Mechanisms of Differential Expression of ESX-1 Secretion System Genes in *Mycobacterium 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

By:

Ryan Peters

May 6th, 2021

Approved by:
Scarlet S. Shell, PhD

Abstract

An important strategy employed by *Mycobacterium tuberculosis* (Mtb) to evade the host immune system is to escape from phagosomes into the cytoplasm of macrophages. To accomplish this and other goals, Mtb secretes various virulence factors through several secretion systems. One such system, ESX-1, is necessary to escape the host phagosome. In particular, EsxA and EsxB are proteins secreted by the ESX-1 system that are important for disrupting the phagosomal membrane. These proteins appear to be encoded in a four-gene operon in the ESX-1 locus. However, previous work has shown that mRNA abundance differs between genes in this supposed operon, in a way that cannot be explained by the known transcription start site (TSS) and endoribonuclease cleavage site. Here we search for additional promotors in this *PE35-PPE68-esxB-esxA* locus and examine the contribution of each TSS and cleavage site to mRNA expression. Using the model organism *Mycobacterium smegmatis*, we characterize a series of mutaitons by quantitative PCR and 5' rapid amplification of cDNA ends. We map two promotors that contribute to expression of *esxB-A*, an mRNA cleavage site upstream of *esxB*, and suggest that at least one additional promotor is present in the locus. These data lay the foundation for a better understanding of the mechanisms behind differential expression of ESX-1 genes in mycobacteria.

Introduction

Tuberculosis is one of the top 10 causes of death globally. In 2019 alone, approximately 1.4 million people died from tuberculosis and 10 million people fell ill (World Health Organization, 2020). The success of the pathogen, *Mycobacterium tuberculosis* (Mtb) which causes tuberculosis, comes in part from its ability to survive inside and ultimately escape host macrophages. Typically, macrophages phagocytose pathogens to kill and degrade them. However, Mtb manages to survive this process and even disrupt the phagosomal membrane to escape into the macrophage cytosol. The EsxA protein, which is secreted by the ESX-1 system, has been shown to mediate this phagosomal membrane rupture (Bosserman & Champion, 2017; Conrad et al., 2017; Houben et al., 2014; Simeone et al., 2009). Understanding the mechanism(s) regulating expression of this protein is therefore crucial to gaining a better understanding of Mtb pathogenesis, and in identifying novel drug targets.

In order to transport substrates across its two membranes, mycobacteria use up to five different Type VII secretion systems named ESX-1 through ESX-5. These systems contain a mixture of conserved components (Ecc), ESX specific proteins (Esp), and previously named proteins including the PE/PPE proteins, small secreted Esx proteins, and MycP (Houben et al., 2014). The genetic makeup of each secretion system varies, but typically contains a conserved secretion machinery, PE/PPE genes in an operon next to a pair of adjacent Esx genes, and varying numbers of genes encoding Esp proteins (Houben et al., 2014). Additional homology is found in the protein structure of the Esx, PE/PPE, and some Esp proteins which form heterodimers composed of four alpha helices with an essential secretion motif on the C-terminus of one dimer. There are also many important differences between the ESX systems. Despite sharing conserved secretion machinery, only ESX-1, ESX-3, and ESX-5 have been proven to

actively secrete proteins (Bosserman & Champion, 2017; Houben et al., 2014). ESX-1 has an unusually high number of secreted Esp proteins, while ESX-5 encodes a greater number of PE/PPE proteins, though these are largely shown to localize at the cell envelope (Houben et al., 2014). ESX-3 and ESX-4 are present in all mycobacteria, but various species have different combinations of the other systems. The presence of the ESX-5 system correlates with slow growing mycobacteria; however, it is not known if the locus impacts growth rates directly (Houben et al., 2014). The ESX-1 system is widely known for encoding virulence factors in Mtb, but interestingly appears to have evolved other roles in some species. Notably, some pathogenic species do not have ESX-1 systems including *M. avium*, and *M. ulcerans*, or are naturally missing certain ESX-1 genes including *M. microti* (Bosserman & Champion, 2017; Houben et al., 2014). Other species, such as *M. smegmatis*, do have an ESX-1 system but are non-pathogenic. In *M. smegmatis*, ESX-1 is essential for establishing which cell is the donor and which is the recipient in conjugation, while ESX-4 controls other aspects of conjugation (Bosserman & Champion, 2017; Houben et al., 2014).

ESX-1 in Mtb was suspected to play a role in virulence when it was found to be located in one of the regions of difference between Mtb and *M. bovis* BCG, the attenuated strain used in vaccination. Restoration of the ESX-1 secretion system partially rescued the microbe's pathogenicity when infecting mice (Simeone et al., 2009). In particular, EsxA (previously known as ESAT-6) was shown to play an important role in virulence. EsxA is secreted as a heterodimer with EsxB (previously known as CFP-10). After secretion, the proteins encounter the acidic environment of the phagosome. A decrease in pH has been shown to cause EsxA to dissociate from EsxB. Early works proposed that in Mtb, EsxA would then directly create pores in the phagosomal membrane, while EsxB acted as a chaperone for secretion. These works show that EsxB, and notably the *M. smegmatis* homolog of EsxA, cannot disrupt membranes. (De Jonge et al., 2007; De Leon et al., 2012; Smith et al., 2008). Recent works have cast some doubt on these findings, but despite these flaws, it still remains that host phagosomal membrane rupture is dependent on ESX-1, in concert with a dependence on membrane lipids (especially DIM) and contact with the phagosomal membrane (Augenstreich et al., 2017; Conrad et al., 2017). It has also been shown that Mtb within macrophages exhibit increased expression of many ESX-1 proteins within 2 days, and it was suggested that the acidic environment of the phagosome triggers this change early on (Rohde et al., 2012).

esxB and *esxA* have long been predicted to be transcribed as part of an operon, with early work predicting a two gene operon (Berthet et al., 1998). Again, the pattern of two genes in an operon that code for a heterodimer seems to be conserved across many secreted substrates in ESX systems (Bosserman & Champion, 2017; Houben et al., 2014; Simeone et al., 2009). This synteny, along with previously described similarities, suggest the five different ESX secretion systems may be the result of duplication events (Houben et al., 2014). Therefore, understanding the regulation of one operon may be more broadly applicable. Unfortunately, previous works have noted that it is difficult to study individual components of the Esx systems at the protein level due to a high level of co-dependence between components. For example, secretion of EspA and EspC are dependent on EsxA and EsxB, and vice versa (Bosserman & Champion, 2017; Houben et al., 2014).

From a regulatory standpoint, several transcriptional regulators have been identified for ESX-1 genes. These regulators include MprAB, EspR, PhoP, and WhiB6 which often co-regulate lipid biogenesis and ESX-1 genes (Abdallah et al., 2019; Bosserman & Champion, 2017; Simeone et al., 2009). It has also been suggested that post-transcriptional regulation plays a role in modulating *esxB-esxA* expression, especially under stress (Bosserman & Champion, 2017). Post-transcriptional regulation occurs through multiple mechanisms that ultimately impact mRNA degradation and/or translation efficiency. In the typical lifetime of an mRNA transcript, it will be bound by ribosomes and translated into its amino acid sequence. At some point it will be cleaved by endoribonucleases before undergoing total disassembly by exoribonucleases. Transcripts begin with three phosphate groups at the 5' end, and cleavage by endoribonucleases generally results in a monophosphorylated transcript. Many factors may impact the final amount of protein made, such as the efficiency of ribosome binding and overall translation efficiency, features of the transcript that impact its ability to be cleaved by endo- or exoribonucleases, and additional molecules that may bind the mRNA to modulate its half-life (Picard et al., 2009; Van Assche et al., 2015). One key feature of a transcript that impacts its half-life is the secondary structure. Certain secondary structures are believed to obscure cleavage sites making it inaccessible to ribonucleases, thereby stabilizing the transcript (Picard et al., 2009).

In addition to endoribonucleolytic cleavage leading to degradation, a number of studies have suggested that cleavage events may actually stabilize certain transcripts, especially those in operons. Studies across a range of bacteria including *Escherichia coli, Clostridium perfringens, Bacillus subtilis*, and even Mtb and *M. smegmatis*, have shown examples of a cleavage event leading to stabilization of a gene. This often allows for differential expression of genes in a polycistronic transcript (Baga et al., 1988; Lodato & Kaper, 2009; Meinken et al., 2003; Obana et al., 2010; Sala et al., 2008). In a prior study on *M. smegmatis*, a combination of 5' end mapping to find cleavage sites and RNA-Seq to quantify expression levels were used to generate a transcriptome wide map of potentially stabilizing cleavage sites and operons (Shell et al., unpublished). One such site is the *PE35-PPE68-esxB-esxA* locus. This potential operon begins with a transcription start site (TSS) at 4,350,613 followed by coding regions for *PE35*, *PPE68*, *esxB*, and *esxA* as shown in Figure 1. 5' end mapping shows that cleavage occurs 70 nucleotides before the *esxB* start codon. RNA-seq shows that *esxA* and *esxB* are expressed at much higher levels than the upstream PE/PPE genes. Given the lack of another known TSS at the time, this cleavage was suggested to stabilize *esxB* and *esxA* transcripts relative to upstream genes in the operon (Shell et al., unpublished). It was suspected this stabilization was made possible by the 5' untranslated region (UTR) upstream of *esxB* which is predicted to have a secondary structure including two prominent hairpin loops. Further work suggested that removal of one loop versus the other had opposite impacts on mRNA abundance and protein expression, while deleting both appeared to increase mRNA abundance (deRivera & Shell, 2016). Additional work has shown that deletion of either individual loop reduced protein expression, while deletion of both increased expression. However, no significant difference was found in the half-life of each transcript (Kelly & Shell, 2021). Given the lack of a clear relationship between these hairpin loops and expression, other mechanisms of differential expression should be examined.

Here we present a novel model of regulation for the *PE35-PPE68-esxB-esxA* locus involving two TSSs and a cleavage site, as shown in Figure 1. This hypothesis is largely based upon the model proposed by Shell et al. (unpublished), where differential expression was achieved through cleavage-dependent stabilization of the *esxB-A* transcript. The addition of a second TSS between *PPE68* and *esxB*, just 3 or 4 nucleotides upstream of the cleavage site in this model allows for an additional mechanism of differential expression, which may not require cleavage-dependent stabilization. Our investigation of strains with mutations in these feature(s) reveal their individual roles, allowing us to form a better understanding of the system as a whole.

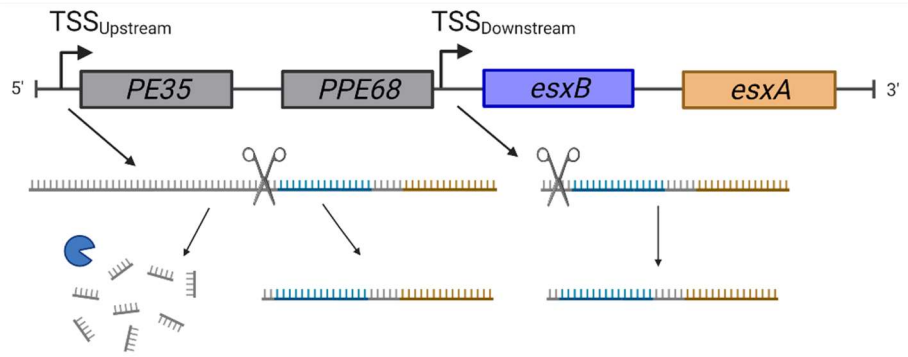

Figure 1. Model of *esxB-A* regulation. A graphical representation of the proposed mechanisms for increased expression of the *esxB* and *esxA* genes relative to the two upstream genes. Cleavage of the transcript from $TSS_{Upstream}$ creates two products which may have different stabilities, resulting in quicker degradation of the *PE35* and *PPE68* transcript. The second putative transcription start site would create more *esxB* and *esxA* transcripts, and may not require cleavage-dependent stabilization to generate differential expression. A bent arrow represents a TSS, and the scissors represent a cleavage site. Created with BioRender.com.

Materials and Methods

Strains, Plasmids, and RNA Extraction

Previous work on this system generated 30 *M. smegmatis* strains with 10 different plasmids, as detailed in Table 1. These strains were created from the mc^2155 strain with the native msmeg_0062-0066 locus deleted and were transformed with the plasmids listed in Table 1. These plasmids differ from those previously studied in the addition of a tsynA transcriptional terminator upstream of the genes of interest. Cultures were grown in Difco™ Middlebrook 7H9 broth with 0.2% glycerol, 0.05% Tween 80, and Albumin Dextrose Catalase (ADC) to have a final concentration of 5 g/L bovine serum albumin fraction V (BSA), 2 g/L dextrose, 0.85 g/L sodium chloride, and 3 mg/mL catalase. Additionally, 250 µg/mL of hygromycin was added to the media to maintain the plasmids, which are episomal. Cultures were grown to an OD_{600} of 0.41 to 0.55. Culture pellets were frozen in liquid nitrogen and RNA was extracted with the Direct-zol™ RNA extraction and purification kit (Zymo Research) using a FastPrep 5G (MP Biomedicals). Each sample was then purified using the RNA Clean & Concentrator™-25 kit (Zymo Research) according to manufacturer's instructions, and measured using a NanoDrop One (Thermo Scientific). This total RNA was used throughout the study.

Table 1. ***Mycobacterium smegmatis*** **strains**. Three clones from mc^{2}155 Δ*msmeg_0062-0066* transformed with each plasmid were examined. tsynA is a transcriptional terminator, mutations to the promotors are at the -10 site which blocks essential activity of the housekeeping sigma factor *SigA* (see Figure 2 for sequence), and mutations in the cleavage site change C to a G (see Figure 2 for sequence).

Strain	Plasmid	Plasmid description
SS-M_0105 – SS-M_0107	pSS176	tsynA, native $TSS_{Upstream}$, *PE35*, *PPE68*, native $TSS_{Downstream}$, native cleavage site, *esxB, esxA*
SS-M_0108 – SS-M_0110	pSS177	tsynA, native $TSS_{Upstream}$, *PE35*, *PPE68*, 30 nt deletion of $TSS_{Downstream}$ and cleavage site, *esxB, esxA*
SS-M_0328 – SS-M_0330	pSS253	tsynA, native $TSS_{Upstream}$, *PE35*, *PPE68*, mutated $TSS_{Downstream}$, native cleavage site, *esxB, esxA*
SS_M_0334 – SS-M_0336	pSS255	tsynA, native $TSS_{Upstream}$, *PE35*, *PPE68*, native $TSS_{Downstream}$, mutated cleavage site, *esxB, esxA*
SS-M_0340 – SS-M_0342	pSS257	tsynA, native $TSS_{Upstream}$, *PE35*, *PPE68*, mutated $TSS_{Downstream}$, mutated cleavage site, *esxB, esxA*
SS-M_0325 – SS-M_0327	pSS252	tsynA, mutated $TSS_{Upstream}$, *PE35*, *PPE68*, native $TSS_{Downstream}$, native cleavage site, *esxB, esxA*
SS-M_0111 – SS-M_0113	pSS178	tsynA, no $TSS_{Upstream}$, *PE35*, *PPE68*, native $TSS_{Downstream}$, native cleavage site, *esxB, esxA*
SS-M_0331 – SS-M_0333	pSS254	tsynA, no $TSS_{Upstream}$, *PE35*, *PPE68*, mutated $TSS_{Downstream}$, native cleavage site, *esxB, esxA*
SS-M_0337 – SS-M_0339	pSS256	tsynA, no $TSS_{Upstream}$, *PE35*, *PPE68*, native $TSS_{Downstream}$, mutated cleavage site, *esxB, esxA*
SS-M_0343 – SS-M_0345	pSS258	tsynA, no $TSS_{Upstream}$, *PE35*, *PPE68*, mutated $TSS_{Downstream}$, mutated cleavage site, *esxB, esxA*

cDNA Synthesis

RNA was diluted to the same concentration (600 ng for qPCR, and 300.3-477.3 ng for 5' RACE depending on the batch) in 5.25 µL before 0.83 µL of 100 mM Tris pH 7.5, and 0.17 µL of 3 mg/mL Random Primers were added. Next, samples were incubated at 70°C for 10 minutes before being snap cooled for 5 minutes. Reverse transcription was carried out by adding 2 µL of 5X ProtoScript II Buffer, 0.5 µL 10 mM each dNTPs, 0.5 µL 100 mM DTT, 0.25 µL 40,000 U/mL RNase Inhibitor Murine (NEB), and 0.5 µL 200,000 U/mL ProtoScript II Reverse Transcriptase (NEB) to the template-primer mix on ice. The same reaction was also carried out with an equal volume of H_2O in place of the reverse transcriptase, to control for the presence of genomic DNA (gDNA). This mixture was then incubated at 25°C for 10 minutes followed by 42°C for at least five hours, and 4°C thereafter. Following cDNA synthesis, alkaline degradation was used to remove template RNA. To accomplish this 5 µL of 0.5 mM EDTA, and 5 µL of 1 N NaOH was added to each sample, and incubated for 15 minutes at 65°C. The reaction was stopped with 12.5 µL of 1 M Tris HCl with a pH of 7.5. cDNA cleanup was carried out with the Monarch® PCR & DNA Cleanup Kit (5 µg) (NEB) according to the manufacturer's directions and measured using a NanoDrop One (Thermo Scientific).

Quantitative PCR

To determine the relative transcript abundance for genes of interest, quantitative PCR was performed on 3 biological replicates. All cDNA samples synthesized from total RNA were diluted in ultrapure water to a concentration of 1 ng/µL, then further diluted to 200 pg/µL. Samples that were not treated with reverse transcriptase during cDNA synthesis were diluted with the same volumes of water as their treated counterparts. Each qPCR reaction contained 400 pg of cDNA or an equivalent volume of no-RT control, 1 µL of a solution with 2.5 µM of each appropriate primer, and 5 µL of iTaq Universal SYBR Green Supermix (BioRad) in a total volume of 10 µL. Reactions were done in 96 well plates. The plate was covered with a clear film and amplified with an Applied Biosystems 7500 qPCR machine. The samples were incubated at 50°C for 2 minutes, 95°C for 10 minutes, then 40 cycles of 95°C for 15 seconds followed by 61°C for 1 minute. At least one well per primer set was filled with an equal volume of H_2O instead of the cDNA template to test for contamination. To compare the transcript abundance of each sample, results were normalized to the housekeeping gene *SigA*. For each strain the number of cycles required to reach a threshold (C_T) of 0.2 was compared for *SigA* and the gene of interest in order to calculate the ΔC_T. Relative expression was defined as $2^{-\Delta ct}$. Primers used for qPCR are listed in Supplemental Table 1.

5' Rapid Amplification of cDNA Ends (RACE)

One set of samples in the converted library were treated with RppH High Concentration (NEB) according to manufacturer's instructions to remove a pyrophosphate from the 5' end of triphosphorylated transcripts, thereby converting them to monophosphates. Nonconverted samples underwent a mock treatment with H_2O instead of RppH. Each set of samples was then purified using the RNA Clean & Concentrator™-5 kit (Zymo Research) according to manufacturer's instructions, and measured using a NanoDrop One (Thermo Scientific). Next, a universal adaptor was ligated to the 5' end of monophosphorylated transcripts. First, all samples were diluted to the same concentration (509.6-720.8 ng depending on the batch) in 8 µL, then incubated with 1 µg of the oligo SSS1016 (see Supplemental Table 1 for sequence) at 60°C for 10 minutes then snap cooled on ice. The following was then added to each reaction: 10 µL 50% PEG8000, 3 µL 10X T4 RNA Ligase I Buffer (NEB), 3 µL 10 mM ATP, 3 µL dimethyl sulfoxide (DMSO), 1 µL RNase Inhibitor Murine (NEB), and 1 µl T4 RNA Ligase I (NEB). These reactions were run at 20°C overnight. Each sample was then purified using the RNA Clean & Concentrator™-5 kit (Zymo Research) according to manufacturer's instructions, and measured using a NanoDrop One (Thermo Scientific). This RNA was then used as a template for cDNA synthesis as described previously, and followed by PCR and DNA recovery as detailed below.

PCR and DNA Recovery

Polymerase chain reactions (PCR) were done in a total volume of 25 µL containing 2.5 µL 10X Taq Reaction Buffer (NEB), 1.25 µL 10 µM primer SSS1017 or SSS2218 (Supplemental Table 1), 1.25 µL 10 µM reverse primer (designed for the gene of interest), 0.5 µL dNTPs (10 mM each), 0.167 µL Taq DNA polymerase (NEB), 1.5 µL 2.25 ng/µL template cDNA with adaptor, and the remaining volume of H_2O. A full list of oligonucleotides used in this study can be found in Supplementary Table 1. PCR reactions were carried out under the following conditions: (i) an

initial step for DNA denaturing at 95°C for 5 minutes, (ii) 35 cycles of 95°C for 30 seconds (denaturing), 52-57°C for 20 seconds (annealing), and 68°C for 12 seconds (elongation), and (iii) a final elongation at 68°C for 5 minutes. These conditions were optimized for each primer set, and are specified in Supplementary Table 2. The annealing temperature was calculated using the online NEB T_m calculator, and the duration of the elongation step was based on the size of the expected product, lasting at least 1 minute per kb.

PCR products were analyzed by gel electrophoresis. A 2.0% Agarose Quick Dissolve LE (Apex) gel was prepared in 1X Tris-acetate-EDTA (TAE) buffer. Bands of interest were sliced from the gel and purified using the Zymoclean™ Gel DNA Recovery Kit according to the manufacturer's instructions and measured with a NanoDrop One (Thermo Scientific). Concentrated DNA was then sent to an external contractor for Sanger sequencing.

Capping-RACE

The Vaccina Capping System (NEB) was used according to the manufacturer's directions to selectively add the eukaryotic 5' cap to triphosphorylated transcripts in total RNA. Next, the product of this reaction was purified using the RNA Clean & Concentrator™-25 kit (Zymo Research) according to manufacturer's instructions, and measured using a NanoDrop One (Thermo Scientific). Then, a modified gene-specific cDNA synthesis was carried out. All samples were diluted to 500 ng in a total volume of 4 µL, then incubated with 1 µL 2 pmol/µL appropriate gene specific reverse primer at 65°C for 5 minutes, and snap cooled on ice for 2 minutes. The following was then added to each sample over ice: 2 µL 5X SuperScript III Buffer (Invitrogen), 1 µL dNTPs (10 mM each), 0.5 µL DTT 100 mM, 0.25 µL RNase Inhibitor Murine (NEB), and either 0.5 µL SuperScript III Reverse Transcript (Invitrogen) or 0.5 µL H_2O for gDNA control samples. These samples were incubated at 50°C for 1 hour. Next 10 µL of a mix containing 2 µL 0.1% bovine serum albumin (NEB), 0.8 µL 50 mM $MnCl_2$, 1 µL 20 µM template switching oligo (see Supplemental Table 1), 2 µL of 5X SuperScript III Buffer, 0.5 µL 100 mM DTT, and either 0.5 µL SuperScript III Reverse Transcript (Invitrogen) or 0.5 µL H_2O (for gDNA control) was added to each sample. After incubating these reactions at 42°C for 90 minutes, they were incubated at 70°C for 10 minutes to inactivate the reverse transcriptase, and kept at 4°C thereafter. Next, RNA was degraded with RNase H (NEB) according to manufacturer's instructions. cDNA cleanup was carried out with the Monarch® PCR & DNA Cleanup Kit (5 µg) (NEB) according to the manufacturer's directions, and measured using a NanoDrop One (Thermo Scientific). Finally, PCR and DNA recovery were carried out as described previously.

Results

Systematic mutations in the PE35-PPE68-esxB-esxA locus to explore regulatory features

In order to better understand the complex regulation of *esxB-A* expression, we sought to identify the mRNA transcripts that contribute to their expression. Since stable transcripts can arise directly from transcription or from transcription and subsequent RNA cleavage, we investigated the impact of several mutations in the promoters and a cleavage site of the *PE35-PPE68-esxB-esxA* locus. To disrupt promotors, the region was either removed, or the -10 consensus sequence was mutated, as detailed in Figure 2. The cleavage site was altered with a single nucleotide mutation at the 3' side of the cleavage site. The individual removal/mutation of each element and various combinations, as detailed in Figure 2, allowed for interpretation of the role each feature plays in expression of the *PE35-PPE68-esxB-esxA* locus.

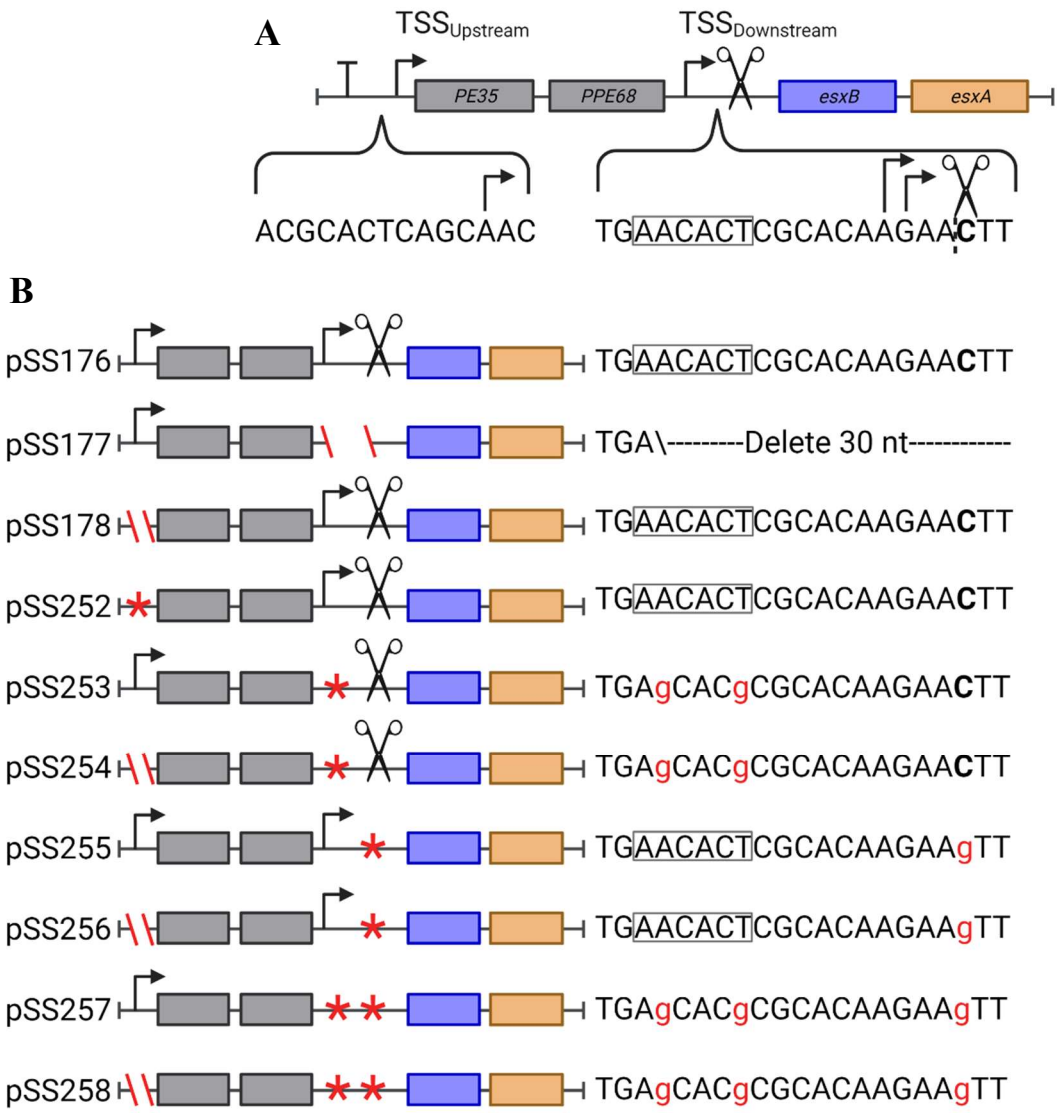

Figure 2. Plasmids for testing the roles of TSSs and cleavage sites in ESX-1. A graphical representation of the plasmids used throughout this paper. A bent arrow represents a TSS and the scissors represent a cleavage site (at the location of the dashed line). A) A key detailing which *Mycobacterium smegmatis* gene each colored box represents in panel B, and showing the sequences surrounding the TSSs and cleavage site. B) The contents of each plasmid are shown next to the corresponding plasmid number, and the sequence in the cleavage site region is shown on the right. A red '-\ \-' indicates a deletion, while a red '*' indicates a mutation predicted to affect transcription or cleavage. Lines and boxes are not drawn to scale. Created with BioRender.com.

$TSS_{Upstream}$ is responsible for expression of PPE68 and contributes to junction region expression

To determine the broad impact each mutation has on expression, we first measured the mRNA abundance at several locations. This provided insight into the effect of each mutation and thus the roles of those components in the native sequence. Thereby these data serve as a guide for future investigation into the mechanisms behind these differences. Figure 3 shows the abundance of *PPE68, esxB*, and *esxA* as well as a portion of the transcript that we refer to as the

junction region (see diagram in Figure 3B) relative to housekeeping gene *sigA*, as determined using quantitative PCR (qPCR). These data show that when the upstream promotor is perturbed the expression of *PPE68* is nearly eliminated. This indicates that this region is solely dependent on $TSS_{Upstream}$. Expression of the junction region, however, appears more complex because it does not follow the same pattern as *PPE68*. The forward primer for the junction region anneals upstream of the $TSS_{Downstream}$, as shown by the green box in Figure 3B, so any transcripts amplified by it must be transcribed from a promotor upstream of this location. As such, it is interesting to observe the continued, albeit lower, expression of the junction region when $TSS_{Upstream}$ is removed. This suggests the at least one additional TSS may be present between $TSS_{Upstream}$ and $TSS_{Downstream}$.

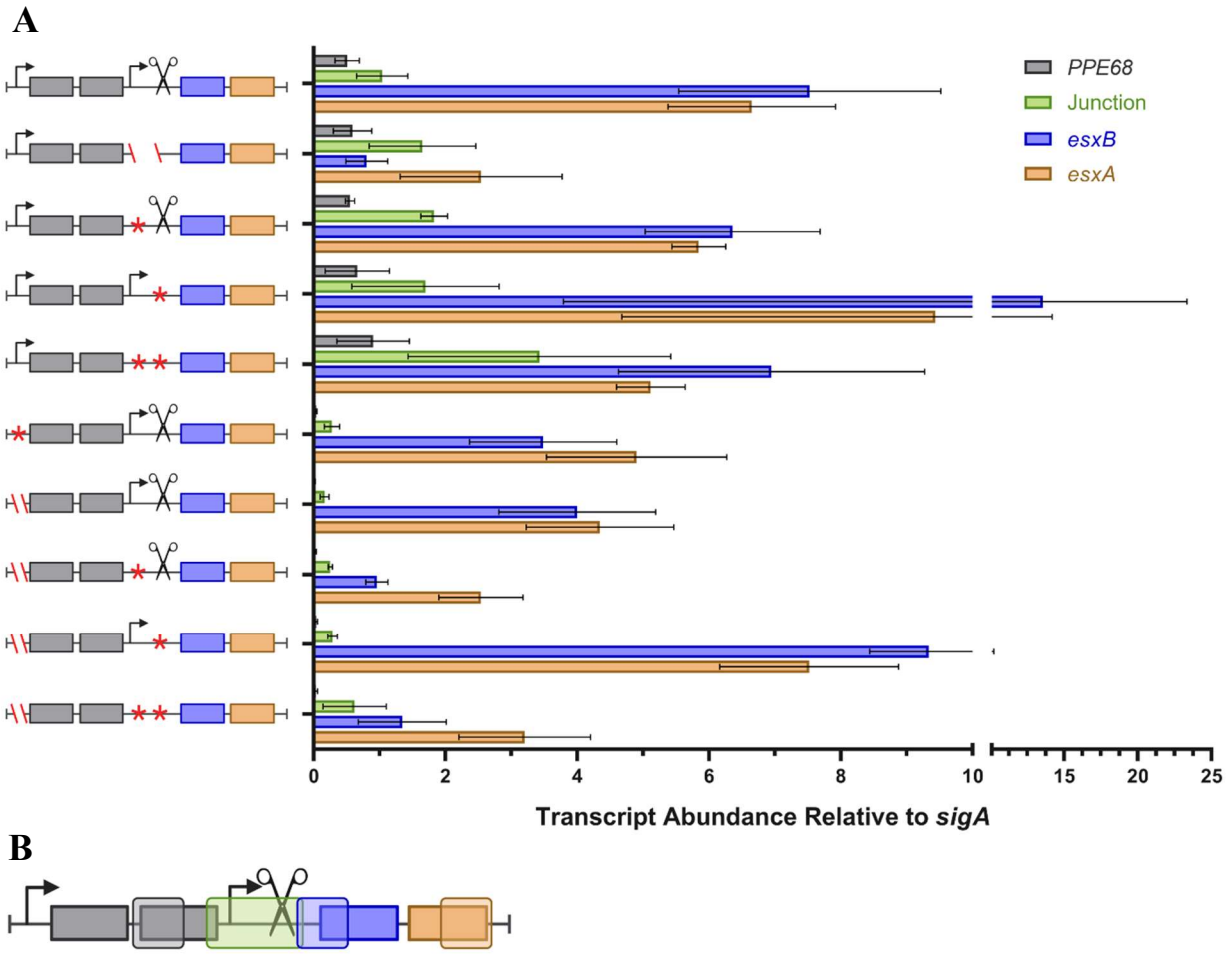

Figure 3. The effects of mutations to $TSS_{Upstream}$, $TSS_{Downstream}$, and $Cleavage_{esxB}$ on Transcript Abundance. A) Analysis by qPCR revealed the transcript abundance of *PPE68*, the junction region, *esxB*, and *esxA* relative to *sigA*. A red '-\ \-' indicates a deletion, while a red '*' indicates a mutation predicted to affect transcription or cleavage. Exact sequences are shown in Figure 2. Error bars show the standard deviation across three biological replicates. B) A graphical representation of the regions amplified to measure expression in panel A. The rounded boxes represent the amplified fragment of *PPE68* (grey), the junction region (green), *esxB* (blue), and *esxA* (amber). Drawing is not to scale but reflects the features encompassed by each primer set. Primer sequences are listed in Supplemental Figure 1. Created with BioRender.com

Mutation of the -10 site is transcriptionally equivalent to promoter deletion

Deletion of the first 30 nucleotides downstream of the *PPE68* coding region showed a dramatic decrease in *esxB* and *esxA* transcript abundance. In order to investigate the sequence contained within those 30 nucleotides more precisely, a more targeted change is needed to disrupt single features. The -10 site is a good target for mutations to disrupt most promoters and is often easy to recognize. This approach was validated using the known $TSS_{Upstream}$. Figure 3 showed that the deletion of the 100 nucleotides upstream of $TSS_{Upstream}$ results in transcript levels that are comparable to mutation of the -10 consensus site of the promotor for $TSS_{Upstream}$. When compared to the construct containing the native promoter for $TSS_{Upstream}$, both the deleted and mutated plasmids nearly eliminated expression of *PPE68* and markedly reduced expression of the junction region. Together, this supports the successful disruption of transcription at $TSS_{Upstream}$ by mutation of the -10 site and suggests this may be an effective approach for disrupting the $TSS_{Downstream}$ as well.

esxB and esxA transcript abundance cannot be fully explained by the proposed features

Examination of *esxB* and *esxA* transcript levels suggests their regulation is complex. Perturbing $TSS_{Upstream}$ does not abolish expression of *esxB* or *esxA*, as it does for *PPE68*, supporting the presence of at least one additional TSS. We previously hypothesized that $TSS_{Downstream}$ could account for this $TSS_{Upstream}$-independent expression. However, disruption of both promotors only reduced expression of *esxB* and *esxA* instead of diminishing them to levels similar to *PPE68*. This partial reduction supports the presence of a promoter giving rise to $TSS_{Downstream}$, but also suggests at least one additional promoter exists in the *PE35-PPE68-esxB-esxA* locus.

Interestingly, mutating the cleavage site appeared to increase expression of *esxB* and *esxA*. It is challenging to draw many conclusions about this mutation from Figure 3 alone due to the large error bars and unknown impact of this mutation on cleavage. A notable takeaway however is that the plasmid with perturbations in all three proposed features produces appreciable expression of *esxB* and even higher expression of *esxA*. This also suggests that these three features alone cannot account for all the expression of these two transcripts.

Addition of a transcriptional terminator reduced esxB-A transcript abundance

The relative expression levels observed here differ in quantity but show similar patterns to those previously obtained using strains without a transcriptional terminator upstream of the *PE35-PPE68-esxB-esxA* cassette. Unpublished data collected previously, shown in Figure 4, depict a dramatically higher transcript abundance of *esxB* and *esxA* relative to *sigA* (Shell et al., unpublished) than is shown in Figure 3. However, the fold change observed when the first 30 nucleotides of the junction region are deleted is similar. The Shell et al. data shows a roughly 10-fold and 5-fold decrease in *esxB* and *esxA* in this mutant. Figure 3 showed an average 9-fold and 3-fold decrease in *esxB* and *esxA* is observed in the 30nt mutant. Possible explanations for the differences include changes to the plasmids used and experimental variability. The Shell et al. data were generated using plasmids similar to the ones used here, with the only difference being the absence of the tsynA transcriptional terminator upstream of the genes of interest. This terminator was inserted to prevent unintentional transcription of the genes of interest from

spurious upstream promotors in the plasmid (Czyz et al., 2014). However, given that *PPE68* expression is similar in the two datasets, it is not obvious why the upstream terminator would affect expression of *esxB* and *esxA*. There may be other, unintended sources of variability between the experiments.

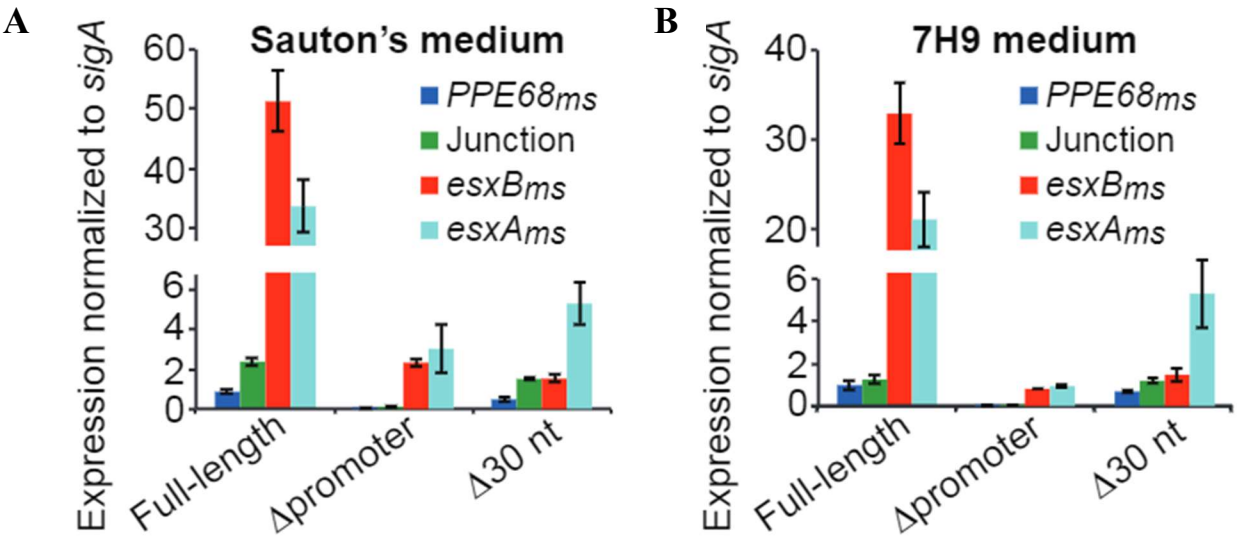

Figure 4. Transcript abundance in similar plasmids shows higher levels but similar patterns. The transcript abundance of *PPE68,* the junction region, *esxB,* and *esxA* were measured relative to *sigA* using qPCR. The plasmids evaluated here are identical to those presented in Figure 2 with the omission of the upstream transcriptional terminator. Cells were cultured in either A) Sauton's medium or B) 7H9 medium. Error bars show the standard deviation of three biological replicates. Credit to S. Shell.

Expression of esxB and esxA is variable between experiments

Expression of the native *PE35-PPE68-esxB-esxA* locus was measured in previous experiments by other researchers in the lab. Interestingly, there was notable variability among the experiments. Figure 5 shows three such datasets where expression of the endogenous *PE35-PPE68-esxB-esxA* locus genes were measured relative to *sigA*. RNA was collected from *M. smegmatis* cultures grown in 7H9, using the same RNA extraction and qPCR protocols, differing only in the strains used and researcher performing the procedures. The 'WT' sample is from strain mc^2155 without any plasmids and was performed by Diego Vargas Blanco. Both the 'CRISPRi control' and 'CRISPRi hyg control' were strain mc^2155 cells transformed with an inducible CRISPRi system in the L5 site, but were not treated with the inducer ATc and contained a non-targeting sgRNA. These samples were measured by Ying Zhou. The 'CRISPRi hyg control' also has an addition of the tetO operator and hygromycin resistance marker adjacent to the endogenous RNase E gene, but again no changes are induced in the absence of ATc. Thus, none of these modifications are expected to change expression of the *PE35-PPE68-esxB-esxA* locus or *sigA*, yet Figure 5 shows considerable variation. Even independent experiments using the same strains performed by the same researcher have yielded notable differences in expression (data not shown). The striking degree of variance seems to indicate seemingly minor differences between experimental setups and their execution by different researchers notably impacts expression of these genes. Regardless of the fluctuations observed in Figure 5, the general pattern of higher expression of *esxB* and *esxA* relative to *PPE68* and the junction region

expression remain, supporting the existence of additional regulatory features beyond $TSS_{Upstream}$ alone.

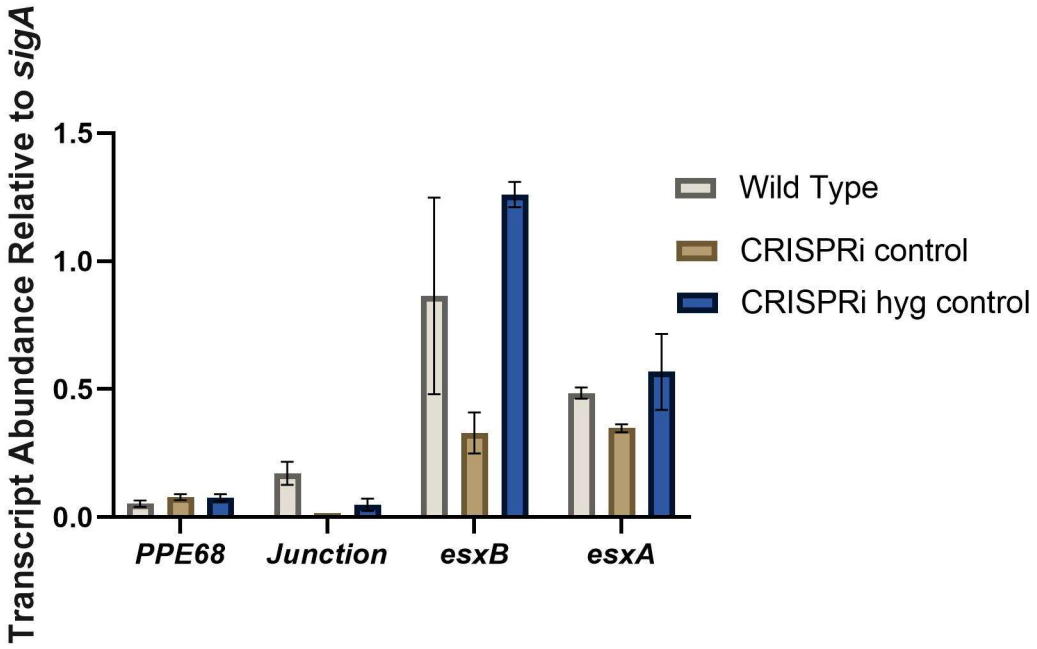

Figure 5. Variation in the abundance of *PE35-PPE68-esxB-esxA locus* transcripts. Analysis by qPCR revealed the transcript abundance of *PPE68*, the junction region, *esxB*, and *esxA* relative to *sigA* in *M. smegmatis*. The "Wild Type" experiment was performed independently from the CRISPRi experiments. The endogenous copies of these genes were identical in all strains shown here, and the genetic differences between the three strains are not expected to change expression of the *PE35-PPE68-esxB-esxA* locus or *sigA*. Data provided by Ying Zhou and Diego Vargas Blanco. The CRISPRi control strain has been previously characterized (Rock et al., 2017).

Validation of 5' RACE through mapping $TSS_{Upstream}$

Given the numerous questions raised by the qPCR data in Figure 3, further analysis was needed to better understand what transcripts constitute the total mRNA expressed for each gene. 5' Rapid Amplification of cDNA Ends (5' RACE) was used to identify the different transcripts present in several of the mutated plasmids. 5' RACE was selected because it can confirm the presence of 5' ends we expected to find and act as a tool for finding additional 5' ends. This method produces amplification of either monophosphorylated transcripts (cleaved) in untreated samples, or both triphosphorylated and monophosphorylated transcripts (primary and cleaved) in samples treated with RppH which converts triphosphates into monophosphates.

To aid in troubleshooting and ultimately validate the procedure, we first looked for the 5' end resulting from $TSS_{Upstream}$. Figure 6 shows several bands of the expected size (192 base pairs) in plasmids with the normal $TSS_{Upstream}$ in place, while those with the deletion show different bands. Sanger sequencing confirmed that the bands running at approximately 200 base pairs are in fact the result of $TSS_{Upstream}$. We note that this band was present in some of the no-RT control samples as well, reflecting apparent contamination of those samples with RT.

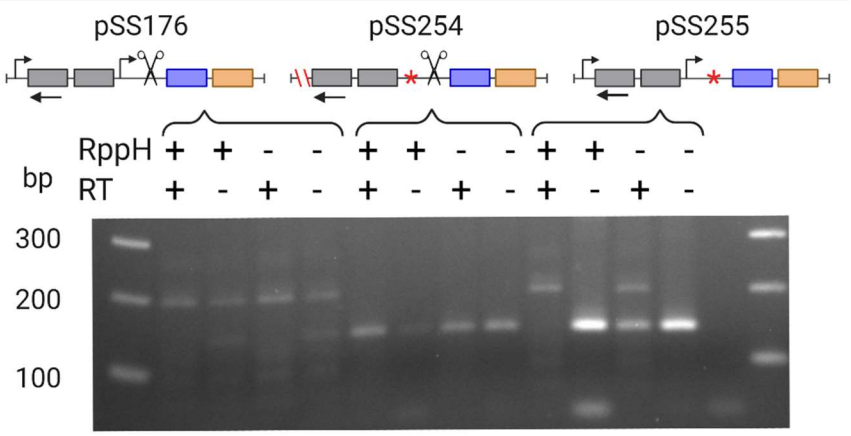

Figure 6. 5' RACE of the upstream region validates the presence of $TSS_{Upstream}$. Separation of amplified 5' RACE products by gel electrophoresis revealed multiple bands. Primer SSS987 (represented by a black arrow under the plasmid diagram) was used to amplify 5' ends in cells expressing pSS176 (left), pSS254 (center), and pSS255 (right). A red '-\ \-' indicates a deletion, while a red '*' indicates a mutation predicted to affect transcription or cleavage. Exact plasmid sequences are shown in Figure 2. Sanger sequencing showed that the bands running at approximately 200 nucleotides are generated by $TSS_{Upstream}$, and the bands running at approximately 150 nucleotides are the result of spurious amplification of 23S ribosomal RNA. Created with BioRender.com.

Amplification in controls without RT due to procedural flaw not gDNA

During cDNA synthesis a mock reaction is carried out without reverse transcriptase (RT) for each sample. Since no new DNA is synthesized in this step, any amplification seen in PCR reactions of the mock sample will reveal the presence of gDNA. During the previous qPCR analysis, the mock samples showed very little expression, however in 5' RACE amplification is seen in nearly all of the samples without RT. In each case, sequencing of the bands in these samples produces the same sequences as bands of equivalent sizes in RT treated samples. Given that the same products are amplified, it seems unlikely to be gDNA because the forward primer used in PCR anneals to the universal adaptor which is selectively ligated to monophosphorylated mRNA transcripts. Without this adaptor sequence gDNA should not be amplified. Close examination of the procedure used in cDNA synthesis and cleanup samples revealed a flaw that can explain these observations. During cDNA cleanup, first the RT samples were purified via column chromatography, then the columns were rinsed and re-used for the corresponding mock sample. When performing this procedure on biological replicates it was modified to use fresh columns for the mock samples. Subsequent PCR and gel electrophoresis, seen in Figure 7, showed no amplification in the mock samples. This supports the conclusion that a flaw in the cDNA cleanup procedure produced the unexpected amplification seen in samples not treated with RT. This contamination of the mock samples by re-use of the columns was not detected in qPCR likely as a result of the highly sensitive and quantitative nature of the procedure. In contrast, the 5' RACE procedure relies on a very high number of amplification cycles, and sanger sequencing is not designed to provide data on abundance, so a small amount of contamination from an RT sample would be hard to distinguish.

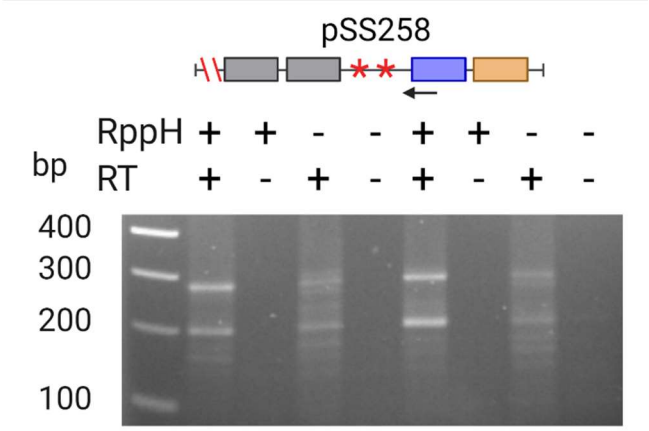

Figure 7. 5' RACE with optimized protocol reveals flaw responsible for amplification in No-RT samples seen in other gels. Fresh columns were used during cDNA cleanup of samples not treated with RT. Subsequent PCR on pSS258 samples using SSS538 shows no amplification in the samples without RT. Created with BioRender.com.

Identification of two 5' ends from $TSS_{Downstream}$ and one from $Cleavage_{esxB}$

Using our validated 5' RACE procedure, we next worked to confirm the presence of the $TSS_{Downstream}$ and $Cleavage_{esxB}$. 5' RACE was employed to identify the exact location of each feature and investigate the outcome of mutations made to the various plasmids.

Due to the close proximity of 5' ends produced by $TSS_{Downstream}$ and $Cleavage_{esxB}$, gel electrophoresis could not separate the resulting bands, which often contained multiple 5' ends. Close examination of the sequencing traces enabled detection of this mixture of 5' ends in many of the bands. One example of this analysis is shown in Figure 8B for a band cut from the gel in Figure 8A. In looking at these sequencing traces, mixed peaks are prevalent before the cleavage site, while high confidence single peaks are observed after the cleavage site. Since sequencing starts with the gene-specific reverse primer downstream of these features, this change in sequencing confidence suggests that the band contains a mixture of 5' ends. Comparing these mixed peaks in Figure 8B to the expected sequence of the cleavage site, and the $TSS_{Downstream}$ if it were to start 3 or 4 nucleotides upstream of the cleavage site, revealed this band is a mixture of all three. This detailed multipeak analysis was carried out for all bands sequenced.

5' RACE with several strains confirmed the presence and dual starting bases of $TSS_{Downstream}$. Figure 8A clearly shows the presence of a product of the expected 179 nucleotides in plasmids pSS176, pSS178, and pSS255 treated with RppH and reverse transcriptase (RT). Sequencing of this band, shown in the lower trace of Figure 8B, confirmed the presence of a mixture of 5' ends including 3 and 4 nucleotides upstream of the cleavage site and the cleavage product itself. Additionally, $TSS_{Downstream}$ was never seen in pSS254, suggesting mutation of the -10 consensus site successfully perturbed this promotor. It is noteworthy that pSS254 shows amplification, despite deletion of the promoter for $TSS_{Upstream}$ and successful mutation of the promoter for $TSS_{Downstream}$, thereby suggesting the presence of a third promoter.

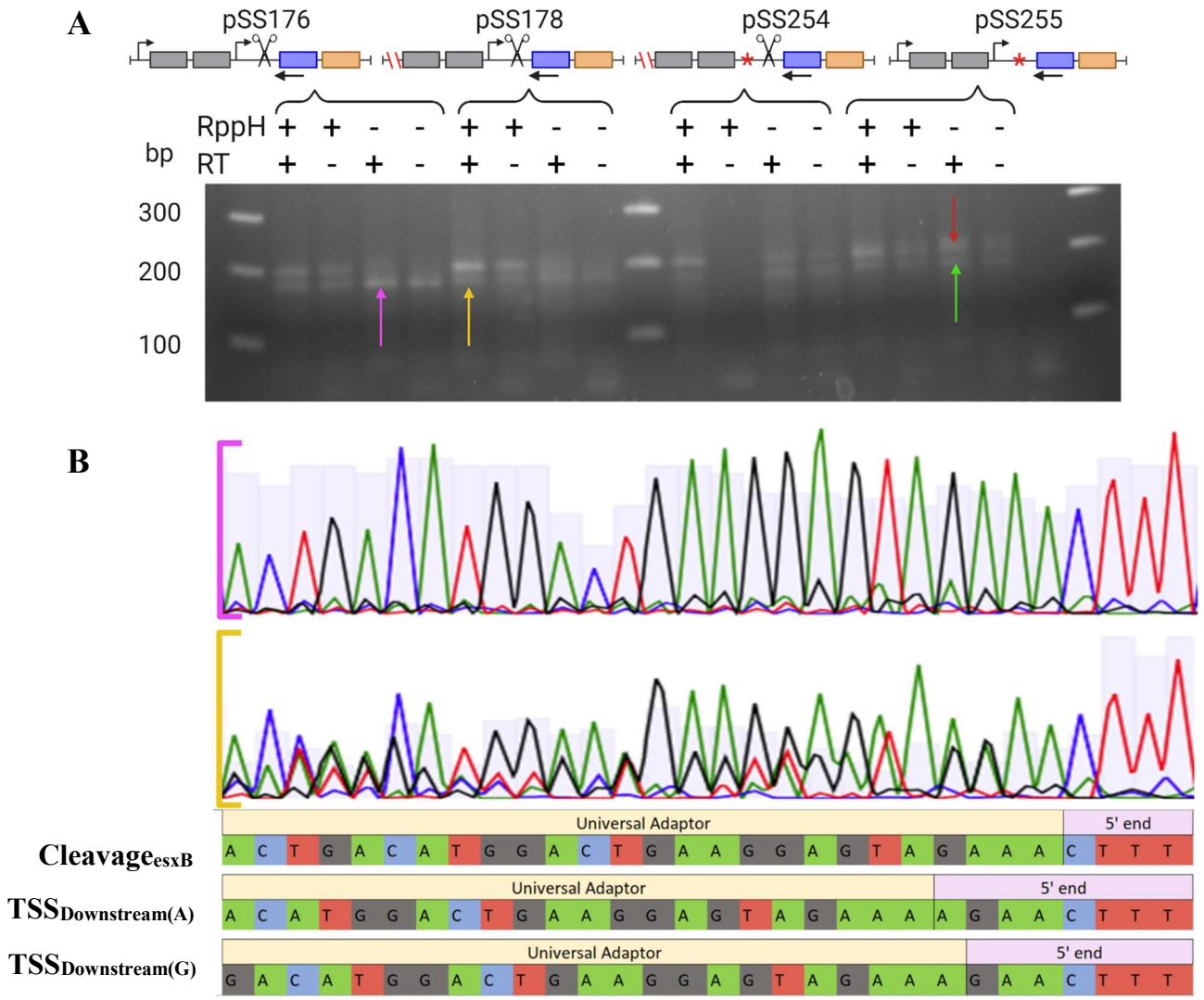

Figure 8. 5' RACE examining the $TSS_{Downstream}$ region reveals two TSSs and the $Cleavage_{esxB}$ site. Gel electrophoresis of 5' RACE products and subsequent analysis of sequencing reads. A) Primer SSS538 (represented by a black arrow under the plasmid diagrams) was used to amplify 5' ends in cells expressing pSS176, pSS178, pSS254, and pSS255 (from left to right). A red '-\ \-' indicates a deletion, while a red '*' indicates a mutation predicted to affect transcription or cleavage. Exact plasmid sequences are shown in Figure 2. B) Sanger sequencing traces from the lower band in pSS176 without RppH and with RT (pink arrow, upper trace), and the lower band in pSS178 with RppH and RT (yellow arrow, lower trace). All of the peaks are explained by alignment with the expected 5' ends resulting from the cleavage site and TSSs 3 and 4 nt upstream of the cleavage site. Created with BioRender.com.

Similarly, 5' RACE of samples not treated with RppH confirmed the presence of the cleavage site, as shown by several bands in Figure 8A. Sequencing of these bands, shown in the upper trace of Figure 8B confirms that cleavage occurs as annotated in Figure 2. Generally, the samples not treated with RppH showed a strong enrichment of the cleaved 5' end relative to the $TSS_{Downstream}$ in the sequencing peaks, as in Figure 8B. In some cases, the 5' ends resulting from the primary transcripts were still mixed into these bands, likely as a result of endogenous triphosphate to monophosphate conversion.

The cleavage site mutation causes a change in cleavage location

Most *M. smegmatis* mRNA cleavage sites were found to occur immediately upstream of cytosines (Martini et al., 2019). We therefore predicted that mutation of the cytosine on the 3' side of the cleavage site to a guanine would impact the cleavage event. Indeed, this single mutation was enough to result in two different cleavage products. The upper band seen in Figure 8A of pSS255 was sequenced and shows that cleavage occurred 28 nucleotides upstream of the normal cleavage site (-28), while the band just below it contained mostly cleavage products with a 5' end 1 nucleotide upstream of the normal cleavage site (-1). Additional PCR and extraction reactions were carried out to confirm this observation, and representative sequence traces are shown in Figure 9. Cleavage at the original position was never detected in this mutant, although the mixed peaks in these traces are challenging to fully interpret and cannot be fully explained by any combination of the 5' ends reported here. The profound impact of this single nucleotide mutation on the cleavage event may impact post-transcriptional regulation of the transcript in a variety of ways.

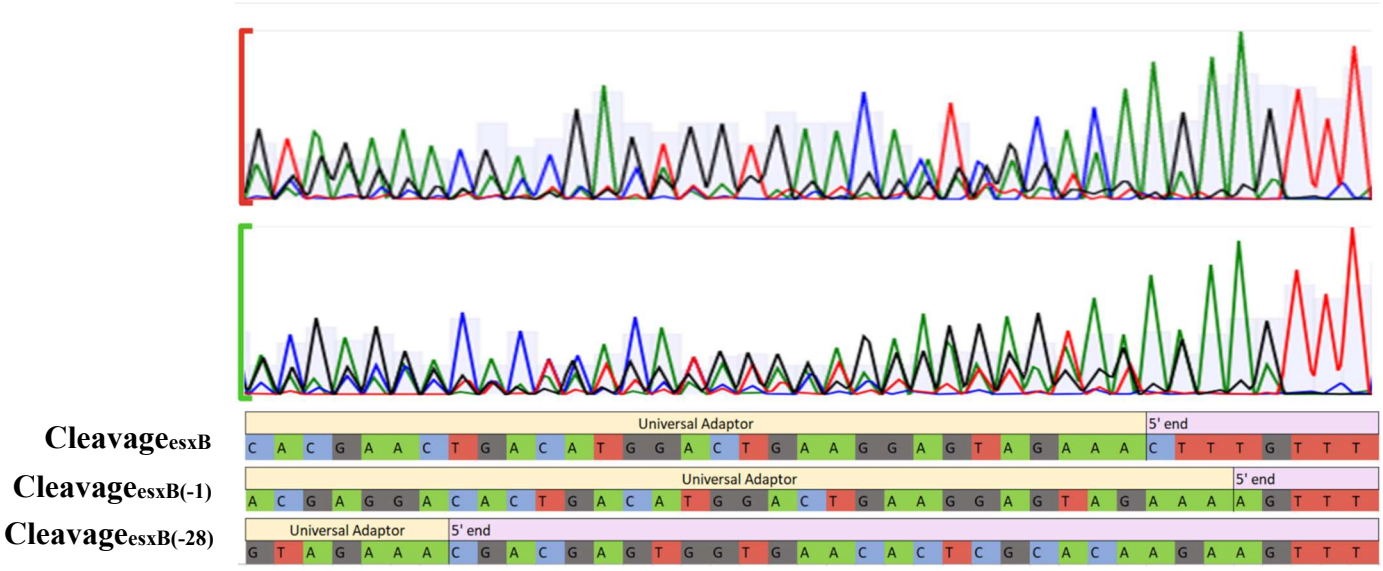

Figure 9. 5' RACE examining Cleavage$_{esxB}$ mutants reveal two alternate cleavage locations. Sanger sequencing traces from pSS255 without RppH and with RT. The upper and lower bands shown in Figure 8A (red and green arrows in Figure 8A, and shown in the same order here) primarily contain Cleavage$_{esxB(-28)}$ or Cleavage$_{esxB(-1)}$ transcripts respectively. The original Cleavage$_{esxB}$ site was not detected in these mutants, but the multiple peaks are challenging to interpret, and cannot be fully explained by known features. Created with BioRender.com.

No TSS was observed in the region upstream of esxA

According to the transcript abundance in Figure 3, *esxA* maintains a higher level of expression than *esxB* when both TSS$_{Upstream}$ and TSS$_{Downstream}$ are perturbed. In order to explain this, we considered the potential for another TSS located just upstream of *esxA*. 5' RACE was performed with a primer annealing in the *esxA* coding region, and several bands were seen. However, sequencing revealed these bands to be the result of imperfect binding of the PCR primer for the universal adaptor to the plasmid sequence directly, not the result of a genuine 5' end. A faint band was seen that mapped to TSS$_{Downstream}$ in all samples where it was not mutated,

validating the methodology. Together this suggests that an additional TSS is not present in the untranslated region upstream of *esxA* or in the *esxB* coding region. It is unclear what mechanism allows for the differential expression of *esxB* and *esxA* observed in the context of plasmids lacking $TSS_{Upstream}$ and $TSS_{Downstream}$, but perhaps it occurs post-transcriptionally. A summary of all the findings from 5' RACE can be found in Figure 10.

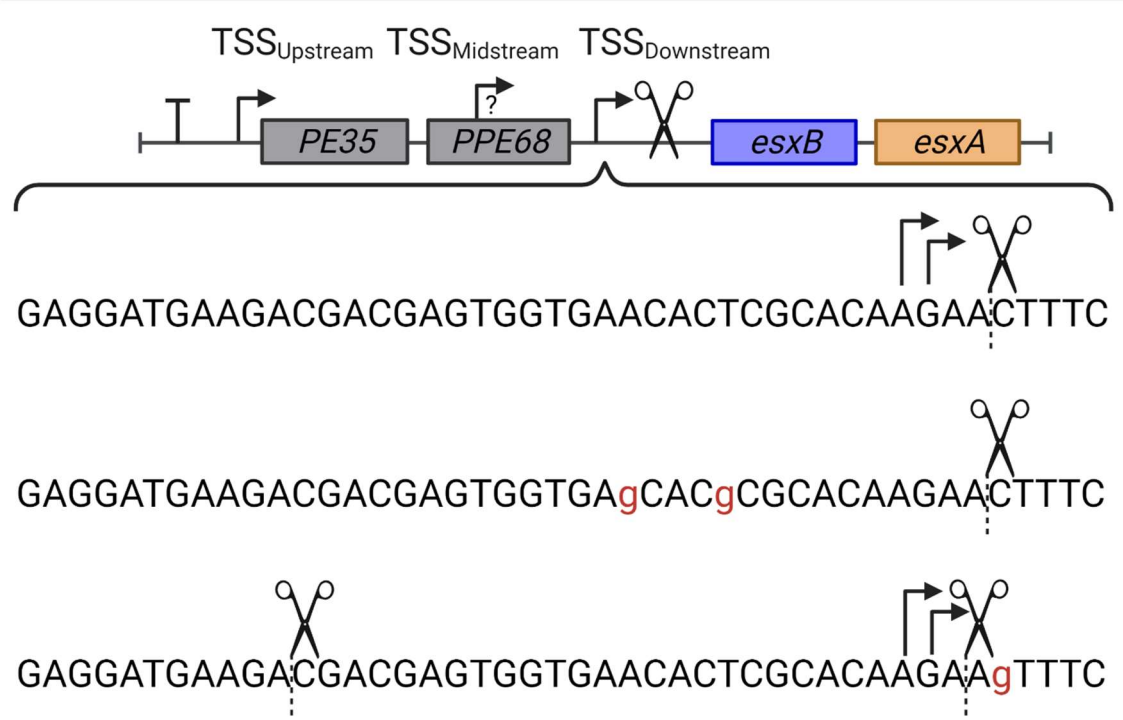

Figure 10. Summary of 5' RACE findings. Multiple amplification and sequencing attempts show different mixtures of 5' ends in the context of different mutations. The bent arrow represents a TSS, the scissors represent a cleavage site (at the location of the dashed line). $TSS_{Midstream}$ has not been successfully mapped, but the data suggest the presence of at least one additional TSS. Created with BioRender.com

The search for an additional TSS in the PPE68 coding region was inconclusive

The qPCR data regarding expression of the junction region and *esxB-A* in the absence of both $TSS_{Upstream}$ and $TSS_{Downstream}$, and 5' RACE mapping of a cleaved transcript in pSS254 samples support the presence of at least one additional TSS in the locus. We attempted to use our 5' RACE method to search for this additional "$TSS_{Midstream}$" in the *PPE68* coding region or 5' untranslated region. Using a primer that anneals just upstream of the $Cleavage_{esxB(-28)}$ we found a number of bands were amplified, as shown in Figure 11A. The band running at approximately 500 base pairs was found to be 16S ribosomal RNA, and the band running at approximately 300 base pairs was not able to be conclusively sequenced. Multiple attempts were made to amplify and sequence this lower band, but all yielded poor quality, mixed sequence traces that did not align with the universal adaptor or plasmid sequence.

Additionally, primers annealing at the beginning of the *PPE68* coding region were used to search for $TSS_{Midstream}$ there. Two primers were used, and each amplified different bands as seen in Figure 11B. Sequencing results for these bands were not available at the time of submission of this document. Overall, several pieces of evidence support the existence of $TSS_{Midstream}$, but it has yet to be mapped.

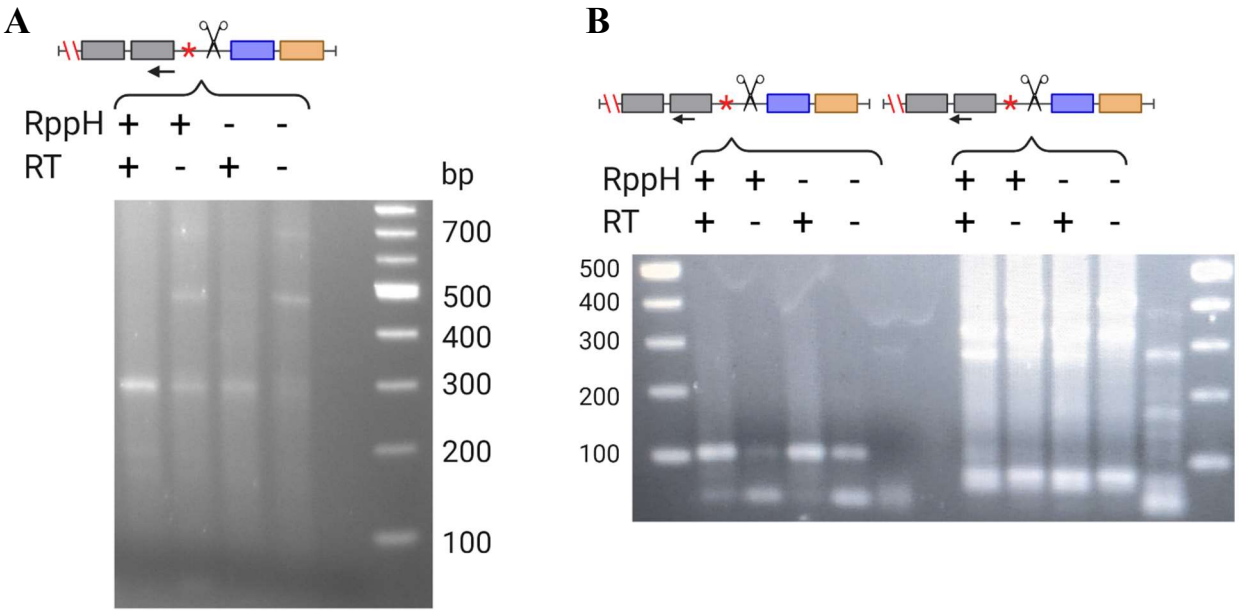

Figure 11. Search for $TSS_{Mistream}$ with multiple primers was inconclusive. We performed 5' RACE with different primers in the *PPE68* coding region to search for one or more TSS in this region using pSS254. A) SSS2255 anneals just upstream of the $Cleavage_{esxB(-28)}$ site, and generates several bands. Sequencing shows the 500 bp band is 16S rRNA, and the band at 300 bp was not interpretable. B) SSS544 (left) anneals about 500 nucleotides into the *PPE68* coding region. SSS546 (right) anneals 32 nucleotides into the *PPE68* coding region. Sequencing results from these bands were not available at the time of submission of this document. Created with BioRender.com.

Further troubleshooting is required for Capping RACE

One limitation of 5' RACE is that it can only isolate cleaved transcripts or provide both primary and cleaved transcripts, but cannot isolate primary transcripts alone. This is especially challenging when features are so close together such that they cannot be separated by gel electrophoresis, as is the case with $TSS_{Downstream}$ and $Cleavage_{esxB}$. Due to this, sequencing many of the bands yielded mixed peaks that were challenging to interpret, as seen in Figure 8 and Figure 9. In order to address this limitation we attempted to perform Capping-RACE as detailed by Liu et al., 2018 who developed this method to map novel TSSs in several prokaryotic species. The procedure selectively captures 5' ends with triphosphates and was intended to be complimentary to traditional 5' RACE. However, our first attempt at the procedure to map $TSS_{Upstream}$ was unsuccessful. Using the same primer which was successful in Figure 6 (SSS987), Capping RACE resulted in two bands in each sample. Most sequencing results contained heavily mixed peaks that were uninterpretable, and those with defined peaks did not align with the plasmid nor the universal template switching oligo. A touchdown PCR with a different reverse primer was attempted but yielded even more bands that could not be interpreted. Given these

results, further troubleshooting is needed on this procedure. If these challenges are resolved this method may prove useful for mapping additional unknown TSS in the *PE35-PPE68-esxB-esxA* locus or other loci. However, since we were able to interpret the mixed peaks provided by 5' RACE to answer the original research question, further work with this procedure was not carried out.

Discussion

The expression of *esxB* and *esxA*, important virulence factors in mycobacterium, cannot be explained by a single promotor. Previous work showed that in *M. smegmatis*, deletion of 30 nucleotides in the intergenic region between *PPE68* and *esxB* greatly reduced expression of *esxB-A* (Shell et al., unpublished). To investigate features of this region that may contribute to gene expression, we explored the impacts of mutating a putative TSS and cleavage site located in that region. In the strains tested, the canonical *PE35, PPE68, esxB,* and *esxA* genes were deleted, and an engineered plasmid was introduced containing a transcriptional terminator followed by the modified genes. Quantification of transcript abundance showed that *PPE68* depends entirely on $TSS_{Upstream}$. On the other hand, *esxB* and *esxA* transcripts were expressed by more than a single TSS, as they still showed some expression when $TSS_{Upstream}$ was deleted. Mutation of the -10 site of $TSS_{Downstream}$ reduced expression to levels similar to the 30-nucleotide deletion, supporting the contribution of $TSS_{Downstream}$ to *esxB-A* expression. However, the combined deletion of $TSS_{Upstream}$ and successful disruption of $TSS_{Downstream}$ did not fully abolish expression of *esxB-A*. Additionally, transcript abundance of the junction region did not match that of *PPE68*. Together, these observations suggest that there may be a third TSS between $TSS_{Upstream}$ and $TSS_{Downstream}$.

Further exploration with 5' RACE identified 5' ends from $TSS_{Upstream}$, $TSS_{Downstream}$, and the cleavage site. Moreover, the $TSS_{Downstream}$ can start at either an adenine or an adjacent guanine. Interestingly, mutation of the cytosine at the cleavage site changes the location of cleavage. The presence of transcripts cleaved at this site in cells with the combined deletion of $TSS_{Upstream}$ and mutation of $TSS_{Downstream}$ further supports the conclusion that at least one additional TSS is located between those explored here.

The findings of this project lay the groundwork for future studies of the expression and regulation of ESX-1 in mycobacteria. Identifying the active TSSs and cleavage site producing transcripts encoding these virulence factors is an essential step in better understanding their regulation. The confirmation of these TSSs provides multiple opportunities for various mechanisms of transcriptional regulation of these crucial virulence factors. The sequence of the promotor, especially the -10 and -35 sites, can be examined for potential regulatory molecule binding sites, siRNA binding sites, secondary structure, and other features that may provide a mechanism for transcriptional regulation(s). Furthermore, an understanding of transcriptional regulation is key to informing other higher-level analyses. For example, prior work on this locus performed at the protein level was hard to interpret due to high expression in all strains including what was intended as a promotorless control. The confirmation of $TSS_{Downstream}$ here provides an explanation for this expression, since all of these plasmids unintentionally contained this

promotor. Now armed with a more detailed understanding of the transcriptional landscape of this locus, protein level studies can be effectively carried out.

With regard to our findings on the cleavage site mutation, prior work has shown that the 5' ends of cleaved transcripts in *M. smegmatis* are very strongly biased towards cytosine (Martini, et al. 2019). It is interesting to see here that mutation of this nucleotide changes the location of cleavage. One of the two new cleavage sites also generates a 5' cytosine, which supports the importance of this nucleotide in the cleavage mechanism. The other alternate cleavage site is only one nucleotide away from the original site and does not leave a cytosine at the 5' end. This may suggest that other features of the sequence or mRNA transcript are important in targeting of the ribonuclease to its target.

Overall, these results shed light on the complex regulatory mechanisms behind expression of the *PE35-PPE68-esxB-esxA* locus in *M. smegmatis*. It is not entirely clear, however, what implications these findings have for Mtb. The answer may lie in the differences that give the ESX-1 system in Mtb such extraordinary virulence in comparison to *M. smegmatis*. It is possible that the two species share similar regulatory features but differ in the protein itself. Studies have shown that inside a phagosome, *M. smegmatis* secretes EsxA, but fails to escape into the host cytosol. Biochemical analysis revealed that the *M. smegmatis* protein is unable to interact with membranes in the same way as the Mtb ortholog (De Leon et al., 2012). Therefore, it is possible that similar regulatory mechanisms in both species enable expression when the cell is inside a phagosome, but the functionally inept EsxA renders *M. smegmatis* less virulent. However, it is also conceivable that there are important variations in the regulation of *esxB-A*. It would be interesting to investigate the regulatory features of the Mtb ESX-1 locus using an experimental design similar to the one presented here.

Supplemental Information

Supplemental Table 1. Oligonucleotides. The description and sequence of all oligonucleotides used throughout the study. Letters preceded by a lowercase "r" are ribonucleic acids, while all others are deoxyribonucleic acids. 'Fw' denotes a forward primer, and 'Rv' denotes a reverse primer.

Description	Name	Sequence (5' -> 3')
Fw *PPE68*	SSS545	TTTCAGCAGCGAGAGTTAGG
Rv *PPE68*	SSS546	GTGTTGATCTCCGGTGGTAG
Fw Junction	SSS549	GACCTCTCCGAGGATGAAGA
Rv Junction	SSS550	CTGTGTCCTCACCAATTCCA
Fw e*sxB*	SSS537	GGTGAGGACACAGGGAAATAAG
Rv e*sxB*, $TSS_{Downstream}$	SSS538	CGGAGATGCGCTCGAAAT
Fw *esxA*	SSS695	CTCCAACGAGCTGAACCT
Rv *esxA*	SSS696	GGCAAACATTCCCGTGAC
5' RACE adaptor	SSS1016	CTGGAGCACGAGGACACTGACATGGACTGAAG GAGTrArGrArArA
Fw adaptor primer	SSS1017	CTGGAGCACGAGGACACTGA
Rv *PE35*, $TSS_{Upstream}$	SSS987	GGATTGTGTGTCATCGGTTG
Rv *PE35*, $TSS_{Upstream}$ nested	SSS554	GTGTGTCATCGGTTGCATTTG
Rv end of *PPE68*	SSS2255	TTCATCCTCGGAGAGGTCGT
Rv start of *PPE68*	SSS544	CGAAGTAGTCCATCTCGTTGAG
Template switching oligo (TSO)	SSS2217	ACACTCTTTCCCTACACGACGCTCTTCCGATCT rGrGrG
Fw TSO primer	SSS2218	ACACTCTTTCCCTACACGACGCTCTTCCGATCT

Supplemental Table 2. PCR conditions. The specific annealing temperature and extension time used for each primer set used in PCR. Any pairs that were not used in this study have a blank entry in the given annealing temperature column.

Oligo Name	Annealing temperature with SSS1017	Annealing temperature with SSS2218	Extension Time
SSS544	53°C		1 minute
SSS538	54°C		12 seconds (mapping $TSS_{Downstream}$) 2 minutes (search for $TSS_{Midstream}$)
SSS696	53°C		42 seconds
SSS987	52°C	58°C	12 seconds
SSS2255	56°C		2 minutes
SSS546	54°C		1 minute
SSS554		63°C – 48°C Touchdown	15 seconds

Works Cited

Abdallah, A. M., Weerdenburg, E. M., Guan, Q., Ummels, R., Borggreve, S., Adroub, S. A., Malas, T. B., Naeem, R., Zhang, H., Otto, T. D., Bitter, W., & Pain, A. (2019). Integrated transcriptomic and proteomic analysis of pathogenic mycobacteria and their esx-1 mutants reveal secretion-dependent regulation of ESX-1 substrates and WhiB6 as a transcriptional regulator. *PLoS ONE*, *14*(1), 1–24. https://doi.org/10.1371/journal.pone.0211003

Augenstreich, J., Arbues, A., Simeone, R., Haanappel, E., Wegener, A., Sayes, F., Le Chevalier, F., Chalut, C., Malaga, W., Guilhot, C., Brosch, R., & Astarie-Dequeker, C. (2017). ESX-1 and phthiocerol dimycocerosates of Mycobacterium tuberculosis act in concert to cause phagosomal rupture and host cell apoptosis. *Cellular Microbiology*, *19*(7), 1–19. https://doi.org/10.1111/cmi.12726

Baga, M., Goransson, M., Normark, S., & Uhlin, B. E. (1988). Processed messenger RNA with differential stability in the regulation of Escherichiacoli pilin gene expression. *Cell*, *52*, 197–206.

Berthet, F. X., Rasmussen, P. B., Rosenkrands, I., Andersen, P., & Gicquel, B. (1998). A Mycobacterium tuberculosis operon encoding ESAT-6 and a novel low-molecular-mass culture filtrate protein (CFP-10). *Microbiology*, *144*(11), 3195–3203. https://doi.org/10.1099/00221287-144-11-3195

Bosserman, R. E., & Champion, P. A. (2017). Esx systems and the mycobacterial cell envelope: What's the connection? *Journal of Bacteriology*, *199*(17), 1–16. https://doi.org/10.1128/JB.00131-17

Conrad, W. H., Osman, M. M., Shanahan, J. K., Chu, F., Takaki, K. K., Cameron, J., Hopkinson-Woolley, D., Brosch, R., & Ramakrishnan, L. (2017). Mycobacterial ESX-1 secretion system mediates host cell lysis through bacterium contact-dependent gross membrane disruptions. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(6), 1371–1376. https://doi.org/10.1073/pnas.1620133114

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), 1–10. https://doi.org/10.1128/mBio.00931-14

deRivera, J. H., & Shell, S. S. (2016). *The Effects of Post-Transcriptional Processing on mRNA Stability in M. smegmatis*. *March*, 1–31.

De Jonge, M. I., Pehau-Arnaudet, G., Fretz, M. M., Romain, F., Bottai, D., Brodin, P., Honoré, N., Marchal, G., Jiskoot, W., England, P., Cole, S. T., & Brosch, R. (2007). ESAT-6 from Mycobacterium tuberculosis dissociates from its putative chaperone CFP-10 under acidic conditions and exhibits membrane-lysing activity. *Journal of Bacteriology*, *189*(16), 6028–6034. https://doi.org/10.1128/JB.00469-07

De Leon, J., Jiang, G., Ma, Y., Rubin, E., Fortune, S., & Sun, J. (2012). Mycobacterium tuberculosis ESAT-6 exhibits a unique membrane-interacting activity that is not found in its ortholog from non-pathogenic Mycobacterium smegmatis. *Journal of Biological Chemistry*, *287*(53), 44184–44191. https://doi.org/10.1074/jbc.M112.420869

Houben, E. N. G., Korotkov, K. V., & Bitter, W. (2014). Take five - Type VII secretion systems of Mycobacteria. *Biochimica et Biophysica Acta - Molecular Cell Research*, *1843*(8), 1707–1716. https://doi.org/10.1016/j.bbamcr.2013.11.003

Kelly, J. M, & Shell, S. S. (2021). *Investigating the relationship between mRNA degradation rates and secondary structure in mycobacteria*. *March*.

Liu, F., Zheng, K., Chen, H. C., & Liu, Z. F. (2018). Capping-RACE: a simple, accurate, and sensitive 5' RACE method for use in prokaryotes. *Nucleic Acids Research*, *46*(21), e129. https://doi.org/10.1093/nar/gky739

Lodato, P. B., & Kaper, J. B. (2009). Post-transcriptional processing of the LEE4 operon in enterohaemorrhagic Escherichia coli. *Molecular Microbiology*, *71*(2), 273–290. https://doi.org/10.1111/j.1365-2958.2008.06530.x

Martini, M. C., Zhou, Y., Sun, H., & Shell, S. S. (2019). Defining the transcriptional and post-transcriptional landscapes of mycobacterium smegmatisin aerobic growth and hypoxia. *Frontiers in Microbiology*, *10*(MAR), 1–17. https://doi.org/10.3389/fmicb.2019.00591

Meinken, C., Blencke, H. M., Ludwig, H., & Stülke, J. (2003). Expression of the glycolytic gapA operon in Bacillus subtilis: Differential syntheses of proteins encoded by the operon. *Microbiology*, *149*(3), 751–761. https://doi.org/10.1099/mic.0.26078-0

Obana, N., Shirahama, Y., Abe, K., & Nakamura, K. (2010). Stabilization of Clostridium perfringens collagenase mRNA by VR-RNA-dependent cleavage in 5′ leader sequence. *Molecular Microbiology*, *77*(6), 1416–1428. https://doi.org/10.1111/j.1365-2958.2010.07258.x

Picard, F., Dressaire, C., Girbal, L., & Cocaign-Bousquet, M. (2009). Examination of post-transcriptional regulations in prokaryotes by integrative biology. *Comptes Rendus - Biologies*, *332*(11), 958–973. https://doi.org/10.1016/j.crvi.2009.09.005

Rock, J. M., Hopkins, F. F., Chavez, A., Diallo, M., Chase, M. R., Gerrick, E. R., Pritchard, J. R., Church, G. M., Rubin, E. J., Sassetti, C. M., Schnappinger, D., & Fortune, S. M. (2017). Programmable transcriptional repression in mycobacteria using an orthogonal CRISPR interference platform. *Nature Microbiology*, *2*(February), 1–9. https://doi.org/10.1038/nmicrobiol.2016.274

Rohde, K. H., Veiga, D. F. T., Caldwell, S., Balázsi, G., & Russell, D. G. (2012). Linking the Transcriptional Profiles and the Physiological States of Mycobacterium tuberculosis during an Extended Intracellular Infection. *PLoS Pathogens*, *8*(6), e1002769. https://doi.org/10.1371/journal.ppat.1002769

Sala, C., Forti, F., Magnoni, F., & Ghisotti, D. (2008). The katG mRNA of Mycobacterium tuberculosis and Mycobacterium smegmatis is processed at its 5′ end and is stabilized by both a polypurine sequence and translation initiation. *BMC Molecular Biology*, *9*, 1–12. https://doi.org/10.1186/1471-2199-9-33

Shell, S. S., Chase, M. R., Gray, T. A., Wade, J. T., Singh, N., Dejesus, M., Sacchettini, J. C., Ioerger, T. R., & Fortune, S. M. (n.d.). *mRNA cleavage shapes mycobacterial transcriptomes and stabilizes the esxB-esxA virulence factor transcript*.

Simeone, R., Bottai, D., & Brosch, R. (2009). ESX/type VII secretion systems and their role in host-pathogen interaction. *Current Opinion in Microbiology*, *12*(1), 4–10. https://doi.org/10.1016/j.mib.2008.11.003

Smith, J., Manoranjan, J., Pan, M., Bohsali, A., Xu, J., Liu, J., McDonald, K. L., Szyk, A., LaRonde-LeBlanc, N., & Gao, L. Y. (2008). Evidence for pore formation in host cell membranes by ESX-1-secreted ESAT-6 and its role in Mycobacterium marinum escape from the vacuole. *Infection and Immunity*, *76*(12), 5478–5487. https://doi.org/10.1128/IAI.00614-08

Van Assche, E., Van Puyvelde, S., Vanderleyden, J., & Steenackers, H. P. (2015). RNA-binding proteins involved in post-transcriptional regulation in bacteria. In *Frontiers in Microbiology* (Vol. 6, Issue MAR). https://doi.org/10.3389/fmicb.2015.00141

World Health Organization. (2020). *Global Tuberculosis Report 2020*.