters. *Frontiers in cellular and infection microbiology*, 4, 148. https://doi.org/10.3389/fcimb.2014.00148

Raghavan, S., Manzanillo, P., Chan, K., Dovey, C., & Cox, J. S. (2008). Secreted transcription factor controls Mycobacterium tuberculosis virulence. *Nature*, *454*(7205), 717-721.

Redon, E., Loubière, P., & Cocaign-Bousquet, M. (2005). Role of mRNA stability during genome-wide adaptation of *Lactococcus lactis* to carbon starvation. *The Journal of biological chemistry*, 280(43), 36380–36385. https://doi.org/10.1074/jbc.M506006200

Ren, G. X., Guo, X. P., & Sun, Y. C. (2017). Regulatory 3' Untranslated Regions of Bacterial mRNAs. *Frontiers in microbiology*, 8, 1276. https://doi.org/10.3389/fmicb.2017.01276

Riba, A., Di Nanni, N., Mittal, N., Arhné, E., Schmidt, A., & Zavolan, M. (2019). Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. *Proceedings of the national academy of sciences*, 116(30), 15023-15032. https://doi.org/10.1073/pnas.1817299116

Rustad, T. R., Minch, K. J., Brabant, W., Winkler, J. K., Reiss, D. J., Baliga, N. S., & Sherman, D. R. (2013). Global analysis of mRNA stability in *Mycobacterium tuberculosis*. *Nucleic acids research*, 41(1), 509-517. https://doi.org/10.1093/nar/gks1019

Russel, J. B., & Cook, G. M. (1995). Energetics of Bacterial Growth: Balance of Anabolic and Catabolic Reactions. *Microbiological Review*, 59(1), 48-62. https://doi.org/10.1128/mr.59.1.48-62.1995

Saito, K., Green, R., & Buskirk, A. R. (2020). Translational initiation in *E. coli* occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing. *eLife*, 9:e55002 DOI: 10.7554/eLife.55002

Sinha, K. M., Stephanou, N. C., Gao, F., Glickman, M. S., & Shuman, S. (2007). Mycobacterial UvrD1 is a Ku-dependent DNA helicase that plays a role in multiple DNA repair events, including double-strand break repair. *Journal of Biological Chemistry*, *282*(20), 15114-15125.

Sinha, D., Matz, L. M., Cameron, T. A., & De Lay, N. R. (2018). Poly (A) polymerase is required for RyhB sRNA stability and function in *Escherichia coli*. *Rna*, 24(11), 1496-1511. https://doi.org/10.1261/rna.067181.118

Stouthamer, A. H. (1979). The search for correlation between theoretical and experimental growth yields, p. 1–47. In J. R. Quayle (ed.), *International Review of biochemistry and microbial biochemistry*, vol. 21. University ParkPress, Baltimore.

Taniguchi, Y., Choi, P. J., Li, G. W., Chen, H., Babu, M., Hearn, J. Emili, A., & Xie1,S. (2010). Quantifying *E. coli* Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *SCIENCE*, 329, 533-538. 10.1126/science.1188308.

Timmermans, J. & Van Melderen, L. (2010). Post-transcriptional global regulation by CsrA in bacteria. *Cellular and molecular life sciences*, 67(17), 2897-2908. https://doi.org/10.1007/s00018-010-0381-z

Terai, G. & Asai, K. (2020). Improving the prediction accuracy of protein abundance in *Escherichia coli* using mRNA accessibility. *Nucleic Acids Research*, 48(14), Page e81. https://doi.org/10.1093/nar/gkaa481

Tuller, T., Waldman, Y. Y., Kupiec, M., & Ruppin, E. (2010). Translation efficiency is determined by both codon bias and folding energy. *PNAS*, 107(8), 3645-3650. https://doi.org/10.1073/pnas.0909910107

Vargas-Blanco, D. A., Zhou, Y., Zamalloa, L. G., Antonelli, T., & Shell, S. S. (2019). mRNA degradation rates are coupled to metabolic status in *Mycobacterium smegmatis*. *MBio*, 10(4), e00957-19. https://doi.org/10.1128/mBio.00957-19

World Health Organization (WHO). (2020, October 14). *Tuberculosis (TB)*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/tuberculosis.

# Part II

A Prototype Biovisualization of Gene Expression Changes in *Mycobacterium tuberculosis* Metabolic Pathways

# A Prototype Biovisualization of Gene Expression Changes in *Mycobacterium tuberculosis* Metabolic Pathways

*Major Qualifying Project*


Written By:

ADRIAN ORSZULAK

Advisor:

DR. LANE HARRISON



A Major Qualifying Project
WORCESTER POLYTECHNIC INSTITUTE

Submitted to the Faculty of the Worcester Polytechnic Institute in partial fulfillment of the requirements for the Degree of Bachelor of Science in Bioinformatics and Computational Biology.

AUGUST 25, 2021 - APRIL 28, 2022

## Abstract

Metabolic pathways are complex, multi-structured organizations of proteins and metabolites that promote the continuation and survival of a living system. Through regulation of the proteins that compose these pathways, a robust system enables greater survival in a variety of stresses and conditions. *Mycobacterium tuberculosis*, the causative agent of tuberculosis, is one such organism. This makes treatment of the pathogen difficult. Generating and using biovisualizations is important for culminating a greater understanding of *M. tuberculosis* and its mechanisms of regulating gene expression. While there are an abundance of biovisualizations mapping metabolic pathways and protein-protein interactions, there is little work involving the ability to visualize changes in protein expression and impacts of regulations on systems. Additionally, many of these visualizations and other existing visualization tools are based on reference pathways, and are not specific to *M. tuberculosis*. In this paper, a prototype biovisualization of *M. tuberculosis* metabolic pathways was constructed consisting of three panels: one visualizing metabolic pathways, one displaying protein-protein interactions, and one showing changes in gene expression from a user uploaded file. The tasks and design components were made based on interviews with *M. tuberculosis* researchers. This work highlights a prospective idea for future metabolic biovisualization work.

# Acknowledgements

I would like to thank my Bioinformatics and Computational Biology advisor Dr. Lane Harrison. His support, wealth of knowledge, and guidance during the course of this project was invaluable.

I would also like to thank my Biology and Biotechnology advisor Dr. Scarlet Shell. Her support during the project and feedback about the final prototype biovisualization were incredibly helpful.

I would like to thank and extend my appreciation to Hilson Shrestha, Yiren Ding, and all graduate students part of the Worcester Polytechnic Institute data visualization community. Their help in troubleshooting errors and problems during the creation of the prototype biovisualization was invaluable.

I would like to extend my thanks and appreciation for all interviews during the goals and tasks analysis portion of this MQP. Your time, ideas, and feedback were all greatly appreciated.

Finally, I would also like to thank members of the Shell Lab. Your feedback during lab meetings and presentations over the course of the project's development was very useful.

# TABLE OF CONTENTS

# List of Figures

# List of Tables

CHAPTER 1

# INTRODUCTION

Researchers in the biological sciences have access to vast amounts of data regarding specific systems and models for study. Metabolism is and the study of metabolic pathways is one such area of interest. Metabolic pathways form the basis for understanding how organisms function and survive and respond to environmental changes[1]. These metabolic pathways are outlined by many concurrent and sequential reactions that occur within cells. These biochemical networks are difficult to visualize because of the vast amount of associated data and the impacts small changes have on the entire pathway or set of pathways[1, 2]. Metabolism and the study of metabolic pathways is, thus, one area of biological intrigue in which visualizations have the greatest need and impact.

To support this intricate web of reactions, most if not all proteins are regulated in their activity in some way. To better understand widespread changes in global gene expression, many researchers use RNA sequencing (RNA-seq) to note changes in expression in order to understand how an organism or system reacts to changes in the environment. Using visualizations, these biological researchers can examine and study the extensive amount of metabolic data and RNA-seq data to visualize where changes and adaptations are occurring. This leads to an enhanced understanding and knowledge to drive further research.

One such research community is the one that studies *Mycobacterium tuberculosis*. *Mycobacterium tuberculosis* is the pathogen that causes tuberculosis in humans [3]. With 1.3 million deaths attributed to *Mycobacterium tuberculosis* in 2020 alone, it is one of the leading causes of death by an infectious disease in the world [3]. There is an extensive amount of work and research being devoted to understanding this pathogen, its mechanism of persistence, and how to treat it [3]. Given the importance of metabolism and its regulation, visualizing such data would

prove useful for these researchers as they work to better study and find means to combat this pathogen. Using visualizations enhances the researchers understanding and to draw inferences regarding their experimental systems and models for a greater push of scientific understanding.

Dang, et al. (2017) and Murray, et al. (2016) have addressed important task taxonomies and design elements for creating protein-protein interaction and metabolic pathway visualizations. These findings have not been examined for researchers studying *Mycobacterium tuberculosis* specifically. Additionally, there is an extensive amount of bioinformatics data and information regarding metabolic pathways that has been generated and are commonly used. Many of these commonly used resources are limited in their representations of specific organisms, lack of gene expression feature change, or limited by some other external factors (Figure 2.1 and figure 2.2 and Figure 2.3). These are key components for many researchers who are trying to elucidate the role of a particular protein or note differences between two or more strains.

Building on the work of Murray, et al. (2016) and Dang, et al. (2017), interviews with researchers studying *M. tuberculosis* and its model organism *Mycolicibacterium smegmatis* were conducted to build a better understanding of the tasks important for a prototype metabolic visualization of *M. tuberculosis*. The responses were recorded and specific tasks that a visualization should accomplish were identified. In regard to the prototype metabolic visualization, the final prototype was created in React.js using d3.js, node.js, and express.js. The final visualization consists of three panels. The first panel visualizes one pathway selected by the user as either a network of connecting proteins or a more traditional view of protein to molecule to protein connections. Transcription factors and other pathway connections are togglable. The second panel is displayed once a protein from the first is selected, a protein's protein-protein interactions are displayed and interaction designated by a color legend. The third and final panel allows a user to upload a file containing a protein or transcription factor and their associated fold change from RNA-seq data. This data is compiled into tables displayed on the panel and is incorporates visually within the first, pathway panel. Broadly, the aim of this work was two-fold. The first goal was to compile a list of the important tasks for a metabolic pathway visualization of *M. tuberculosis* from researchers working to understand the pathogen. The second goal was to create a prototype visualization of the metabolic pathway visualization of *M. tuberculosis*.

Thus, the contributions of this project are as follow:

- A compiled data set of two pathways and one collected set of proteins grouped together by similar function with a number of dimensions created specifically for this visualization
- A list of notable tasks relevant to a metabolic pathway visualization of *M. tuberculosis* from researchers studying *Mycobacterium tuberculosis*
- A prototype biovisualization of the metabolic pathways of *Mycobacterium tuberculosis* using d3.js that includes:

- The ability to visualize two pathways and one collected set of proteins grouped together by similar function
- The ability to visualize the protein-protein interactions of each protein in the visualized pathway
- The ability to upload a file containing a list of proteins or transcription factors and their associated fold changes of expression generated from RNA-seq data that is then incorporated onto the pathways being visualized

CHAPTER 2

# BACKGROUND

## 2.1 Importance of Visualizing Metabolic Pathways and Gene Expression Changes

Metabolic pathways are complex, biological systems that are central in the survival of a variety of organisms. Each pathway consists of several different interacting proteins and metabolites. Researchers organize these components into pathways to better understand their relation and map their impacts on living systems. These pathways are multivariate and contain numerous associated datapoints, inherent through the nature of these living systems and expressed in the multivariate statistical analysis that is taken to analyze them[1, 4, 5]. When studying the impacts of different protein-coding genes, groups of protein-coding genes, or their gene products, expression is often varied to determine their role in their specific living system. These changes may have great implications on the system and its survival by impacting one or more metabolic pathways. These changes in protein and metabolite concentrations are known to easily perturb and affect metabolic systems[2]. In the context of gene expression changes, living systems are easily and potentially significantly affected by those changes. Understanding those changes, in a neighborhood and global context, is important to researchers. In the context of *M. tuberculosis*, the changes to its metabolism are central in the bacteria's ability to cause disease, survive various stressors, and become tolerant to drug exposure[6]. Given the multivariate nature, number of interactions, and significance of changes that are present in a metabolic pathway, visualizing metabolic pathways and their effects will aid researchers in their efforts to better understand the systems they study. For researchers studying *M. tuberculosis*, a visualization tool that can map gene expression changes to the metabolic pathways of the organism would understand the

various aspects of *M. tuberculosis* pathogenesis, persistence, and survivability.

## 2.2 Goals and Task Taxonomy Regarding Metabolic Pathway Visualizations

It is important to know what members of the scientific community and those seeking to use this visualization in order to understand the goals and tasks of a visualization. In a general approach, Lee, et al. (2006) describes a number of tasks for different types of network visualizations. In an attribute-based network, nodes should have specific values that allow for users to act on one or more of the following tasks: filtering the nodes, computing data information, determining a range, or characterize the distribution of nodes for and over a given dataset[7]. For metabolic pathway biovisualizations, it could be argued that noting correlations could also be vital for attribute-based nodes, using Lee, et al. (2006) as a reference. Links should be specific and defined by some relationship between one or more nodes[7]. This general view is, as mentioned, not specific to biovisualizations or metabolic pathway biovisualizations.

A couple of studies have previously examined a list of task taxonomies related to metabolic pathway visualizations and protein-protein interaction visualizations. In the former, Murray, et al. (2016) interviewed many different biomedical researchers to create a list of visualization task taxonomy for metabolic pathway biovisualizations. Highlighted from their analysis are three categories of tasks: attribute associated, relationship associated, and data curation associated[8]. Attribute associated tasks focused on the multivariate nature of the data, the ability to compare data between pathways, denoting experimental evidence for each entry, and noting the level of confidence between relationships[8]. Relationship associated tasks highlighted an understanding of the relationships and interactions between different pathway components in terms of the reaction, directionality, grouping, casualty, and feedback[8]. Finally, data curation tasks focus on modifying and updating the dataset in order to match new findings, discoveries, and clarification of protein functions and interactions in order to maintain accuracy in the biovisualization[8]. Current implementations or methods for implementing each of these tasks was provided[8].

In the latter, protein-protein interaction, Dang, et al. (2017) highlighted some key tasks for a solely protein-protein interaction visualization in cancer cells. Though specific to cancerous cells, these insights can extend to protein-protein interaction visualizations related to biomedical research data. These tasks focused on presenting accurate and relevant data that allowed users to investigate protein interactions and protein neighborhoods. Additionally, pointing out relevant and conflicting literature was a key task identified in their work[9]. Overall, these previous works serve as a great resource for building visualizations incorporating both metabolic pathways and

protein-protein interactions. No specific insight has been examined in researchers studying *M. tuberculosis* or other bacterial species. Investigating these individuals could provide insights into tasks, goals, and features specific to this and other related communities of researchers and their research goals.

## 2.3 Existing Metabolic Pathway Databases for *Mycobacterium tuberculosis*

There are a number of databases that curate and store metabolic information for *M. tuberculosis*. When discussing proteins, the UniProtKB database, a collection of protein entries for most organisms, is commonly used[10, 11]. Each protein entry lists a number of structural, functional, and taxonomic data associated with the protein of interest[10, 11]. This makes identifying proteins easy. The database has many annotations to other databases. For example, protein-protein interactions for each entry direct the user to the STRING network[10, 11]. There are not many features specific to visualizing pathways or noting gene expression changes with this database.

The Systems Bio MtB database is a large-scale, multi-omics curated dataset that seeks to generate a predicted network for gene regulation in *M. tuberculosis*, especially for host-pathogen interactions[12]. Each gene is given an entry that denotes its sequence data, regulation and co-regulation overview, and existing quantitative proteomics data [12]. Annotations to other databases are also included for metabolic pathway and protein-protein interaction data[12]. The most notable feature of this database is its integration of regulation data and information between transcription factors and proteins. Users are able to better define where changes of gene expression can occur and which specific regulation elements, whether that be proteins or transcription factors, are likely to be causing any noted changes[12]. No user data is able to be uploaded to entries within the database.

The Pathosystems Resource Integration Center (PATRIC), is a database specifically created for collecting, organizing, and integrating data for use in biomedical research against bacterial infectious diseases[13]. Taking a more global approach, this database highlights large-scale transcriptomics, protein-protein interactions, sequence analysis, protein family grouping, and metabolic pathways[13]. Using experimental evidence, interactions between proteins and their effects are noted. Large-scale transcriptomics data highlights changes in expression indicated through the literature. No direct changes in expression can be manipulated or inputted to any of the sites directly. The metabolic pathway collection uses what appear to be hand-drawn reference pathways designed by the Kyoto Encyclopedia of Genes and Genomes Database (KEGG) onto which proteins known to be found in *M. tuberculosis* are mapped[13]. Proteins can then be

selected, but the goal and main function of this database appears to be storing and presenting the data to the user.

The BioCyc database is the most extensive of the databases examined, in regard to data and features. Proteins from an extensive number of organisms, *M. tuberculosis* included, can be investigated for sequence, function, and structural data[14]. In conjunction with the database, BioCyc includes a metabolic map for a number of organisms[14]. This map organizes reactions into specific pathways where applicable[14]. As opposed to other visualizations discussed, the reference metabolic map is based on that particular organism rather than one, general reference metabolic pathway where applicable[14]. In this particular network, nodes of varying shapes represent different components of a reaction whereas links represent the reactions themselves[14]. A zoom feature is used to examine the metabolic map[14]. Using the Omics tool, users are able to upload a file, specifying gene or protein name and a numerical value, to see the degree of gene expression change for metabolic experiments particularly[14, 15]. Limiting a user's investigation, however, is a subscription to use the database[14]. The MetaCyc database, a derivative of the BioCyc database, has many features specific to metabolism and metabolic pathways[16]. From this database, protein entries can be examined to obtain sequence, function, and structural data about a specific protein. This database does not feature a metabolic map or similar functionality.

In summary, these databases are useful and robust collections of curated data that help researchers easily reference that available data. Many of the investigated databases have some degree of specificity to *M. tuberculosis* and are useful to easily find proteins of interest. Two, specifically BioCyc and PATRIC, present metabolic pathways in a visualized way to some degree[13, 14]. In regard to BioCyc, this database incorporates a metabolic map tool that allows users the ability to note experimental gene expression changes from user inputted data, all based on a *M. tuberculosis* reference pathway[14]. BioCyc, though a useful tool that does address metabolic gene expression changes, is often limited in use for its not complete representation of *M. tuberculosis* metabolic pathways and subscription based service[14]. Aside from BioCyc, other investigated databases are limited in their representation of protein-protein interactions and gene expression changes. These conclusions are summarized in Figure 2.1. Thus, facilitating the creation of a prototype biovisualization to denote differences in gene expression in the *M. tuberculosis* would serve to benefit the larger tuberculosis research community. Additionally, creating a prototype biovisualization would present the novel idea incorporating a metabolic pathway visualization with the ability to denote gene expression changes and protein-protein interactions specific to *M. tuberculosis*.

| | Visualize Pathway | Specific to *M. tuberculosis* | Identify Particular Proteins | Allow Uploading Data | Indicate Impacts of Gene Expression Changes | Protein-Protein Interactions |
|---|---|---|---|---|---|---|
| Systems Bio MtB | | | | | | |
| BioCyc | | | | | | |
| MetaCyc | | | | | | |
| PATRIC | | | | | | |
| UniProtKB | | | | | | |

**Figure 2.1:** ***A summary of the features present and not present in existing databases of M. tuberculosis containing metabolic pathway information.*** *The "Visualize the Pathway" feature refers to the database containing a visualization of a metabolic pathway. The "Specific to M. tuberculosis" feature refers to the database containing data specific and inherent to M. tuberculosis and not simply orthologous information. The "Identify Particular Proteins" feature refers to the database allowing users to find and search for proteins or pathways. The "Allow Uploading Data" feature refers to the database allowing users to enter data in some form, regardless if that data is specific to gene expression datasets or simply a list of proteins to highlight, The "Indicate Impacts of Gene Expression Changes" feature refers to the database providing information related to gene expression changes, with or without user file uploading. The "Protein-Protein Interactions" feature refers to the database indicating interactions between proteins and the nature of those interactions. Features that are present in a specific biovisualization tool are noted in green. Features that are present in a specific biovisualization tool to some capacity are noted in brown. Features that are not present in a specific biovisualization tool are white. Any component listed as annotation to another database is not included as the database containing the feature or containing the feature to some capacity.*

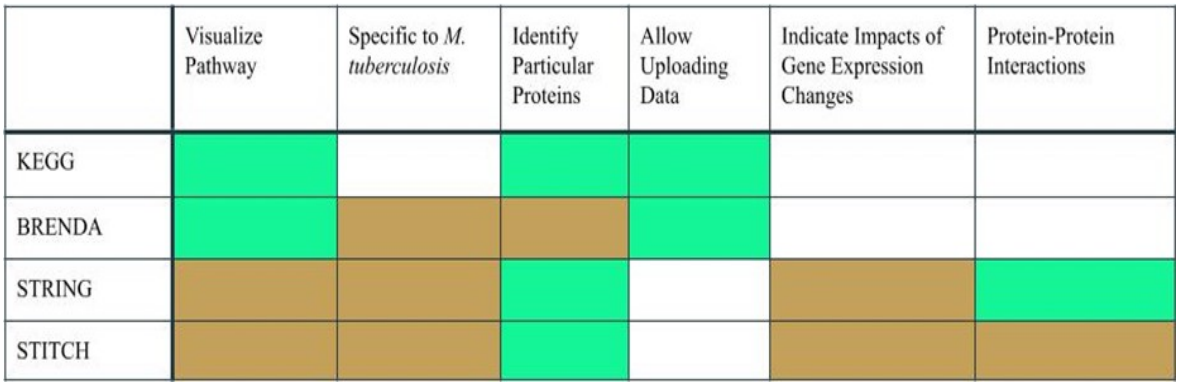
## 2.4 Existing Metabolic Pathway Visualizations of *Mycobacterium tuberculosis*

There are a handful of different, dedicated visualizations of metabolic pathways that already exist and include information and data for *M. tuberculosis*. The KEGG database contains one such visualization[17–19]. Using a base reference map of metabolic pathways, the KEGG visualization highlights proteins or orthologous proteins of a specific organism that construct the metabolic pathways for that organism or particular isolate of that organism[17–19]. The user is then able to highlight specific pathways of interest and specific proteins[17–19]. Clicking on a node or link in the pathway directs the user to another page with more details regarding a particular protein or metabolite[17–19]. The KEGG biovisualization is robust and extensive in its representation of metabolic pathways for a number of organisms, but lacks specificity as a result[17–19]. Certain

proteins and protein collections specific to *M. tuberculosis* are absent from the base reference map and are not represented. Additionally, while the pathway does allow users to upload data, this upload feature is restricted to highlighting specific proteins in the reference protein only[17–19]. No changes nor notes of gene expression changes can be submitted alongside a protein in question.

Another such biovisualization is found in the BRENDA database[20]. BRENDA uses its extensive source of data and a reference map of the metabolic pathways in order to construct the metabolism of many different organisms[20]. A number of different strains and isolates of *M. tuberculosis* are represented within this metabolic pathways biovisualization[20]. Using a zoom feature, a user is able to get a greater insight into each specific pathway, being able to denote proteins, molecules, and some cofactors involved in the many reactions of a metabolic pathway[20]. Once an organism is selected, the visualization generates a coverage score and highlights proteins or orthologous proteins within the reference pathways that are found in the organism of interest[20]. Clicking on a protein node directs the user to another page with more information specific to that protein of interest[20]. While extensive, the BRENDA biovisualization trades specificity for scope per its intended goal, and is not specific to *M. tuberculosis* or any particular organism[20]. Thus, certain protein or protein collections are not present in the visualization. The upload feature allows users to highlight specific proteins only[20].

While not specific to displaying the metabolic pathways of any particular organism, the STRING and STITCH visualizations provide an interesting viewpoint into the metabolism of an organism of interest through protein-protein interactions and protein-molecule interactions respectively[21–23]. In STRING, a user is able to enter a protein of interest within a particular organism to generate a network of protein-protein interactions at varying levels of confidence[22, 23]. In STITCH, a user is able to enter a protein of interest within a particular organism to generate a network of protein-protein interactions and protein-molecule interactions at varying levels of confidence[21]. Both visualizations generate these networks as representations of the confidence of these interactions rather than the specificity[21–23]. In the STRING visualization, users are able to identify physical interactions, the formation of a complex, and functional interactions, some unspecified interaction, between proteins[22, 23]. In the STITCH visualization, users are able to note the molecular action and binding affinity of protein-protein and protein-molecule interactions[21]. Inherent in their representation and robust dataset, these visualizations are able to show metabolic pathways through the interactions of proteins, but not in a style commonly associated with metabolic pathway representations[21–23]. Furthermore, this dataset allows the visualization to be robust in its representation of many different proteins for a variety of organisms without being made specific for any one, including *M. tuberculosis*[21–23]. As a representation of interactions between proteins and molecules respectively, STRING and STITCH have some degree of indicating gene expression changes using outside information, but



| | Visualize Pathway | Specific to *M. tuberculosis* | Identify Particular Proteins | Allow Uploading Data | Indicate Impacts of Gene Expression Changes | Protein-Protein Interactions |
|---|---|---|---|---|---|---|
| KEGG | | | | | | |
| BRENDA | | | | | | |
| STRING | | | | | | |
| STITCH | | | | | | |

**Figure 2.2:** ***A summary of the features present and not present in existing biovisualizations of M. tuberculosis metabolic pathways.*** *The "Visualize the Pathway" feature refers to whether the visualization visualizes a metabolic pathway. The "Specific to M. tuberculosis" feature refers to the visualization containing data specific and inherent to M. tuberculosis and not simply orthologous information. The "Identify Particular Proteins" feature refers to the visualization allowing users to find and search for proteins or pathways. The "Allow Uploading Data" feature refers to the visualization allowing users to enter data in some form, regardless if that data is specific to gene expression datasets or simply a list of proteins to highlight, The "Indicate Impacts of Gene Expression Changes" feature refers to the visualization providing information related to gene expression changes, with or without user file uploading. The "Protein-Protein Interactions" feature refers to the visualization indicating interactions between proteins and the nature of those interactions. Features that are present in a specific biovisualization tool are noted in green. Features that are present in a specific biovisualization tool to some capacity are noted in brown. Features that are not present in a specific biovisualization tool are white. Any component listed as annotation to another database or visualization is not included as the visualization containing the feature or containing the feature to some capacity.*

have no means to represent those changes within the visualization[21–23]. Likewise, no ability to upload data is present[21–23].

In summary, these biovisualizations are extensive and robust in the information and detail that they present regarding the metabolism and protein-protein interactions of a variety of organisms, including *M. tuberculosis*. The metabolic visualizations created in KEGG and BRENDA are based on a reference pathway that, while effective at displaying the majority of metabolism in *M. tuberculosis*, is not specific to that or any organism[17–20]. Thus, certain proteins and pathways specific to the organism are not able to be found within this visualization. While STRING and STITCH do include many proteins specific to *M. tuberculosis*, their goal is not to indicate the whole metabolic pathways of these proteins[21–23]. STRING and STITCH are able to indicate impacts of gene expression changes to some capacity, but the goal of these biovisualizations does not allow the user to draw inferences about how these gene expression changers impact the pathway as a whole or other pathways[21–23]. Finally, aside from the ability to highlight a set of proteins if those proteins are found in the KEGG and BRENDA biovisualizations, these visualizations do

not allow the user to upload data or datasets denoting gene expression changes[17–20]. These conclusions are succinctly detailed in Figure 2.2. Thus, creating a prototype biovisualization of *M. tuberculosis* metabolic pathways would address a novel idea of incorporating and visualizing gene expression changes in a pathway alongside protein-protein interaction data.

## 2.5 Existing Metabolic Pathway Visualization Tools

While not necessarily specific to *M. tuberculosis*, a number of visualization tools have been developed to better map, understand, and draw inferences from metabolic visualization data. A select number of other visualization tools referenced in Murray, et al. (2016) and review of the literature were chosen to understand the capabilities of these tools for use with *M. tuberculosis*. VisANT is one such visualization tool that utilizes networks to represent large scale datasets for researchers to draw inferences regarding the nature of the interactions and gene expression ratios[24–26]. Interactions with the node can reveal further increasing complexity between the interactions present[24–26]. Edgers represent different experimental methods and evidence through use of a color visual channel[24–26]. Many annotations allow the user to easily find other entries and further information more easily, especially when paired with the Predictitome database[24–26]. Layout and graph operations are determined from user input on various filters[24–26]. While not displaying metabolic pathways explicitly, the inherent displaying of interaction networks shows the metabolic relationship of proteins[24–26]. No explicit feature for denoting gene expression changes is found within this visualization tool[24–27].

Cytoscape is another visualization tool originally designed for assessing and analyzing metabolic data[28]. Using the various layout tools, different networks with different degrees of details can be shown[28]. This ranges from displaying pathways based from a public source database, protein interaction networks, and protein-protein interactions[28]. Filters are present in order to reduce or increase the magnitude of data displayed at any one time[28]. Furthermore, the users are able to upload gene expression data and draw inferences about the changes in protein expression. There is no specific data to *M. tuberculosis* regarding metabolomics data, but access to specific *M. tuberculosis* sequence and genome data is present[28].

PTools is the most extensive and robust of the available metabolic visualization tools discussed. This tool is used in conjunction with and found as a part of the BioCyc database, as aforementioned[14, 15]. Given this relationship, PTools reconstructs a metabolic pathway using BioCyc for a variety of organisms, including *M. tuberculosis*[15]. As noted with BioCyc, the representation is not entirely complete in regard to *M. tuberculosis*[14]. Using this tool, specific proteins, molecules, and reactions can be found within organized pathways, where applicable, and

in intracellular or extracellular space as they would exist in a living system[15]. A highlighting feature allows users to identify proteins easily within the visualization tool[15]. Furthermore, an omics aspect to the visualization tool permits the user to upload gene expression data and note changes in expression on the visualized pathway[15]. A scale is used to be able to denote differences in expression changes[15]. As previously mentioned with regard to BioCyc, this Omics tool does require a subscription for continued use[14].

Another metabolism visualization tool, BiologicalNetworks, incorporates the ability to upload and analyze microarray gene expression data[27]. The BiologicalNetworks visualization tool visualizes metabolic pathways, protein-protein interactions, and protein-DNA interactions through a variety of networks[27]. Through filtering features and hierarchical organization of networks within the tool, users are able to specify what is exactly visualized[27]. Using those visualized networks, users are able to find and search for proteins, pathways, common targets, common regulators of gene expression, and intersections with other pathways[27]. Furthermore, this visualization tool supports the upload of user gene expression data to reveal the impact of gene expression changes from microarray data in a static or dynamic display [27]. Capabilities to analyze and compare datasets of gene expression changes are included in this tool allow users to better understand and define the relevance of their gene expression datasets[27]. The BiolgicalNetworks tool has no inherent specificity to *M. tuberculosis* provided its sources of data integration[27].

The GeneVis visualization tool takes a different approach to understand genetic regulatory networks in prokaryotic organisms using a simulation[29]. Using a circle to represent a bacterial chromosome, smaller circular nodes representing protein-coding genes[29]. About each protein-coding gene node, activity state of a gene and whether regulatory proteins are bound to the gene are found[29]. Using this approach, regulatory and activity dynamics of these protein-coding genes are visualized[29]. The degree of detail and exact genes, with their associated protein products, are able to be toggled through the use of different views and lenses [29]. Interactions between gene-coding proteins and protein products amass to display a hierarchical system of regulation based on the chromosomal ring[29]. The impacts of gene expression shown are limited to effects of selected proteins products on genes already within the visualization as opposed to a more global approach [29]. The exact nature of a global application is not discussed. While designed for prokaryotic organisms, *Escherichia coli* is the organism of interest explored within the initial publication of this visualization tool[29]. No known work has been done using data from *M. tuberculosis* with GeneVis.

The final metabolic visualization tool discussed in this report is VANTED. The VANTED tool allows researchers to visualize biochemical data regarding proteins and metabolites in the context of metabolic pathways[30]. Using the KEGG pathways, another independent pathway

file, or a pathway built by the tool itself from the data, the visualization tool is able to display the metabolic pathways of an organism [30]. On the final constructed pathway, user experimental data can be uploaded and mapped to the proteins or metabolites of interest[30]. For each specified experiment or condition, the metabolite or protein concentration is displayed at each node which allows for easy comparison within a pathway[30]. Plotting expression across a similar timeframe can also be done within a pathway using a neuronal-network algorithm: the self-organizing map (SOM)[30]. Finally, the tool allows users to perform a correlation analysis between metabolite expression and protein expression [30]. Though there is a lack of specificity with regard to the *M. tuberculosis* metabolic pathway, VANTED may construct its own pathway if the KEGG pathway is not complete[17–19, 30]. The tool is able to represent these pathways with greater specificity to *M. tuberculosis* should such data be uploaded and used.

In summary, existing visualization tools are quite robust in their data integration and functionality. Alongside visualizing metabolic pathways to some degree, the majority of these tools are able to show indications of gene expression changes from a user created dataset. By proxy, identifying particular proteins is an inherent and basic function within all these tools. The use of different filters and lenses specific to each tool make searching easy and allow users to draw conclusions and inferences while still maintaining the multivariate nature of the data. Additionally, all visualization tools discussed provide some degree of statistical analysis to them. These visualization tools, however, are not specific to *M. tuberculosis*, providing a niche for another visualization tool designed for the *M. tuberculosis* research community. These conclusions are summarized in Figure 2.3. Through this project, integrating the aspects of a metabolic pathway visualization, a protein-protein interaction visualization, and noting changes of gene expression for the *M. tuberculosis* research community is the main goal. Many of these features and design choices, especially with regard to filtering and expressing data, should be carried into a biovisualization that this project seeks to design and create.

| | Visualize Pathway | Specific to *M. tuberculosis* | Identify Particular Proteins | Allow Uploading Data | Indicate Impacts of Gene Expression Changes | Protein-Protein Interactions |
|---|---|---|---|---|---|---|
| VisANT | | | | | | |
| Cytoscape | | | | | | |
| PTools | | | | | | |
| BiologicalNetworks | | | | | | |
| GeneVis | | | | | | |
| VANTED | | | | | | |

**Figure 2.3:** ***A summary of the features present and not present in existing tools that visualize and inform researchers about metabolic pathways.*** *The "Visualize the Pathway" feature refers to the tool containing a visualization of a metabolic pathway. The "Specific to M. tuberculosis" feature refers to the tool containing data specific and inherent to M. tuberculosis and not simply orthologous information. The "Identify Particular Proteins" feature refers to the tool allowing users to find and search for proteins or pathways. The "Allow Uploading Data" feature refers to the tool allowing users to enter data in some form, regardless if that data is specific to gene expression datasets or simply a list of proteins to highlight, The "Indicate Impacts of Gene Expression Changes" feature refers to the tool providing information related to gene expression changes, with or without user file uploading. The "Protein-Protein Interactions" feature refers to the tool indicating interactions between proteins and the nature of those interactions. Features that are present in a specific biovisualization tool are noted in green. Features that are present in a specific biovisualization tool to some capacity are noted in brown. Features that are not present in a specific biovisualization tool are white. Any component listed as annotation to another database or visualization is not included as the visualization containing the feature or containing the feature to some capacity.*

# Chapter 3

# Experimental Setup

## 3.1 Compiling the Dataset for the Prototype Metabolic Pathway Biovisualization

The pathway and other related data was gathered from a number of databases. The KEGG, PATRIC, and Mycobrowser databases were used to obtain the protein name, substrate and product details, protein connections, branch point status, and pathway connections[13, 17–19, 31]. No hypothetical proteins were included in this dataset. Protein-protein interaction data was gathered using the STRING and STITCH databases[21–23]. The MTB Network Portal database was used to obtain transcription factor information and to determine the gene name of each protein[12]. Only those interactions with the highest confidence were included. This data was compiled into a csv file for use in the final prototype visualization. The final dataset can be found in the GitHub repository alongside the final prototype visualization (https://github.com/ao-joker/V2-BCB-MQP-BioVis). All data recorded was from the *M. tuberculosis* H37Rv strain.

| Researcher Title and Research Areas |
|---|
| Title: Bioinformatics Graduate Research Assistant<br>Research: *M. tuberculosis*, *M. smegmatis*, mRNA degradation, Machine learning |
| Title: Biology and Biotechnology Graduate Research Assistant<br>Research: *M. smegmatis*, RNA-binding proteins |
| Title: Post-Baccalaureate Research Assistant<br>Research: *M. smegmatis*, esxB, ribosomal proteins |
| Title: Undergraduate MQP Research Assistant<br>Research: *M. smegmatis*, mRNA half-life, mRNA concentration |

**Table 3.1:** ***Researchers Interviewed.*** *A description of the researchers who were interviewed.*

## 3.2 Understanding and Compiling the Important Goals and Tasks for the Prototype Metabolic Pathway Biovisualization

To understand the important tasks for this prototype metabolic pathway visualization, four interviews were conducted with researchers of the *M. tuberculosis* community (Table 1). Prior to conducting the interviews, oral consent was provided. These researchers (Table 1) were asked to think about what a visual image or description of a metabolic pathway is. Then, they were asked to utilize some existing metabolic pathway visualizations and note both effective and less effective aspects or functions. At the end, interviewees were asked to consider important features and tasks after exploring existing visualizations. The final questions used in all interviews can be found in Appendix A. The goal of these interviews were to understand what aspects are great in existing data visualizations, understand what elements are lacking in existing data visualizations, to identify what are the goals for a data visualization of *M. tuberculosis* metabolic pathways would be, to note tasks that a biovisualization should accomplish, and to define some task abstractions from the interviewee's response. The responses were then recorded for use when building this prototype metabolic pathway visualization. Responses were then organized by order of occurrence to denote the importance of a task or feature.

## 3.3 Creating and Improving on the Prototype Metabolic Pathway Biovisualization

The bulk of this prototype visualization was built using JavaScript and CSS. The d3.js version 6 and node.js libraries were used to handle the structuring and display of data seen in all three constructed panels. Express.js and axios library were utilized in the file input mechanism. The entire visualization is hosted on a React.js server. As the final prototype was

being constructed, conversations regarding the type of information and exact detail choices were had with both advisors. The completed prototype can be found at the following GitHub repository:https://github.com/ao-joker/V2-BCB-MQP-BioVis.

CHAPTER 4

# RESULTS

## 4.1 Building an Understanding Regarding the Important Tasks and Goals for a Prototype Metabolic Pathway Biovisualization of *Mycobacterium tuberculosis*

Interviews with *M.tuberculosis* researchers (Table 1) were performed in order to better understand both the tasks and goals of a prototype metabolic pathway biovisualization of *Mycobacterium tuberculosis*. The insights of these interviews can be denoted into 3 distinct categories: key actions, important data features, displaying the data. The goals of the biovisualization focused on displaying a metabolic pathway without overwhelming the user and adding a mechanism to receive inputted data which would then be visualized alongside the pathway to be able to draw inferences about larger changes of gene expression. This information was important in structuring and building the prototype metabolic pathway biovisualization.

### 4.1.1 Key Actions

There were three classes of key actions that the interviewees found important: discovering data, searching options, and uploading data. Firstly, in regard to discovering data, many interviewees expressed desire to take the data presented from a visualization and make new inferences about gene expression changes. For example, one researcher noted the importance of having all kinds of data present in a given visualization for ease of access. Metabolic pathway data is inherently multivariate in nature[1, 4, 5, 8]. Using interactivity to display data per the user's

interest, such as in the use of "on-click" mechanisms and tooltips, would be useful to show all the data in a meaningful way without being too overwhelming. Additionally, doing so will provide a sense of discovery of the data and information.

Secondly, the ability to search and find data of interest was highly important to the interviewees. When provided sample metabolic pathway biovisualizations, interviewees often called attention to the degree of ease in finding particular data of interest. A biovisualization of a metabolic pathway should present some ability to find data of interest, whether it be a more global view focused on pathways or a more specific view focused on individual proteins. Many noted the limited reach and ability of existing biovisualizations to both find identifiable pathways and proteins in those pathways. Databases do have this well-received feature, but are more text and information heavy by nature. A prototype metabolic pathway biovisualization should work, thus, to combine the ability to search per pathway and per protein. With a greater variability in searching, researchers will be able to use a metabolic pathway biovisualization most effectively. Current implementations of this task would include the ability to locate, browse, and specifically search for the data in question.

The final action noted was the ability to input data to note changes in gene expression. As aforementioned, many researchers, including the interviewees, are interested in understanding how changes in expression of particular proteins may be able to impact a biological system. These changes in the expression of genes is a large area of interest that fuels research in *M. tuberculosis*. No visualizations known or used by the interviewees allowed data to be implemented and compared to the present metabolic pathway network. All interviewees expressed interest in implementing the ability to insert some data to understand how it impacts metabolic pathways. The use of different visual channels would especially be vital in this role, as users would want to notably see those differences. Current implementation would likely involve the use of an upload mechanism that then impacts the visualization as a whole, providing data and editing the visualization in some way discernable and useful to the user. Organizing the provided data in a table would be an easy solution that would be both familiar to the user and easy to reference when identifying specific changes with the greater visualization as a whole.

### 4.1.2 Important Data Features

Metabolic pathways, by their complexity, are inherently multivariate in nature. All the interviewees and researchers as a whole are well aware of this notion. Thus, when compiling a dataset to represent the *M. tuberculosis* metabolic pathway, there is a variety of data that needs to be conveyed in order to form a meaningful and useful visualization. Interviewees noted that a metabolic pathway should include not only protein to protein connections in a

pathway, but the molecules that serve as substrates and products. Doing so would provide users a greater sense of familiarity in the pathway structure. Furthermore, molecules were noted by some interviewees to be important to the understanding impacts of gene expression changes. Additionally, the interviewees noted the importance of including gene name alongside the protein name. With different identifiers, including all possible varieties will make the visualization more apt in the key actions of discovery and search. As final note, ensuring the data within the visualization was accurate and supported with high confidence was noted as essential for any good visualization. One interviewee specifically added that potentially linking proteins and their interaction to academic papers would provide an easy way to validate the visualization in high confidence and find the original literature. Given the importance of accuracy and peer-review in the scientific community, this is a logical and important feature that should be inherent to all data. Current ideas and implementations would present the confidence of a particular result alongside a particular data point. Consequently, only including data with high confidence could also address the aspect of validity and accuracy of the data. Linking the original and other potential literature, as in the case of database annotations, would address the thoughts of the aforementioned interviewee. More specifically about the pathway itself, the inclusion of connections between molecules and other pathways could prove strenuous on many devices, as in the case with the BRENDA pathway[20]. To condense, current implementation ideas could focus on providing different layouts to allow the user to investigate different aspects and perspectives of the pathway.

### 4.1.3 Displaying the Data

As in all visualizations, displaying the data is itself an intricate design choice. Apart from important data, this area delves into how the user would see and interact with the data benign visualized. Interviewees all noted that metabolic pathway data is inherently extensive in nature, with there being countless values and items to represent at one time. In one specific example, an interviewee indicated that displaying all data at one would be overwhelming, impeding the ability to draw inferences and conclusions for a visualization. With current techniques, any tabular and otherwise structured data should remain uncondensed and easy to identify. Furthermore, using tooltips and other features focused on interactivity would make the visualization more appealing and prevent the user from being overwhelmed with the vast amount of information. Additionally, implementing a layout system to show different kinds of data at once is a viable option and would allow users to adjust their perspective on their own discretion. On a final note, using the visual channels of size and color will be essential to discerning all the different kinds of data. Most interviewees agreed on this statement. Implementing color and size channels will aid the user in quickly and readily identifying differences in data values and features of the visualization.

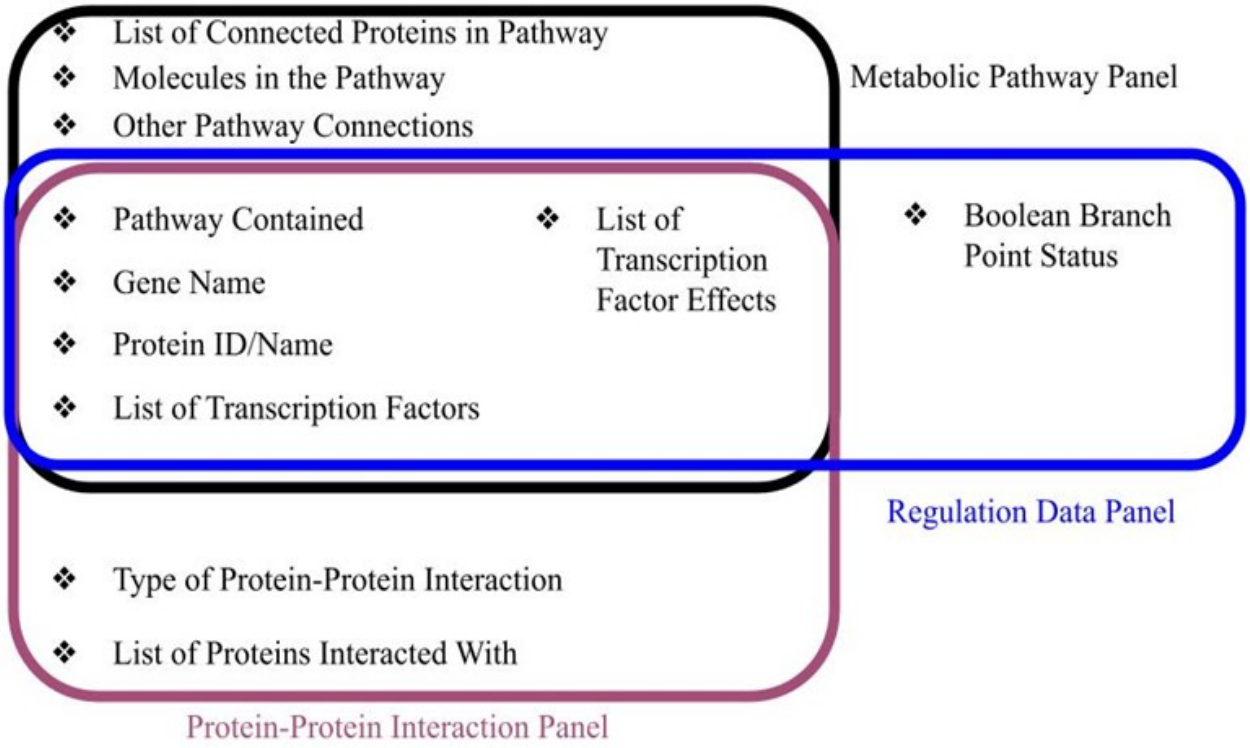



**Figure 4.1:** ***A representation of the data involved in each panel of the biovisualization.*** *The black bordered data are used in the metabolic pathway panel. The pink-red bordered data are used in the protein-protein interaction panel. The blue bordered data are used in the regulation data panel.*

## 4.2 Compiling a Dataset for the Prototype Metabolic Pathway Biovisualization of *Mycobacterium tuberculosis*

In order to create a visualization of M. tuberculosis metabolic pathway, the data for this pathway needed to be sought and organized. For the purposes of this prototype, data was compiled for the Glycolysis/Gluconeogenesis and Citrate Cycle pathways and a collection of proteins involved with T-receptor signaling grouped together for similar function. All are specific to *M.tuberculosis*. The final dataset consists of the following data categories: protein name, protein ID, pathway, proteins connected to in the pathway, molecules connected to by substrate and product relation, pathways connected to, a list of transcription factors that impact expression of a protein (if applicable), the impact of the transcription factor (if applicable), boolean status as a branch point, high confidence proteins that interact with a given protein in some way, and the nature of the interaction of those high confidence proteins. The importance of this data to each panel of the biovisualization can be found in Figure 4.1.

Protein name provides the gene name whereas protein ID denotes the name of the protein and enzyme commission (EC) number. Proteins connected to in the pathway refers to the organization of the proteins in which one protein takes in a substrate and outputs a product that a connected protein uses as a substrate in its reaction. The layout of the metabolic pathway would only show these proteins and not include molecules. Molecules connected to by substrate and product relation refers to creating a layout where both the molecules and proteins accounting for the change of substrates and molecules are included along with the proteins. The category "Pathways connected to" refers to proteins that lead into other proteins not included in the current pathway of interest. The transcription factor list and the impact of the transcription factor, denoted by a positive or negative one, indicates if other proteins impact the expression of a particular protein in a positive or negative way. Branch point status refers to proteins that can make two different products from a given substrate. Given limited information on the branch point status, this value was characterized as a boolean so as to present this information to the user. Finally, protein interactions were included in a list with a separate, corresponding list concerning the nature of those protein interactions in order to construct a protein-protein interaction network. This data indicated the type of interaction as specific as possible given the information available. Only proteins with high confidence were included. All the data here was used in the construction of the final visualized metabolic pathway, displaying a protein-protein interaction map, and indicating changes in gene expression by file upload. In order to compile this dataset, a number of databases were used to compile all the necessary data. No one specific database had all the necessary data needed for this visualization. Thus, the dataset was entirely compiled by hand, scoping through a number of databases (Found in experimental setup). Given this method of construction, the addition of papers to link users to the original literature was not included.

## 4.3 The Prototype Metabolic Pathway Biovisualization of *Mycobacterium tuberculosis*

Guided by the tasks and goals analysis and using the data collected, a prototype metabolic pathway visualization of M. tuberculosis was constructed. The final, prototype visualization consisted of three, linked panels that comprise the one larger visualization. These three panels are aptly named the metabolic pathway panel, the protein-protein interaction panel, and the regulation data panel. The final prototype visualization, its code, and the datasets used can be found at this GitHub repository: https://github.com/ao-joker/V2-BCB-MQP-BioVis. The panel colors are not a design choice, but rather a designation choice.

### 4.3.1 The Metabolic Pathway Panel

As implied for the title of this panel, this portion of the visualization focuses on displaying the actual pathway. For this visualization, only one pathway is displayed at a time. A drop down menu allows the user to alternate and change the pathway being shown. The pathway here is constructed using a force-directed network. With a particular pathway, radio buttons allow the user to alternate the pathway layout. One layout shows only connecting proteins, where each node represents a protein and the edges connect proteins in the metabolic pathway. Most reactions are in equilibrium and this visualization implies equilibrium between reactions.While this is not true for many cases, all are left undirected. The second layout includes the molecules, substrates and products, and proteins in the pathway. Additional nodes, smaller and gold in color, represent the molecules in the pathway. Edges represent the progression of the pathway as with the first layout. Regardless of the layout, each node always includes the name of the molecule or the gene name of the protein in question. A third radio button allows the user to toggle transcription factors as part of the pathway. These transcription factors are represented as smaller, red nodes with the gene name of the transcription factor positioned next to the node similarly to the proteins and molecules. This function mainly lets the user see which transcription factors impact which proteins in the pathway to draw inferences from changes in gene expression if applicable. Edges for transcription factors are acting on the connected node and are designated as undirected. A fourth and final radio button allows the user to toggle other pathway connections. If a protein is connected to another protein found in a different pathway, then a node to present that pathway is drawn. Other pathways are represented as larger, orange nodes with the name of the pathway positioned next to the node similar to the other nodes described. This allows the user to draw further inference and denote where impacts of gene expression changes one pathway may cascade to other pathways. Hovering over a particular protein reveals a tooltip with further information about the protein. This includes the gene name, protein name, and EC identification number. Clicking on a protein node will draw the protein-protein interaction network for that protein in the protein-protein interaction panel. The function of clicking on a node is subsequently reserved for creating protein-protein interaction networks. Hovering over a molecule node will just indicate the molecule's name. Hovering over a transcription factor node will display the transcription factor name and the effect it has on its connected protein. Hovering over a pathway node will display the pathway itself and indicate users to explore that pathway, if applicable, in the biovisualization. Clicking on a molecule node, transcription factor node, or pathway node will produce an alert that informs the user that no protein-protein interaction network exists for this particular node. Figure 4.2 shows a sample image the metabolic pathway panel described.

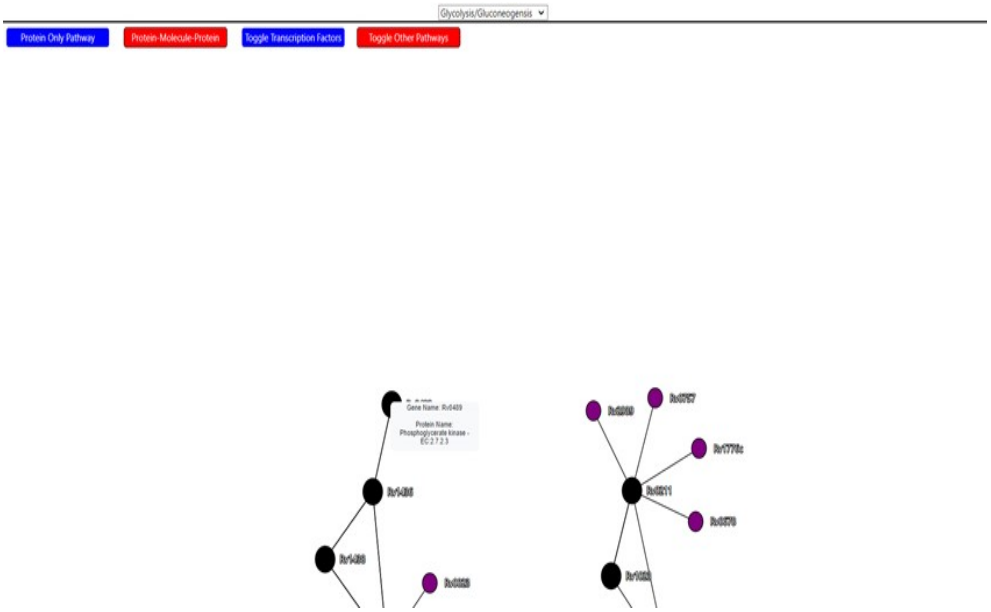



**Figure 4.2:** ***A sample image of the metabolic pathway panel.*** *This sample image displays a portion of the metabolic pathway panel within the prototype biovisualization. The panel is much larger in the prototype biovisualization. A dropdown menu selects the Glycolysis/Gluconeogenesis. The transcription factors are toggled on. The Rv0489 node is hovered over to display a tooltip.*

### 4.3.2 The Protein-Protein Interaction Panel

This panel displays the protein-protein interactions for specific proteins from the pathway. In order to display this information, a protein node must be clicked in the metabolic pathway panel. Once completed, a force-directed network is used to draw this network. The central protein node is the protein of interest whereas the surrounding protein nodes represent proteins with high confidence that interact with the central node protein. The gene names of all proteins are positioned to the side of the node. The edges represent the type of interaction between a protein of interest and other proteins. The possible interactions are constructed from the dataset, allowing any changes and additions of new interactions not previously listed to be included in the visualization at a later date. Each interaction is represented with a different color and an associated legend is drawn alongside this network to indicate what each color edge represents to the user. Any protein with more than one interaction is defined as multiple in this key. Hovering over any node will pull up a tooltip that indicates the gene name of the protein and the type of interaction. For any edge marked as having multiple interactions, the tooltip will provide further clarity on what those interactions are. Hovering over the central protein node will indicate that there is no interaction of the protein on itself. Figure 4.3 displays a sample image of the protein-protein interaction panel described.

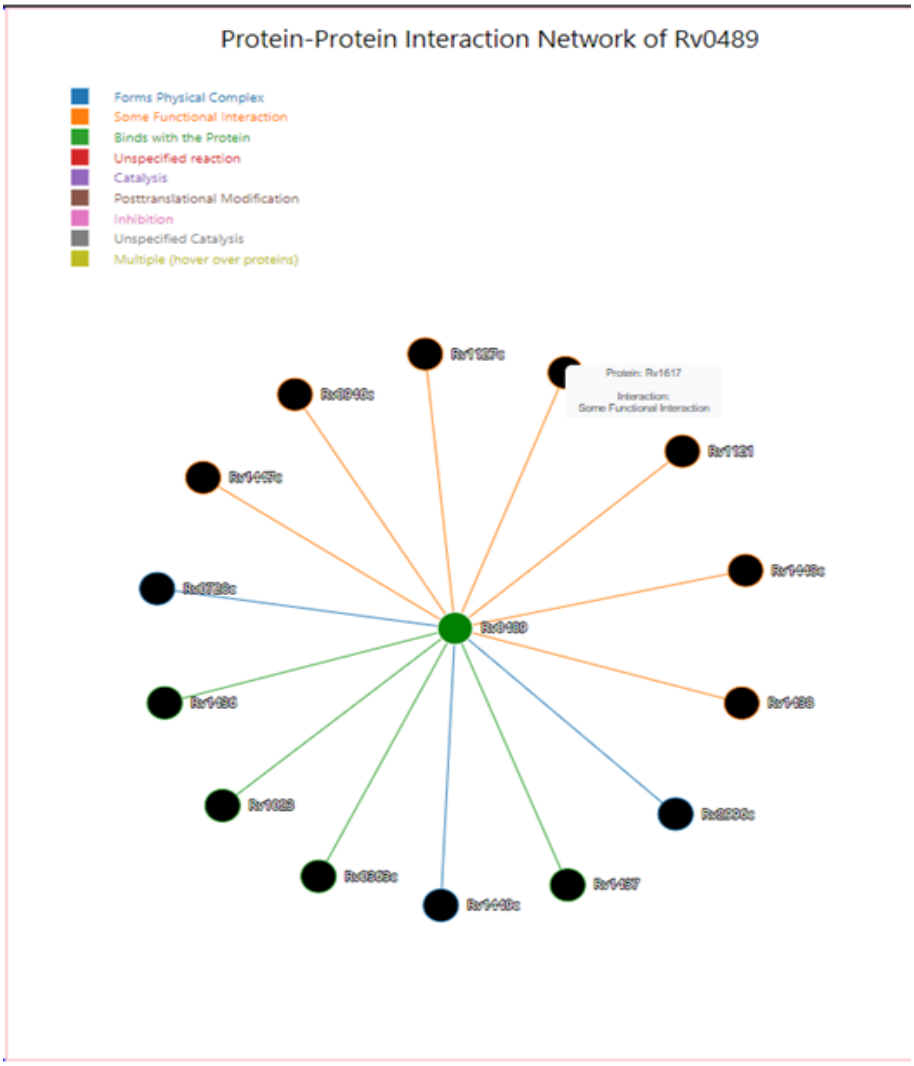



**Figure 4.3:** ***A sample image of the protein-protein interaction panel.*** *This sample image displays the protein-protein interaction panel within the prototype biovisualization. The protein-protein interaction network of the protein product of Rv0489. A legend to decipher legend color with the specific type of interaction is found within the panel. The Rv1617 node is hovered over to display the tooltip.*

### 4.3.3 The Regulation Data Panel

This panel displays data that the user inputs for noting changes in gene expression. Attached to this panel, the user is able to upload a csv file from the client side to the server. The csv file in question specifies two parameters: protein name and fold change, as would be obtained from data such as RNA sequencing. When uploaded, the panel will display a maximum of two data tables. The first table contains the protein name, pathway or pathways the protein is located in, whether the resulting regulation is positive or negative, and the exact degree of fold change taken from the data inputted by the user. The second table contains the transcription factor name, pathway or pathways the protein is located in, whether the resulting regulation is positive or negative, and the exact degree of fold change taken from the data inputted by the user. Additionally, using this uploaded data, the metabolic pathway panel would be modified. Any protein or transcription factor listed in the selected pathway would result in that specific node being outlined in green for upregulation or red for downregulation. If the fold change was zero or not included, no change to the protein or transcription factor node would be seen. Currently, if a different pathway should be examined, a user would have to re-upload the dataset for regulation again. Another limitation of the design is that only data part of the final data set compiled is included in the tables mentioned and the changes in gene expression. Figure 4.4, Figure 4.5, and Figure 4.6 displays a set of sample images that display an active regulation data panel, the mechanism for file upload, and the changes a file upload has on the metabolic pathway panel respectively.

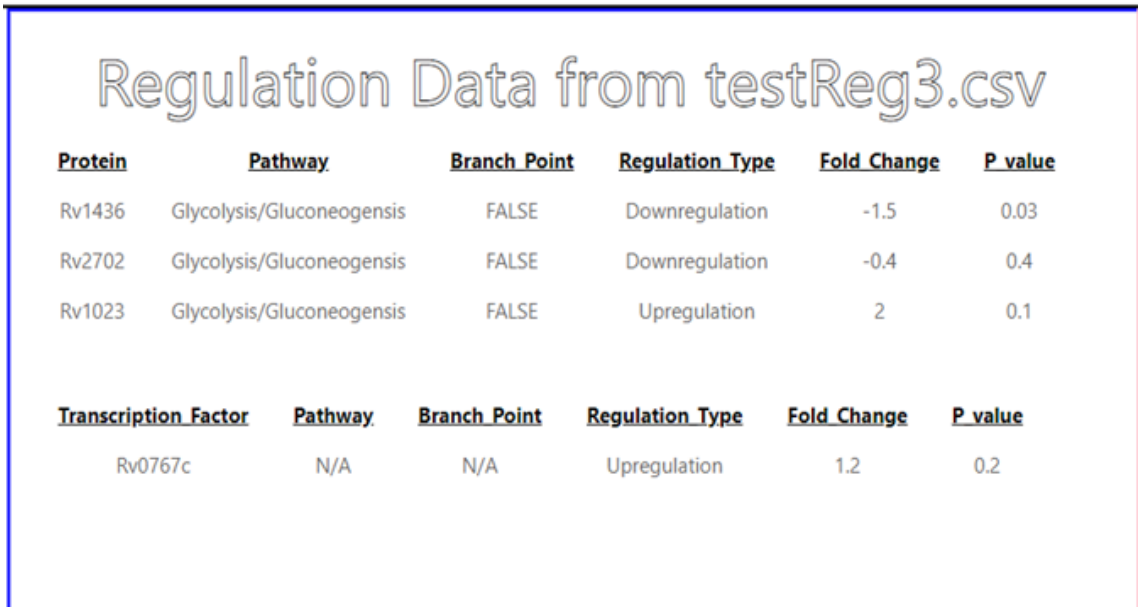



**Figure 4.4:** ***A sample image of the regulation data panel.*** *The sample image shows the panel after a sample dataset, testReg3, is uploaded to the prototype biovisualization. Two tables, one from proteins and one for transcription factors, are displayed that contain data from the uploaded file and from dataset the prototype biovisualization uses.*


**Figure 4.5:** ***A sample image of the file upload mechanism tied to the regulation data panel.*** *A file is selected and is then uploaded using the respective button.*

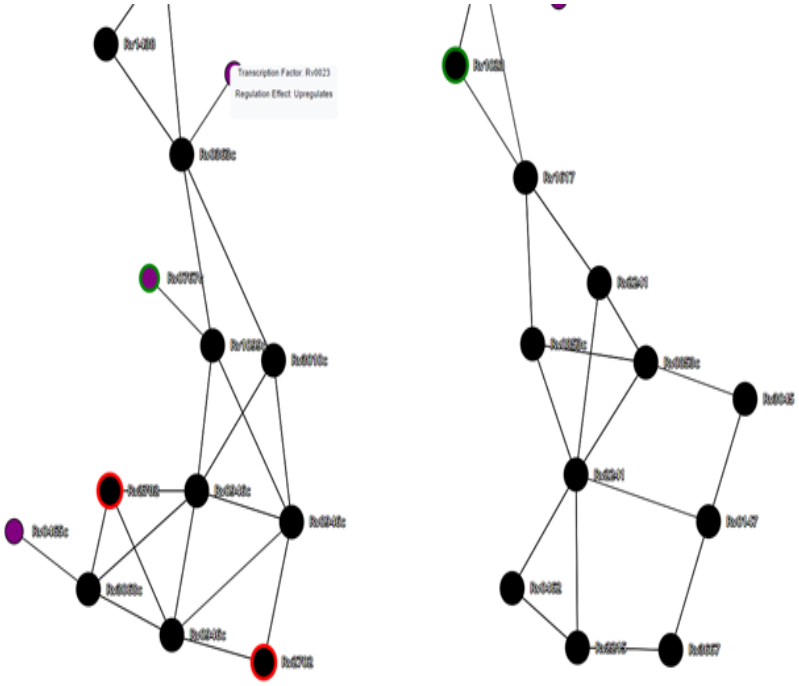


**Figure 4.6:** ***A sample image of the impacts of the file upload on the metabolic pathway panel.*** *Displayed is a portion of the Glycolysis / Gluconeogenesis pathway in the metabolic pathway panel after uploading a sample gene expression dataset. The panel is much larger in the prototype biovisualization. The uploaded dataset is testReg3. Upregulated genes proteins are outlined in green whereas downregulated proteins are outlined in red. Exact references regarding the fold change and significance are done by referring to the tables in the regulation data panel.*

# Chapter 5

# Discussion

Presented in this paper is a prototype biovisualization of *M. tuberculosis* metabolic pathways. Using a variety of databases, a dataset was compiled that incorporated data specific to the aspects of this visualization. A set of interviews with researchers in the *M. tuberculosis* community were conducted to build an understanding of the goals and tasks that would be important for building a metabolic pathways biovisualization. With this data and direction, a prototype biovisualization of *M. tuberculosis* metabolic pathways was constructed. This visualization consists of three distinct panels: one that showcases one specific pathway, one that indicates protein-protein interactions, and one that allows users to input gene expression data and use it to draw inferences from the first panel. The final prototype is, as the name implies, a representation of what a final product would look like. There are some noteworthy limitations with the dataset and the interviews that should be noted. Likewise, there are a number of aspects of the biovisualization that should be addressed or added in order to improve the user experience with the biovisualization as a whole. Despite these limitations, this prototype presents a tool for visualizing gene expression datasets alongside protein-protein interactions and metabolic pathways specific to *M. tuberculosis* from which researchers can use and draw inferences about the effect of gene expression changes on the overall system.

## 5.1 Data Limitations and Future Applications

As is true for any set of compiled or collected data, there are limitations to the one created in this paper. In an effort to create the dataset for a few pathways, the data was taken from many

different databases that were of high confidence and accompanied with relevant citations. The original literature was not checked to ensure the validity of the data if there was any discrepancy. Furthermore, the amount of available data from these databases was limited. There were a few instances where access to a database and its information was locked unless a subscription was purchased. While going to the original literature is another option, this process is time intensive and can also be prevented without a subscription as well. Using a number of databases and collaborating with researchers who have access to these locked databases, especially if the data is important to the visualization, could prove useful to obtaining the necessary data. One notable case can be found in the protein-protein interaction data and panel. This panel displays interactions of a protein of interest with high confidence. In the construction of this dataset, only data of high confidence from primarily STRING and STITCH was selected. This limits the ability of the users to explore deeper protein neighborhoods as with these two databases and as highlighted in Dang, et al. (2017). Selecting high confidence data in the construction of the dataset limits this ability in favor of greater empirical evidence.

In terms of methodology, the data collected was taken and investigated by hand. Any and all data with high confidence was recorded individually and exported as a csv formatted file for use in the visualization. Part of this effort was to understand what data should be included as an important step in creating the base dataset the visualization draws from. In order to expand the use of the visualization, however, more pathways will need to be added to this dataset. There may be a more efficient method that can be used to comb through one or multiple databases at a time and collect the necessary data. Doing so will save time in building the dataset that can be allocated to improving the visualization.

## 5.2 Interview Limitations and Future Applications

As noted, previous work such as in Murray et al, (2016) and Dang, et al. (2017), have identified many key goals and tasks that are important to researchers in general. To gain a greater insight into the specific needs and tasks for a biovisualization of *M. tuberculosis* metabolic pathways, interviews were conducted with researchers in the *M. tuberculosis* community. In order to get a better understanding of the goals and tasks that are specific to this group of researchers, more interviews should be performed. Adding questions specific to improving the current prototype biovisualization would be useful to understand areas not highlighted in the results or discussion that can be improved for both efficiency and user experience.

## 5.3 Biovisualization Limitations and Future Applications

Limitations of the biovisualization are related to both implementation and design choice. Starting with the former, the biovisualization is defined as a prototype, implying further development and improvement should be completed. One aspect of this development that should be of key interest is related to the uploading of data in the regulation data panel. In its current implementation, uploading a file will only indicate changes of gene expression related to the current pathway layout visualized in the metabolic pathway panel. This limits the ability of the user to actively investigate, discover, and draw inferences about their data on metabolic pathways as a whole. Furthermore, if the layout is changed and the user reuploads the same data set, no gene expression changes will be shown as the information has already been uploaded. Thus, finding a way to remove datasets from the program and to hold onto their information between layouts is an important next step in the development of this biovisualization.

When uploading data to the regulation panel, the user is expected to reference the regulation data panel to understand the degree of change of a protein or transcription factor's expression. While a small set of data is manageable, larger datasets will make this task less efficient and manageable if the user would like to easily denote changes in expression between two or more proteins. Implementing a colored scale would be easier for the user and make denoting differences in expression more efficient. Additionally, the current set-up of the regulation data panel involves incorporating the uploaded data in a set of tables for the user to easily reference. As more data is included in uploaded files, the panel will become filled with long lists of tables. To better facilitate this process, making the regulation data panel svg scrollable would easily facilitate the additional data without losing the ability to see the information.

In regard to design choice, there are three key aspects that limit the presentation of understanding of the data: accessibility to the colorblind, the use of force-directed networks, and directionality of the edges in those networks. Firstly, the biovisualization does not take into account colorblindness. While panel backgrounds will change, as they simply exist for designation, node coloration and the legend in the protein-protein interaction panel will be important for denoting different items. Identifying and using colors that are color-blind friendly will be an important step for improving the biovisualization for a wider audience. This idea should be at the forefront with other recommendations aforementioned. Secondly, force-directed graphs are limiting in their ability to convey meaningful structure. Many pathways are drawn to convey more of an ordered progression that makes recognizing and identifying the pathway much easier. Glycolysis and Gluconeogenesis, for example, are often depicted as a linear chain of reactions. Force-directed networks palace nodes at a certain distance from each other by some force metric and length distance that is irrespective of commonly drawn organizations used by researchers.

Using this network structure was found to be the best method for implementing many different protein data points together at the cost of this lost structure. While the different layouts highlight the organized flow, looking at other network structures and algorithms may improve the user experience and help them draw better inferences from the visualization as a whole. Thirdly, the current biovisualization has edges which are undirected. Future improvements should add direction, especially for reactions where equilibrium cannot be assumed and is not the case. Likewise, adding direction for transcription factor action would make visualizing the action simpler.

CHAPTER 6

# CONCLUSIONS

Metabolic networks are complex systems that are always in flux because of changes in gene expression within a living system. To better understand these living systems, visualizing these changes will allow researchers to better understand how these changes in these pathways impact growth, survival, and persistence. In this paper, a prototype metabolic pathway visualization for *M. tuberculosis* was constructed based upon the goals and tasks from published work and interviews with researchers in the *M. tuberculosis* community. While nothing inherently new in design and task implementation, such as a new network design, was constructed with this prototype, this visualization combines metabolic pathway data and protein-protein interaction with the ability to upload gene expression datasets to allow researchers to visualize and draw inferences from their data. Furthermore, this prototype creates a visualization tool with a specific focus on *M. tuberculosis* for use in that research community. It is the goal of this paper to showcase these ideas in order to be able to implement them in a more integrated and robust visualization.

## APPENDIX A: INTERVIEW QUESTIONS

1. Can you please tell me about your research and work?
2. Have you ever used any databases or visualizations specific to biology generally or *M. tuberculosis* specifically? What have you used them for? (Continues if applicable)
    a) With regard to this visualization, what aspects of the visualization is useful for understanding *M. tuberculosis* metabolic pathways?
    b) Is there anything that is extraneous, ineffective or missing, regarding the visual or the data present?
3. Using your knowledge of the metabolic pathway (and the visualization you mentioned if applicable), what are the key elements when visualizing metabolic or protein data?
    a) Could you list some particular actions that are associated with those key components?
4. *Shows the KEGG database OR from answer to question 2 and provides time for exploration*
    a) What are your thoughts when first exploring this visualization? *If needed* Examples would include the design, the data, the interface, etc.
    b) With regard to this visualization, what aspects of the visualization is useful for understanding the *M. tuberculosis* metabolic pathway?
    c) Is there anything that is extraneous, ineffective or missing, regarding the visual or the data present?
5. *Shows the STRING PPI visualization and provides time for exploration* (Using a goal - like find confidence and most strong evidence for a functional and physical connection).
    a) What are your thoughts when first exploring this visualization? *If needed* Examples would include the design, the data, the interface, etc.
    b) With regard to this visualization, what aspects of the visualization is useful for understanding the *M. tuberculosis* metabolic pathway?
    c) Is there anything that is extraneous, ineffective or missing, regarding the visual or the data present?
6. In visualizing the *M. tuberculosis* metabolic pathway and given from the examples shown, what are some aspects that you particularly liked or disliked across all the visualizations?
7. After observing some sample visualizations, what would you determine are the most important parts of a visualization? This could be data to include, features, details, etc.

8. It is my goal to create a biovisualization that incorporates the *M. tuberculosis* metabolic pathway, a PPI network, and a method to input RNA-seq data to identify changes in protein levels and regulation.
    a) Given what you have seen and identified, are there other key components that should be noted when comparing and visualizing these three aspects together?
    b) Are there areas that you believe should receive more or less detail?
    c) How would a visualization like this be useful in your research and work, if at all?
9. Are there any other details you would like to add or questions you would like to ask?

## Appendix B: Informed Oral Consent Form

Thank you again for responding and accepting my request for an interview with you. In order to proceed with the interview, I ask that you kindly provide oral consent to this interview.

You, the participant, are being invited to participate in an interview with Adrian Orszulak. This interview is part of a Major Qualifying Project (MQP) to understand the goals and tasks for creating a biovisualization of the *M. tuberculosis* metabolic pathways. In this interview, I am seeking to define what you, the researcher, would define as important goals for a biovisualization of the *M. tuberculosis* metabolic pathways. Using this information, I will be able to list a set of goals for building the said visualization, and identify the corresponding tasks needed to accomplish those goals. The interview will last from 1 hour to 1.5 hours. Your name will not be published, but your specific field of work (such as metabolomics) will be.

Do you have any questions?

This process is a voluntary one. If you do not want to answer any particular question, you are not obligated to do so. We will simply move on should you make this known. Additionally, if you no longer feel open to an interview or wish to stop the interview, you have the right to withdraw from the interview at any point.

Do I have your consent to move forward with this interview?

In addition, I would like to ask if I could have your consent to record the interview. You do not need to agree to have the interview recorded. I will act as both interviewer and notetaker if you do not consent to the recording. Do I have your consent to record the interview?

Do you have any questions before we begin?

# BIBLIOGRAPHY


[1] B. Worley and R. Powers, "Multivariate analysis in metabolomics," *Curr. Metabolomics*, vol. 1, no. 1, pp. 92–107, 2013.

[2] D. B. Kell, M. Brown, H. M. Davey, W. B. Dunn, I. Spasic, and S. G. Oliver, "Metabolic footprinting and systems biology: the medium is the message," *Nat. Rev. Microbiol.*, vol. 3, no. 7, pp. 557–565, Jul. 2005.

[3] World Health Organization, "Tuberculosis (TB)," https://www.who.int/publications/i/item/9789240037021., 2021.

[4] N. Levin, "Multivariate statistics and the enactment of metabolic complexity," *Soc. Stud. Sci.*, vol. 44, no. 4, pp. 555–578, Aug. 2014.

[5] S. Ren, A. A. Hinzman, E. L. Kang, R. D. Szczesniak, and L. J. Lu, "Computational and statistical analysis of metabolomics data," *Metabolomics*, vol. 11, no. 6, pp. 1492–1513, Dec. 2015.

[6] D. F. Warner, "Mycobacterium tuberculosis metabolism," *Cold Spring Harb. Perspect. Med.*, vol. 5, no. 4, pp. a021 121–a021 121, Dec. 2014.

[7] B. Lee, C. Plaisant, C. S. Parr, J.-D. Fekete, and N. Henry, "Task taxonomy for graph visualization," in *Proceedings of the 2006 AVI workshop on BEyond time and errors novel evaluation methods for information visualization - BELIV '06*. New York, New York, USA: ACM Press, 2006.

[8] P. Murray, F. McGee, and A. G. Forbes, "A taxonomy of visualization tasks for the analysis of biological pathway data," *BMC Bioinformatics*, vol. 18, no. Suppl 2, p. 21, Feb. 2017.

[9] T. Dang, P. Murray, and A. Forbes, "BioLinker: Bottom-up exploration of protein interaction networks," in *2017 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, Apr. 2017.

[10] Y. Wang, Q. Wang, H. Huang, W. Huang, Y. Chen, P. B. McGarvey, C. H. Wu, C. N. Arighi, and UniProt Consortium, "A crowdsourcing open platform for literature curation in UniProt," *PLoS Biol.*, vol. 19, no. 12, p. e3001464, Dec. 2021.

[11] UniProt Consortium, "UniProt: the universal protein knowledgebase in 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D480–D489, Jan. 2021.

[12] S. Turkarslan, E. J. R. Peterson, T. R. Rustad, K. J. Minch, D. J. Reiss, R. Morrison, S. Ma, N. D. Price, D. R. Sherman, and N. S. Baliga, "A comprehensive map of genome-wide gene regulation in mycobacterium tuberculosis," *Sci. Data*, vol. 2, no. 1, Dec. 2015.

[13] J. J. Davis, A. R. Wattam, R. K. Aziz, T. Brettin, R. Butler, R. M. Butler, P. Chlenski, N. Conrad, A. Dickerman, E. M. Dietrich, J. L. Gabbard, S. Gerdes, A. Guard, R. W. Kenyon, D. Machi, C. Mao, D. Murphy-Olson, M. Nguyen, E. K. Nordberg, G. J. Olsen, R. D. Olson, J. C. Overbeek, R. Overbeek, B. Parrello, G. D. Pusch, M. Shukla, C. Thomas, M. VanOeffelen, V. Vonstein, A. S. Warren, F. Xia, D. Xie, H. Yoo, and R. Stevens, "The PATRIC bioinformatics resource center: expanding data and analysis capabilities," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D606–D612, Jan. 2020.

[14] P. D. Karp, R. Billington, R. Caspi, C. A. Fulcher, M. Latendresse, A. Kothari, I. M. Keseler, M. Krummenacker, P. E. Midford, Q. Ong, W. K. Ong, S. M. Paley, and P. Subhraveti, "The BioCyc collection of microbial genomes and metabolic pathways," *Brief. Bioinform.*, vol. 20, no. 4, pp. 1085–1093, Jul. 2019.

[15] S. Paley, R. Billington, J. Herson, M. Krummenacker, and P. D. Karp, "Pathway tools visualization of organism-scale metabolic networks," *Metabolites*, vol. 11, no. 2, p. 64, Jan. 2021.

[16] R. Caspi, R. Billington, I. M. Keseler, A. Kothari, M. Krummenacker, P. E. Midford, W. K. Ong, S. Paley, P. Subhraveti, and P. D. Karp, "Karp the MetaCyc database of metabolic pathways and enzymes - a 2019 update," 2019.

[17] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[18] M. Kanehisa, "Toward understanding the origin and evolution of cellular organisms," *Protein Sci.*, vol. 28, no. 11, pp. 1947–1951, Nov. 2019.

[19] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, "KEGG: integrating viruses and cellular organisms," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D545–D551, Jan. 2021.

[20] A. Chang, L. Jeske, S. Ulbrich, J. Hofmann, J. Koblitz, I. Schomburg, M. Neumann-Schaal, D. Jahn, and D. Schomburg, "BRENDA, the ELIXIR core data resource in 2021: new developments and updates," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D498–D508, Jan. 2021.

[21] D. Szklarczyk, A. Santos, C. von Mering, L. J. Jensen, P. Bork, and M. Kuhn, "STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D380–4, Jan. 2016.

[22] D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, L. J. Jensen, and C. Mering, "January) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res*, vol. 47, pp. D607–61 349, 2019.

[23] ——, ") the STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets," *Nucleic Acids Res*, vol. 49, no. D1, pp. D605–612, 2021.

[24] Z. Hu, J. Mellor, J. Wu, and C. DeLisi, "VisANT: an online visualization and analysis tool for biological interaction data," *BMC Bioinformatics*, vol. 5, p. 17, Feb. 2004.

[25] Z. Hu, E. S. Snitkin, and C. DeLisi, "VisANT: an integrative framework for networks in systems biology," *Brief. Bioinform.*, vol. 9, no. 4, pp. 317–325, Jul. 2008.

[26] B. R. Granger, Y.-C. Chang, Y. Wang, C. DeLisi, D. Segrè, and Z. Hu, "Visualization of metabolic interaction networks in microbial communities using VisANT 5.0," *PLoS Comput. Biol.*, vol. 12, no. 4, p. e1004875, Apr. 2016.

[27] M. Baitaluk, M. Sedova, A. Ray, and A. Gupta, "BiologicalNetworks: visualization and analysis tool for systems biology," *Nucleic Acids Res.*, vol. 34, no. Web Server issue, pp. W466–71, Jul. 2006.

[28] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003.

[29] C. Baker, M. Carpendale, P. Prusinkiewicz, and M. Surette, "Genevis: visualization tools for genetic regulatory network dynamics," in *IEEE Visualization, 2002. VIS 2002.*, 2002, pp. 243–250.

[30] B. H. Junker, C. Klukas, and F. Schreiber, "VANTED: a system for advanced data analysis and visualization in the context of biological networks," *BMC Bioinformatics*, vol. 7, no. 1, p. 109, Mar. 2006.

[31] A. Kapopoulou, J. M. Lew, and S. T. Cole, "The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes," *Tuberculosis (Edinb.)*, vol. 91, no. 1, pp. 8–13, Jan. 2011.