INVESTIGATING THE ROLE OF SMALL RNAS IN TRANSCRIPTOME REGULATION OF GENETICALLY DIVERSE CLINICAL STRAINS OF *MYCOBACTERIUM TUBERCULOSIS*

Divenita Govender

Submitted in fulfilment of the requirements for the degree of Master of Science in the Discipline of Microbiology, School of Life Sciences, College of Agriculture, Engineering and Science, University of KwaZulu-Natal (Westville Campus)

Submission: 6 December 2021

Supervisor: Dr N. E. Mvubu

Co-Supervisor: Prof. M. Pillay

As the candidate's supervisor I agree to the submission of this dissertation.

Supervisor: Signed: _____ Name: _____ Date: _____

PREFACE

The research contained in this dissertation was completed by the candidate while based in the Discipline of Microbiology, School of Life Sciences of the College of Agriculture, Engineering and Science, University of KwaZulu-Natal, Westville Campus, South Africa from January 2019 - December 2021 under the supervision of Dr. N. E. Mvubu and Prof M. Pillay. The contents of this work have not been submitted in any form to another university and, except where the work of others is acknowledged in the text.

COLLEGE OF AGRICULTURE, ENGINEERING AND SCIENCE

DECLARATION 1 - PLAGIARISM

I, Divenita Govender, declare that:

1. The research reported in this thesis, except where otherwise indicated, is my original research.

2. This thesis has not been submitted for any degree or examination at any other university.

3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
   a. Their words have been re-written but the general information attributed to them has been referenced
   b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.

5. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

……

Signed; Divenita Govender

Date: 6 December 2021

COLLEGE OF AGRICULTURE, ENGINEERING AND SCIENCE

DECLARATION 2 - PUBLICATIONS

DETAILS OF CONTRIBUTION TO PUBLICATIONS that form part and/or include research presented in this thesis (include publications in preparation, submitted, *in press* and published and give details of the contributions of each author to the experimental work and writing of each publication)

Publication 1: Title: *Mycobacterium tuberculosis* complex transcriptome is regulated by lineage-specific small RNAs mutations and abundance - *International Journal of Molecular Sciences* (Revision invited by the journal) (Chapter 2 of dissertation).

Author contributions: Divenita Govender- Methodology, validation, software selection, formal analysis, investigation, data curation, original draft preparation and visualization; Dr N. E. Mvubu- Conceptualization, methodology, validation, investigation, review and editing, visualization, supervision, and funding acquisition; Prof M. Pillay- Conceptualization, validation, investigation, resources, review and editing, supervision and funding acquisition.

Signed: Divenita Govender

Date: 6 December 2021

ACKNOWLEDGEMENTS

I would like to acknowledge the following people for the successful completion of this dissertation, in no particular order:

God, for making this possible and giving me strength when I felt my worst.

My Father and Guardian Angels for always guiding and watching over me.

My family for their unwavering support, especially my mother, Glendelle Naidoo, grandparents, Kesava and Theresa Naidoo, Lynn and Tyger Naidoo and Tivon Reddy for being my motivation and strength. Your unconditional love and ever-willingness to help has always been appreciated.

My supervisor, Dr Nontobeko Mvubu for all her guidance, patience and inspiration throughout this degree. You have nurtured the researcher within and have made an incredible impact on my life that I will forever be grateful for.

My co-supervisor, Professor Manormoney Pillay for her expertise, assistance and advice navigating the scientific world.

Asanda Nyide and Kynesha Moopanar for always being available to help and push me in the right direction when times were difficult.

The staff and students from the Medical Microbiology Department, UKZN, Nelson R. Mandela School of Medicine and the Microbiology Department, UKZN, Westville Campus.

The National Research Foundation for funding this degree.

This project was part of the EDCTP2 programme supported by the European Union (Grant number: TMA2020CDF-3167

ABSTRACT

Tuberculosis (TB), caused by the human adapted members of the *Mycobacterium tuberculosis* complex (MTBC), is a threat to global health. Understanding the regulatory network of the MTBC members may reveal novel vaccine candidates and drug targets. The small RNAs (sRNAs) have only recently been investigated for their role in *Mycobacterium tuberculosis* (*M. tb*) transcriptome regulation with none being explored in clinical strains or within the MTBC lineages. The present study aimed to investigate the regulatory role of sRNAs on the *M. tb* transcriptome in a lineage-specific manner, with emphasis on the clinical strains most prevalent in South Africa. *In silico* whole genome sequence alignment of strains belonging to the eight MTBC lineages was performed to identify sRNAs containing lineage-specific mutations and their respective potential targets. To elucidate transcriptome regulation in clinical strains of *M. tb* belonging to the Beijing and F15/LAM4/KZN lineages, mRNA and sRNA sequencing were performed followed by Hisat-Ballgown Bioinformatics analysis to identify novel sRNAs and their respective targets. The sRNAs discovered from sRNA sequencing were confirmed through real time qPCR. The *in silico* data revealed several sRNAs that may play a role in transcriptome regulation at a lineage-specific level, such as those involved in macrophage entry, lipid biosynthesis pathway, adaptation mechanisms during antibiotic exposure, and environmental stress. They may also be able to disrupt genes that are detrimental and restore functions to those that are beneficial. The mutated and consensus sRNAs were identified to target the same function, but one pathway may be more efficient than the other. Novel sRNAs were discovered from sRNA sequencing of the Beijing and F15/LAM4/KZN clinical strains, with their predicted targets absent from the mRNA sequencing results, indicating these sRNAs may elicit an inhibitory function. Real time-PCR analysis revealed significant fold change differences between the clinical strains belonging to the Beijing, F15/LAM4/KZN, F11 and Unique families suggesting an underlying regulation of these transcripts at a family level. This data could explain the underlying phenotypic differences observed within the MTBC and understanding of the regulatory function of these sRNAs, may identify novel alternative strategies in the fight against *M. tb*.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

Tuberculosis (TB) is one of the top 10 causes of death across the world, with an estimated incidence of 9.9 million in 2020 (World Health Organization, 2021). Approximately 1.3 million TB deaths were reported among HIV-negative people and 214 000 among HIV-positive people in the same year (World Health Organization, 2021). An estimated 150 000 people started treatment for multi-drug resistant (MDR) and rifampicin- resistant tuberculosis (RR-TB) in 2020, a 15% decrease since 2019 (World Health Organization, 2021). The causative agent, *Mycobacterium tuberculosis* (*M. tb*), is spread via the respiratory system (Guirado *et al.*, 2013). The bacterium is phagocytosed/engulfed by innate cells lining the alveolus, and the microorganism is eventually disseminated to other organs (Guirado *et al.*, 2013).

Despite the use of 6 diagnostic tests recommended by the World Health Organization (WHO), 25 drugs and 14 vaccine candidates in clinical trials, the decrease in infection rates, mortality and the number of cases was largely due to the COVID-19 pandemic (WHO, 2021). A 1.3 million drop was noted in the number of newly diagnosed people, resulting from reduced access to TB treatment and testing facilities, which subsequently lead to an increase in TB related deaths (WHO, 2021). This has caused a reversal of progress that has been made for the eradication of TB (World Health Organization, 2021).

Both tuberculosis and COVID-19 are infectious diseases that primarily affect the lung, with both primary dispersal methods via aerosols (Jayaweera *et al.*, 2020). The lockdowns imposed to combat COVID-19, had serious repercussions to the battle against TB (Meo *et al.*, 2020). Due to the similarity of symptoms between both diseases, there may have been a delay in infected people suspecting TB and preferred to wait till the symptoms subsided. With both the TB stigma and the now added stigma regarding COVID-19, people were more reluctant to get tested especially in those of a lower socio-economic standard as the quarantine requirement would result in the inability for them to work and obtain basic needs, such as food. The lockdown also resulted in the confinement of families to close quarters, increasing the spread of TB. Non-essential services and limited private sector healthcare also decreased the ability of those who wanted to get tested from being tested (Kant & Tyagi, 2021). Those that have been diagnosed suffered the consequences as departments dedicated to TB were non-functional and testing laboratories were dedicated to processing COVID-19 samples. Follow-up treatment requiring both sputum microscopy and culturing of the patient samples was not possible during the lockdown period resulting in deterioration of those suffering from drug resistant TB, relapse and treatment failure. Counselling required for those undergoing treatment to combat the stigma attached to the disease were halted.

Medication received via direct observation had stopped, causing premature discontinuation resulting in the development of resistance in patients (Kant & Tyagi, 2021).

The pandemic had also helped in decrease the spread of TB through the implementation of physical distancing and mask wearing. There has also been a general increase in the air quality due to fewer vehicles on the road and the closure of factories, indirectly assisting those affected with respiratory ailments. The health sector has had a tremendous upgrade in equipment which will assist in patient care for various illnesses other than COVID-19. Protocols implemented to reduce aerosol production can also be used for TB wards and hospitals. Treatment plans have been adapted to allow for treatment to be given for a longer period of time to reduce the number of visits. Online, telephonic and appointment -based protocols have been implemented to assist with non-emergency situations and conditions reducing exposure to overcrowded healthcare facilities. Following COVID-19, a spotlight has been placed on the healthcare sector, forcing governments to allocate more funds to this department, increasing the standard of health globally (Kant & Tyagi, 2021).

The only vaccine available against TB is the Bacillus Calmette-Guerin (BCG) vaccine, which has been in use for decades and only provides partial protection in adolescents and adults (de Gijsel & von Reyn, 2019). Current treatment regimens last between 6-9 months and have severe side-effects. These factors result in treatment reluctancy, development of extensively drug-resistant TB (XDR-TB) and RR-TB cases as well as reduced protection against the bacterium (World Health Organization, 2021). There is a move toward discovering novel methods to target and eradicate TB, in particular, multi-drug resistant TB (Miotto *et al.*, 2012b).

One of these methods includes targeting the mechanisms involved in the regulatory network of the cells, and an excellent candidate for this are small RNAs (sRNAs) (Miotto *et al.*, 2012b). Previously, gene expression and transcript control have been explored at the DNA level, however, in the last 20 years, studies have revealed a more layered regulation, particularly in bacteria (Haning *et al.*, 2014b) These small transcripts have been identified in several bacteria and are considered important regulators which act on targets that are independently expressed (Storz *et al.*, 2011). Many act via base-pairing to the complementary targets, however, there are a few that perform regulatory functions by mimicking the secondary structures of other nucleic acids (Gottesman & Storz, 2011).

The investigation of sRNAs in *M. tb* has only become successful in recent years (Ostrik *et al.*, 2021). Changes in *sRNAs* have been largely investigated in the laboratory H37Rv strain (Liu *et al*., 2016), with over 200 sRNAs being discovered in *Mycobacteria* (Haning *et al.*, 2014b). There have been no reports, however, on the role of these non-coding RNA species in clinical

strains of *M. tb*. The differences in transcript expression of these sRNAs, particularly under environmental stress, suggests that these transcripts may be relevant to the pathogenesis of *Mycobacteria* (Haning *et al.*, 2014b). The study proposes to elucidate the role of sRNAs on transcriptome changes during *in vitro* growth of genetically diverse clinical strains of *M. tb*. Understanding the role of these sRNAs might offer novel insights into crucial gene regulatory networks that are exhibited by these strains, and might ultimately contribute to strain virulence. Novel sRNAs identified to play an important role in TB pathogenesis may be suitable targets for the design of more effective intervention strategies.

1.2 Epidemiology

Among the 5.4 million infected with TB in 2020, 56% were adult males, 33% were adult females, 11% were children, under the age of 15 and 6.9% were diagnosed with HIV. Two thirds belong to eight countries: India (26%), China (8.5%), Indonesia (8.4%), the Philippines (6%), Pakistan (5.8%), Nigeria (4.6%), Bangladesh (3.6%) and South Africa (3.3%). These eight countries, along with 22 other countries, form part of the 30 high TB burden countries by WHO, contributing to 86% of all cases, with South Africa being the 8th highest. In 2019, it was estimated that there were 5.8 million new cases, a 1.3 million decrease from 2019. There have also been an estimated 1.4 million TB deaths and of those, 209 000 deaths were those infected with HIV in the same year. In South Africa, the total TB incidence is 328 000 with a TB mortality of 63 000 in 2019 and a XDR and RR-TB contribution of 6 800 cases (World Health Organization, 2021).

The incidence rate of TB has been decreasing by 1.6% per year between 2000 and 2018, with a cumulative decrease of 44% between 2015–2020. In the same time period, the number of TB deaths decreased by 4.9% globally. The drastic reductions are due to TB prevention and care programmes, however, complete eradication of TB remains a global challenge (World Health Organization, 2021).

Almost all TB infections are caused by the inhalation of droplet nuclei (Fitzgerald *et al.*, 2015) and can result in two medically characterized states: active disease or latent infection (Cadena *et al.*, 2017). The latent infection is defined as the presence of immunological sensitivity to the mycobacterial antigen, without the development of the clinical symptoms such as cough, fever and loss of weight, which is responsible for 90% of human infections (Cadena *et al.*, 2017). The active disease is identified as the presence of clinical symptoms as well as the microbiological evidence of the causative agent of TB, *M. tb* (Cadena *et al.*, 2017). This evidence can be obtained from a sputum sample, where an estimated 10 000 organisms/mL in the sputum smear indicate positivity, by nucleic acid testing, acid-fast staining or fluorescent light-emitting diode microscopy, which will allow for identification of the infecting bacilli (Fitzgerald *et al.*, 2015;

Cadena *et al.*, 2017; World Health Organization, 2021). Among the TB infected individuals, many have presented with the latent form, with only a few developing the progressive lung disease, which is responsible for its transmission (Arnvig & Young, 2012).

1.3 Characteristics of *M. tuberculosis*

Several pathogenic traits are distinctive for *M. tb* and aid the bacterium to be the successful pathogen that we know today. This slow-grower has a characteristic cell wall structure providing the microorganism with a protective barrier against antibiotics and environmental stress, and to a certain degree, contribute to the pathogen's virulence (Daffe & Draper, 1997; Cole *et al.*, 1998). The cell wall comprises of a large number of glycolipids, which interferes with the host defence mechanisms allowing survival within phagosomes. The uniqueness of this bacterium stems from the high lipid and mycolic acid content within the cell wall (Fratti *et al.*, 2003; Alderwick *et al.*, 2007). The antigen 85 complex is an exported protein that is characteristic of *M. tb*. It is a prominent virulence factor and a potent and protective antigen for the bacterium (Belisle *et al.*, 1997). The other most important virulence factors of *M. tb* are the five type 7 protein secretion systems (ESX1-5) of. The ability of *M. tb* to persist within macrophages contributes to its success as a pathogen. ESX1 assists with this by translocating the bacterium from the phagosome of the cytosol of the macrophages (van der Wel *et al.*, 2007; Romagnoli *et al.*, 2012). The ESX3 secretion system is involved in iron and zinc acquisition, while ESX5 is involved in the immune system interaction within mammals (Serafini *et al.*, 2009; Houben *et al.*, 2012). ESX2 and ESX4 are involved in survivability and conjugation, respectively (Roy *et al.*, 2020). The dormancy survival regulon (Dos) controls over 50 genes responsible for the hypoxic response that is experienced within the macrophages and granulomas (Voskuil *et al.*, 2004; Kumar *et al.*, 2007). The dormant state is then activated during those environments causing the metabolism of the bacterium to be down-regulated, multiplication to cease, anaerobic metabolism to be initiated and the production of stress proteins to commence (Korch *et al.*, 2009). The bacterium can restore its active state by resuscitating promoting factors (rpf), triggering a cascade of events but can persist in this dormant state for a very long time (Hett *et al.*, 2010; Mukamolova *et al.*, 2010).

1.4 Evolution and lineages

The *Mycobacterium* genus compromises of 170 species, which are divided into two categories, namely fast-growers, those that form colonies in less than 7 days and slow-growers

that take more than 7 days to form colonies. An example of a fast-grower is *Mycobacterium abscessus*, a so-called non-tuberculous *Mycobacterium* causing the disease in immune-compromised individuals. The other bacteria falling into this category include the slow-growing *Mycobacterium avium*, *Mycobacterium marinum*, *Mycobacterium xenopi*, *Mycobacterium gordonae* and *Mycobacterium kansasii*. Other slow-growers include human mycobacterial pathogens such as the human-adapted *M. tb* complex (MTBC), *Mycobacterium leprae* and *Mycobacterium ulcerans*. The human-adapted MTBC are the only human pathogens that have no environment or animal reservoir and can only be passed on from person to person whereby their virulence is directly linked to transmission (Gagneux, 2018). The MTBC comprises of *M. tb, M. africanum, M. canetti, M. bovis, M. caprae, M. pinnipedii, M. microti, M. suricattae, M. mungi, M. dassie and M. oryx* (Velayati & Farnia, 2017).

Due to the limited amount of genetic variation found within *M. tb* strains as compared to other bacteria, methods such as multilocus sequence typing, which have been used to determine sequence variation, generally in genetically diverse microorganisms, have proved uninformative in the case of this pathogen (Gagneux, 2018). However, through whole genome sequencing of strains found globally, advances have been made in uncovering more genetic diversity (Borrell *et al.*, 2019). The MTBC consists of seven phylogenetic lineages, differing in their distribution throughout the globe. Strains within these lineages can differ by ~2000 single nucleotide polymorphisms (SNPs)(Borrell *et al.*, 2019).

It has been hypothesised that the ancestor of *M. tb* was *Mycobacterium prototuberculosis* (Demay *et al.*, 2012). Several genetic changes had to occur in order for the members of the human-adapted MTBC to be the professional pathogens identified to date (Gagneux, 2018). When compared to the distant relatives, *M. marinum* and *M. kansasii*, there was a significant decrease in the genome size, allowing for the bacterium to adopt a pathogenic lifestyle as well as gaining new genes via horizontal gene transfer (Veyrier *et al.*, 2011). The *M. tb* genome is 4.4 Mb while the genome size of *M. marinum* and *M. kansasii* are 6.6 Mb and 6.4 Mb, respectively (Stinear *et al.*, 2008; Wang *et al.*, 2015). Some of the genes acquired include those encoding transferases and those associated with adaptation to anaerobic environments (Becq *et al.*, 2007; Veyrier *et al.*, 2009; Boritsch *et al.*, 2014; Reva *et al.*, 2015). The virulence factors such as the PhoPR two-component system (transcription factors), the dormancy survival regulator system (DosR/S/T) regulon (involved in the latent TB infection), *mce*-associated genes (macrophage entry) and components of the ESAT6 (6 kDa early secretory antigenic target) secretion system (a protein secretion system involved in virulence) were identified in members of the MTBC (Wang *et al.*, 2015). Moreover, the protein families of proline-glutamic acid (PE) and proline-proline glutamic acid are found in larger numbers in pathogenic *Mycobacteria* as compared to the non-pathogenic

Mycobacteria. The protein subfamily PE_polymorphic guanine-cytosine-rich sequence (PE_PGRS) proteins are found exclusively in pathogenic *Mycobacteria* (Gey van Pittius *et al.*, 2006). A disproportionately large number of toxin-antitoxin system genes have been observed in *M. tb*, and some were even found to be differentially expressed in clinical strains. It can be stated that no one variation in terms of genetic makeup causes the virulence and transmissibility that is observed in the MTBC, but rather a wide array of epistatic interactions between genes and transcriptional activities (Sala *et al.*, 2014; Wang *et al.*, 2015; Gagneux, 2018).

The *M. tb*, has co-evolved and adapted to its human host over the centuries, with many strain families and lineages. These lineages have been associated with specific virulence factors and different forms of the disease (Reed *et al.*, 2009). However, the recent strains of *M. tb* follow a clonal characteristic, which is due to unidirectional evolution where divergence is observed through intragenomic variation and horizontal gene transfer (Demay *et al.*, 2012). The mycobacterial lineage has been subdivided in the phylogenetic tree with both humans and other mammals as hosts (Figure 1.1a). The human host mycobacteria include *M. tb sensu stricto* and *Mycobacterium africanum* and can be subdivided into seven lineages (L): L1 – L7. Lineage 5 (L5) and Lineage 6 (L6) are known as *M. africanum* (AFRI) West African 1 and 2, respectively(Gagneux, 2018). Lineage 1 (L1) is also known as the East-African-Indian (EA1) lineage (Conceicao *et al.*, 2019). Lineage 2 (L2) is known as the East- Asian lineage and includes the Beijing family strains while the L3 lineage is also known as Central Asian (CAS)/Delhi lineage (Shuaib *et al.*, 2020). Lineage 4 (L4), also known as the Euro-American lineages comprises of the Cameroon, Ghana, Haarlem, Latin American_Mediterranean (LAM), X and S type sublineages and families (Yimer *et al.*, 2015; Thain *et al.*, 2019). The final lineage 7 (L7) is predominantly restricted to Ethiopia and Ethiopian immigrants (Senghore *et al.*, 2020). The evolutionary history of the MTBC still remains unclear as not all mutations for each lineage follow the molecular clock hypothesis of the set accumulation of genetic diversity over time and the limited knowledge on the effects of the latency stage on this evolution. The geographic location of the lineages differs greatly with some lineages being found globally and others more geographically restricted (Figure 1.1b). It is hypothesized that the restricted lineages are adapted to their local human populations. This local adaptation is the ability of the microorganism to adapt to a single host species without the ability to infect other host species. This is expected as the MTBC is an obligate pathogen that can manipulate the human immune system, promoting its replication and transmissibility (Gagneux, 2018). The TbD1 (*M. tb* specific deletion 1) deletion can classify these lineages into two clades, namely the “modern” and “ancestral” clade. The “modern”, TbD1- deficient clade is composed of the L2, L3 and L4 lineages which are more

recent strains. The remaining “ancestral” strains with an intact TbD1 region form a paraphyletic group (Coscolla & Gagneux, 2014; Orgeur & Brosch, 2018).

Phenotypic variation has been identified among the clinical strains in terms of transcriptomic profiles, methylation profiles, drug susceptibilities, protein and metabolite levels and cell wall structure(Borrell *et al.*, 2019). Most importantly, variation has been observed in the infectivity of the bacterium to human host. “Modern” lineages have been found to have a faster disease progression as compared to the “ancestral” lineages as stated by Borrell *et al.* (2019).

These strains can be further classified using various molecular methods, with one based on *katG (*Catalase-peroxidase-peroxynitritase T)-*gyrA* (DNA gyrase (subunit A) SNPs, having separated the *M. tb* isolates into three principal genetic groupings (PGG). PGG1 being considered as the oldest and PGG3 as the youngest which evolved from PGG2 (Demay *et al.*, 2012). Through other characterization methods, such as the structure of the RD locus and spoligotyping, the PGG groups were able to be subdivided into PGG1 and PGG2/3. PGG1 includes the ancestral (EAI) and modern lineages (CAS lineage and Beijing family) while PGG2/3 includes only the modern sublineages (Haarlem, LAM, T, and X). SNPs and large sequence polymorphisms have also allowed for the grouping of “Euro American” which encompasses PGG2/3 sublineages including Haarlem, LAM, T and X (Demay *et al.*, 2012).

From an ecological point of view, the lineages can be subdivided into specialists and generalists (Gagneux, 2018). Specialist have a small niche targeting a specific human population and generalists have a wide host range, targeting many populations such as L4 (Figure 1.1b). Recent whole genome studies have found that L4 can be further divided into sublineages (Figure 1.2a). These sublineages can also be divided into generalists and specialists, the generalists being Haarlem, LAM and PGG3 and the specialists being L4.5, Uganda, Cameroon. A genotyping study completed by Stucki *et al.* (2016) exhibited the distribution and sublineage groupings of 3366 L4 clinical isolates from 100 countries (Figure 1.2b).

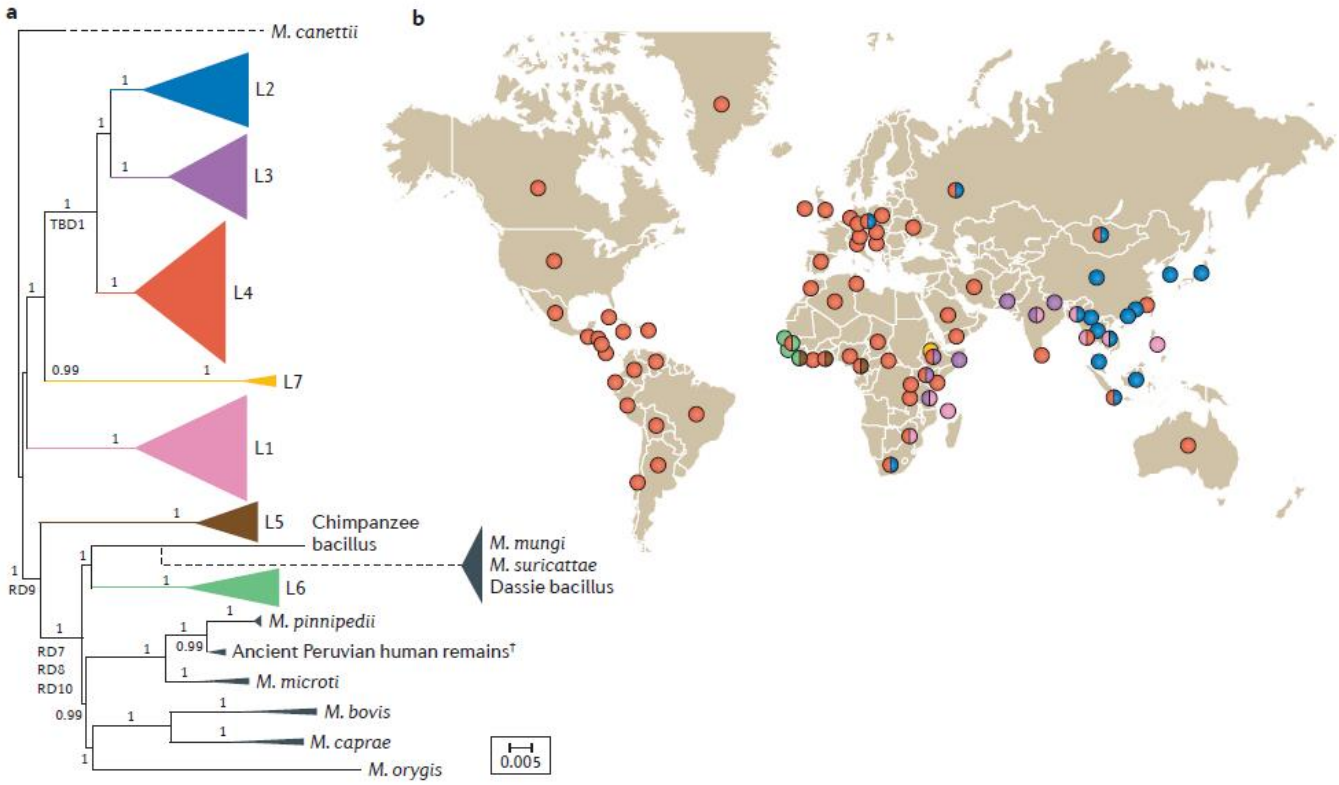


Figure 1.1: Global phylogeography of the members of the MTBC. a) Phylogenetic tree of the MTBC exhibiting the different human and animal host lineages, in colour, and species that cause TB in other mammals, in grey. b) The geographic distribution of the various lineages globally with the colour of the lineages corresponding to the colours displayed on the map (Gagneux, 2018).

In the study conducted by Demay *et al.* (2012), it was observed that the strain families and sublineages LAM, Beijing, T, X and Haarlem were found in North America, in similar proportions (~20%) with CAS (Central Asian spoligotype) and AFRI lineages as well as the S strain family found below 4% of the studies isolates. In the Caribbean, Haarlem, LAM and T families were found at similar proportions (~27%) and the X family at 8.2%. LAM (49.3%), T (26.7%) and Haarlem (15.7%) families were found in South America. Europe was divided into Northern, Southern, Western and Eastern Europe with the T strain family being represented in Northern, Southern and Western Europe at 35%. The Haarlem sublineage represented at 24% in Eastern Europe, EAI lineage (24.9%) in Northern Europe and S strain family (5.8%) in Southern Europe (Demay *et al.*, 2012).

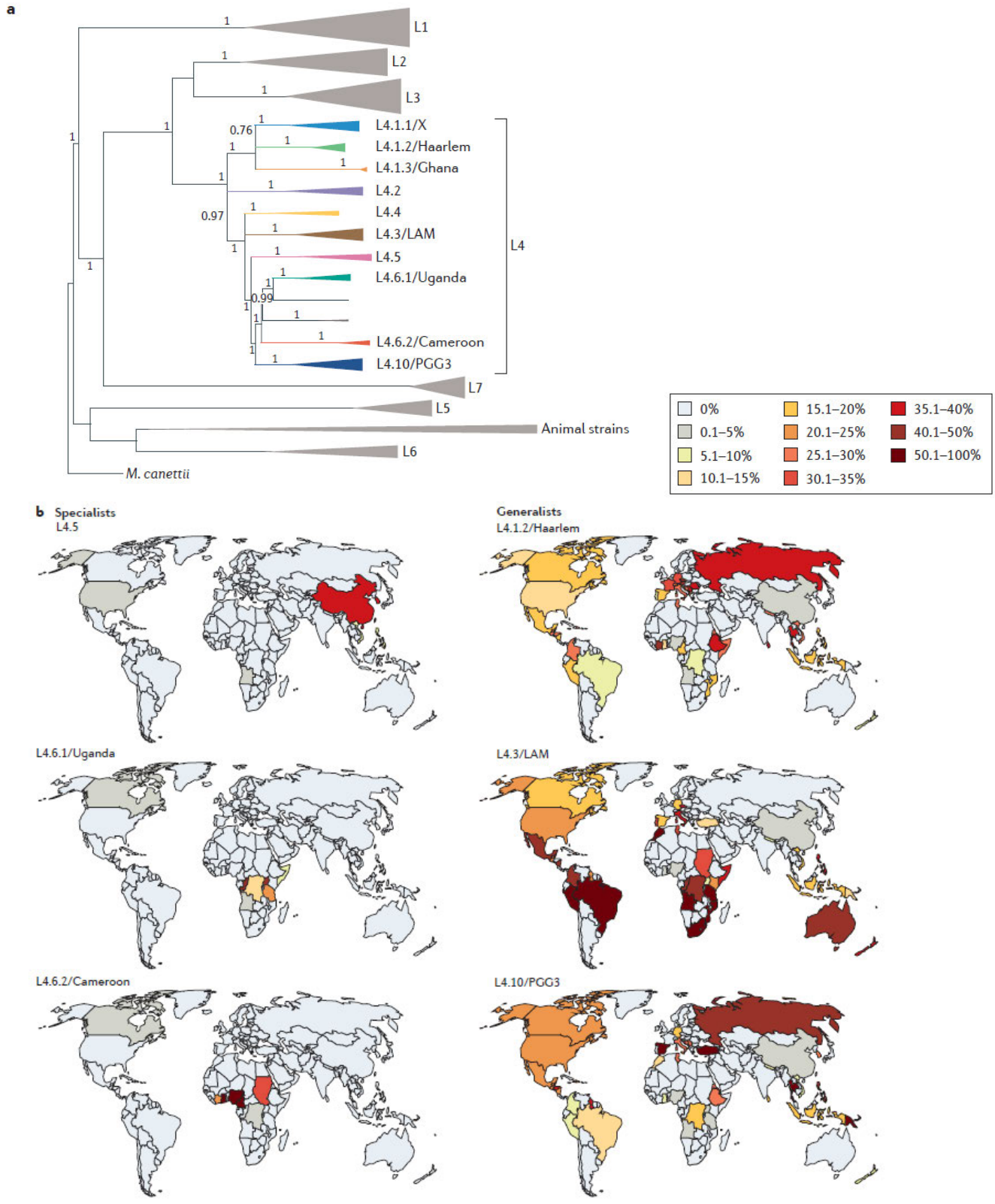


Figure 1.2: Further phylogenetic subdivision of the L4 lineage of the *M. tb* and its global distribution. a) The L4 subdivisions are presented in colour with their common names. b) The global distribution of the L4 subdivisions that are classified as generalist and specialist. The

specialists exhibit a more localized distribution while the generalists are found globally (Gagneux, 2018).

Africa was also divided into sub-regions, namely, Northern, Southern, Eastern, Western and Middle Africa. LAM was the most predominant sublineage in the Northern, Southern and Eastern sub-regions with the West African 1 (L5) lineage being the most predominant in Western (37%) and Middle Africa (7.3%). The T strain family was identified in all regions at ~20%, the Beijing family was identified in Northern, Southern and Eastern Africa, with the highest percentage calculated in Southern Africa at 21.5%. The CAS lineage isolates were observed in Eastern (11.8%) and Northern Africa (7.2%), while the S strain family was identified in Northern (8%) and Southern (5.8%) Africa, with the X strain family being present at 17.2% in Southern Africa and the EAI lineage in Northern (6.8%) and Eastern Africa (8.2%). The Beijing family was found in 51.20% of the Northern Asia isolates, with Central, Eastern and South-Eastern Asia representing ~50% of the isolates from the Beijing family. Isolates collected from South-Eastern Asia (37.6%) and Southern Asia (33%) belong to the EAI lineage, with the CAS lineage representing 30.1% in Southern Asia as well. Western Asia mostly comprised the CAS lineage at 9.6% and T strain family at 35.4%. T and LAM families contributed to ~20% in Australasia (Demay *et al.*, 2012).

1.5 ***M. tuberculosis*** **infection**

M. tb is spread via the respiratory system (Guirado *et al.*, 2013). The first stage occurs when a few tubercle bacilli in an active TB infected patient reach the alveoli. Non-phagocytic cells present in the alveolar space, such as M (microfold) cells, alveolar endothelial cells as well as type 1 and 2 epithelial cells (pneumocytes) can also be infected by *M. tb*. The bacilli are then phagocytosed by macrophages and dendritic cells. *M. tb* then replicates intracellularly and these bacteria-loaded cells eventually disseminate to other organs (Ahmad, 2011; Delogu *et al.*, 2013; Guirado *et al.*, 2013). The infection is dependent on the genetics of the infected patient as well as the strain of *M. tb* that is inhaled. This initial infection causes mild inflammation in the lungs. Generally, the macrophages destroy the bacteria as soon as they enter due to the innate immune system, however, some bacteria are able to escape this first line of defence (Delogu *et al.*, 2013).

The surviving bacteria replicate in the macrophages and can spread to the epithelial and endothelial cells as well as to other organs through the lymphatic system and by haematogenous dissemination to infect other cells. When the adaptive immune system is activated, neutrophils, lymphocytes and other immune cells are directed to the primary infection site, which later forms the granuloma (Delogu *et al.*, 2013). At this point there is limited tissue damage. After 6-8 weeks post-infection, antigen-presenting dendritic cells travel to the lymph nodes, recruiting T

lymphocytes to the infection site forming an early stage granuloma where the macrophages are activated to destroy intracellular *M. tb*. The development of the granuloma marks the persistent or latent stage of the infection (Sasindran & Torrelles, 2011). The granuloma is then calcified using fibrotic components, which cover the surface and the bacilli are then encased in this structure, protected from the immune response of the host. This structure is then called the Ghon complex. This complex was thought to be the housing centre of *M. tb* during the latent infection with the bacteria, maintaining a metabolically inactive, dormant state where growth and the spread of the bacterium is limited. Approximately 90% of the people who have been infected are asymptomatic, however, Mycobacteria can persist in the macrophages (Sasindran & Torrelles, 2011; Delogu *et al.*, 2013). It was also hypothesized that the *M. tb* infection was reactivated from this source, however, studies completed by Hernandez-Pando *et al.* (2000) and Neyrolles *et al.* (2006) reported that latent infections can be reactivated in tissues and cells that are not associated with the Ghon complex, granuloma or the primary infection site.

The final stage includes the reactivation of the disease. Due to the dynamic relationship between the host and *M. tb* and changing environments, it has been proposed that the majority of the bacteria are found in the dormant state, with a few being in the active, replicating state and these bacteria are known as "scouts". When the host immune system is active, these scouts are destroyed by the defences and induce memory T cells to target antigens on the surface of the bacilli. When these bacteria are killed, the dormant bacteria replenish these sources by becoming active. The problem then arises when the host immune system is compromised, which may be due to genetic or environmental causes, resulting in the uncontrolled growth of the scouts and the disease becoming active and persistent (Delogu *et al.*, 2013). Another possibility for the reactivation of the disease is the failure to maintain immune signals, resulting in the eruption of the granuloma, which progresses to a disease (Sasindran & Torrelles, 2011).

1.6 TB treatment and vaccines

There are various types of TB such as MDR-TB, XDR-TB and totally drug-resistant (TDR-TB). MDR-TB is characterized by the resistance of the bacilli to isoniazid (INH) and rifampicin (RMP) (Velayati *et al.*, 2013). The XDR-TB exhibits resistance to INH, RMP, a fluoroquinolone and a second-line injectable drug and TDR-TB exhibits resistance to all first- and second- line drugs (Velayati *et al.*, 2013). The TB of most concern is TDR-TB as there are currently no drugs to treat this type, therefore the need for novel ways to target and eradicate this disease is of the utmost importance (Velayati *et al.*, 2013; Fitzgerald *et al.*, 2015). Despite the wide variety of drugs that are available to cure TB, many resistant strains have emerged, especially over recent years. There is a move toward discovering novel methods to target and

eradicate TB, in particular, MDR-TB (Miotto *et al.*, 2012b). The first line of drugs used includes INH and RMP. INH is a base-line drug included in all TB treatments unless the bacilli are resistant to the drug. It can be used when *M. tb* is resistant to it at low-levels but is effective at higher concentrations and if the bacilli are resistant to other anti-tuberculosis (anti-TB) drugs. The side effects of this drug include hepatitis and liver injury. RMP is the second most widely used drug and can cause hepatitis at a lower risk than INH. Another concern is the drug-drug interactions that may occur if the patient is on other types of treatment, resulting in an increase in the dosage or a change in the anti-TB treatment. The alternative to first line drugs include rifapentine, pyrazinamide, ethambutol, streptomycin, fluoroquinolones and bedaquiline, each with various side effects and consequences such as a short half-life, high hepatotoxicity rates, rash, ocular toxicity and negative reactions with HIV drugs (World Health Organization, 2019).

If the first line treatment fails, there are second- and third-line agents, both increasing in toxicity with decreasing effectivity. The second-line agents include, ethionamide, cycloserine, terizidone, amikacin, kanamycin, capreomycin, thiacetazone and para-aminosalicylic acid (PAS). Third-line agents are only prescribed to patients with XDR-TB and have not been investigated thoroughly as a treatment for TB. Agents such as clarithromycin, amoxicillin-clavulanate, clofazimine and linezolid are considered as third-line agents (World Health Organization, 2019).

Several drugs have specific bacterial cell targets, such as inhibition of the β-subunit of RNA polymerase by rifampin. Rifapentine binds to and inhibits the β-subunit of the DNA-dependent RNA polymerase while fluoroquinolones target the bacterial DNA gyrase preventing essential supercoiling required for chromosomal replication. The ATP synthase enzyme is directly inhibited by TMC207, a diaryquinoline, and nitroimidazopyran, PA-824, when activated, inhibits protein and cell wall lipid synthesis. Another nitroimidazopyran, OPC-67683 inhibits mycolic acid biosynthesis and the diamine, SQ109, prevents cell wall synthesis (van den Boogaard *et al.*, 2009).

Several agents are being developed and are under investigation to be prescribed for MDR-TB. Three drugs namely, bedaquiline, delamanid and pretomanid have received regulatory approval and seven drugs have been repurposed to treat TB and are undergoing further testing. These include clofazimine, levofloxacin, linezolid, moxifloxacin, nitazoxanide, rifapentine and a higher dose of rifampicin (World Health Organization, 2019).

The current TB treatments are faced with many issues such as nonadherence due to the complexity and length of the treatment resulting in failure, relapse and resistance of the bacilli. The adverse effects of the treatment on the patient contributes to the treatment regime not being followed. Drugs are being developed to overcome these obstacles to allow for the treatment

duration to be shorter, with fewer side-effects. Changes in the regime components as well as the duration are also being investigated with several moving to phase II of clinical trials. However, this process will take time as the drugs would need to be monitored over several years (Goletti *et al.*, 2018).

A preventative approach to the TB infection is the immunisation of the population with a vaccine. However, the Bacille Calmette-Guérin (BCG) vaccine used currently was developed in the 1920s and is not as effective in protecting adults from all forms of TB, however does provide protection to children that may be exposed to severe forms of TB (World Health Organization, 2019). Three vaccines that are in phase III of trials are MIP/Immuvac, Vaccae and VPM1002. Vaccae, being part of the largest TB vaccine trial in the past 10 years, uses a lysate composed of the heat-killed environmental *M. vaccae*, aiding in decreasing the TB treatment time for patients with drug-susceptible TB. The safety and immunogenicity of the VPM1002 vaccine was being assessed in a phase II trial in South Africa with HIV-exposed and unexposed newborns (World Health Organization, 2019). MIP/Immuvac is a vaccine used to treat patients with multibacillary leprosy and to prevent transmission of leprosy to individuals in close proximity to the infected patients. The phase III trial tests the effectiveness and safety of the vaccine in preventing pulmonary TB in TB-positive cases (World Health Organization, 2019).

1.7 Potential approaches for drug development

Studying large regulatory networks and identifying genes and biomolecules, which form part of the *M. tb* networks, might identify potential targets for the treatment of TB. Processes such as lipid metabolism, iron uptake, oxidative stress resistance, cell wall synthesis, inhibition of apoptosis and protein secretion are being dissected in order to profile virulence mechanisms specific to *M. tb* using transcriptomic approaches. The complete genome of *M. tb* and global gene microarrays can be used in conjunction to identify genomic expression profiles at varying conditions, which would reveal novel transcripts which can be targeted for anti-TB drug development (Mukhopadhyay *et al.*, 2012).

Using transcriptomics, the *desA3* gene, which codes for the linoleolyl-CoA desaturase, was found to be upregulated in the human lung granuloma, resulting in the subsequent identification of the anti-TB thiourea drug isoxyl (Phetsuksiri *et al.*, 2003; Rachman *et al.*, 2006). The *kasA* and *kasB* genes, which code for β-keto-acyl carrier protein synthase were targets for thiolactamycin and platensimycin, both anti-TB treatments showing promising results with the potential to develop novel synthetic analogs (Brown *et al.*, 2009). The upregulation of the *mma-4* gene product in lung granulomas wase found to be a target for thiacetazone analogs such as SRI-224 (Alahari *et al.*, 2007). These are only a few of the many drugs and drug targets which

have been developed as anti-TB treatment due to the use of the transcript expression profiles and the transcriptome data generated by *M. tb* studies (Mukhopadhyay *et al*., 2012). However, not much research has been directed to understanding the molecular regulatory networks that control *M. tb* growth and virulence.

A potential approach to this is targeting the mechanisms involved in the regulatory network of the cells, and sRNAs present ideal target candidates (Miotto *et al.*, 2012b). Previously, gene expression and transcript control have been explored at the DNA level, however, in the last 20 years, studies have revealed a more layered regulation, particularly in bacteria (Haning *et al.*, 2014b) These small transcripts have been identified in several bacteria and are considered important regulators which act on targets that are independently expressed (Storz *et al*., 2011). Many act via base-pairing to the complementary targets, however, there are a few that perform regulatory functions by mimicking the secondary structures of other nucleic acids (Gottesman & Storz, 2011).

1.8 Small RNAs

The sRNAs and micro RNAs (miRNAs) have been found to control transcription, translation and gene silencing by changing RNA conformations, protein and RNA binding and DNA interactions (Miotto *et al.*, 2012b). The bacterial sRNAs are differentiated from eukaryotic miRNAs as the position of the sRNA binding with its target frequently occurs on the 5' end of the target whereas, miRNAs frequently pair on the 3' untranslated region of the target mRNA (Bloch *et al.*, 2017).These are small transcripts that range between 50 and 500 bases (Gottesman & Storz, 2011). sRNAs that have been involved in base-paring, are divided into two groups: cis-encoded sRNAs and trans-encoded sRNAs (Arnvig & Young, 2009). Cis-encoded RNAs are transcribed in the antisense direction to the targeted mRNA or opposite open reading frames, while trans-encoded sRNAs are transcribed from the intergenic regions of the DNA, which generally have several targets (Figure 1.3) (Arnvig & Young, 2009).

Since only a few sRNAs have their physiological roles defined, these can be used as models for what cis-encoded antisense sRNAs could do. The most prevalent role is the repression of toxic protein coding genes, which was first identified for plasmid-encoded sRNAs (Gottesman & Storz, 2011). These sRNAs also have the ability to affect translation efficiency as well as mRNA stability by binding to one or more mRNAs (Figure 1.4) (Arnvig & Young, 2012). However, in the case of 6S RNA, which binds directly to RNA polymerase, attachment results in the downregulation of the polymerase itself, inhibiting transcription (Wassarman, 2007b). These sRNAs can also upregulate translation by changing the conformation of the sRNA therefore

preventing an inhibitory secondary structure. This change can allow for the ribosomal binding site to become accessible (Figure 1.4)(Gottesman & Storz, 2011).

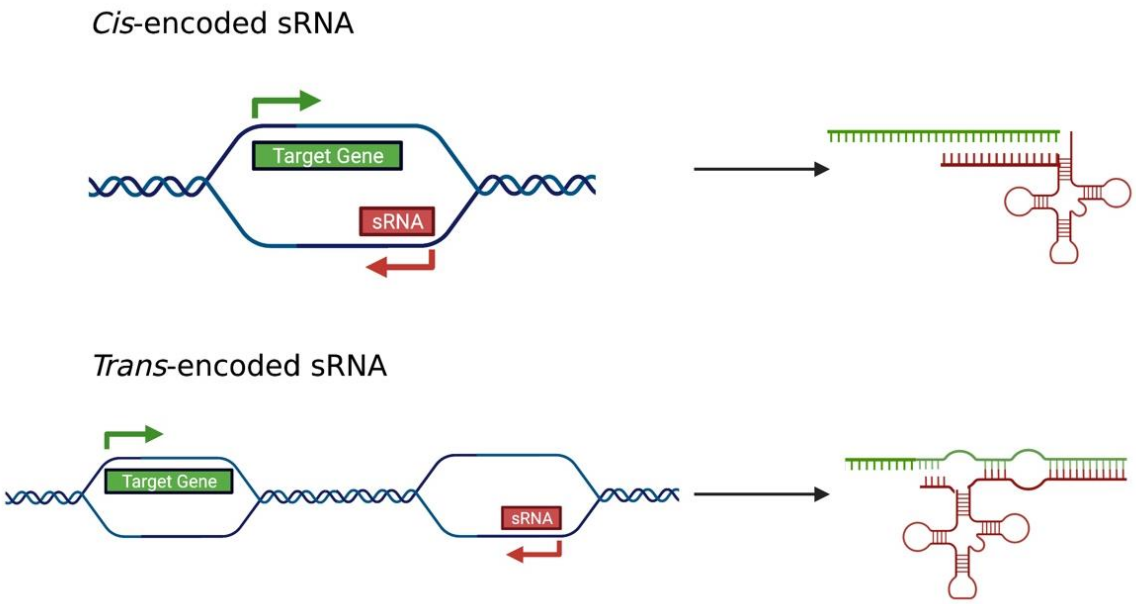


Figure 1.3: Genetic location and mode of action of cis- and trans-encoded sRNAs. Cis-encoded sRNAs have a high degree of complementarity while trans-encoded sRNAs have limited complementarity due to their multiple targets (Created with BioRender.com).

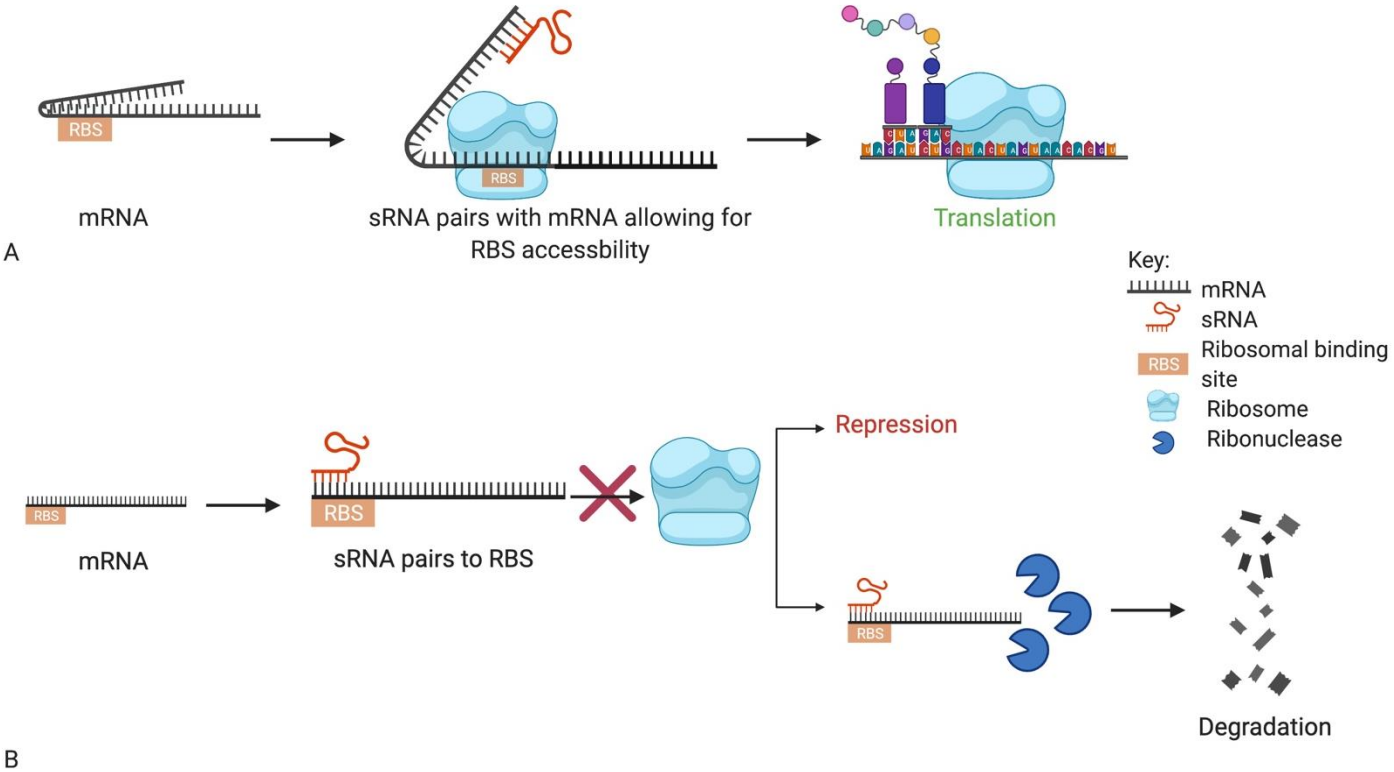


Figure 1.4: sRNA mode of action. A) Activation of translation by action of sRNA. Stem loop structure formation of an mRNA can cause the ribosomal binding site to be inaccessible . After

binding to a specific sRNA, the mRNA changes conformation, becoming more accessible, allowing for translation to proceed, and upregulating gene expression. B) sRNA inhibiting translation. Due to the sRNA binding to a portion of the mRNA target that contains the ribosomal binding site (RBS), the ribosome cannot attach to RBS, preventing translation of the mRNA, resulting in a decrease in gene expression and can result in mRNA degradation (Created with BioRender.com).

Initially, sRNAs were screened using computational searches, looking for intergenic regions with conserved sequences or orphan promoter and terminator sequences (Gottesman & Storz, 2011). Over time, the exploration methods improved to whole genome expression profiling, however the exact number of sRNAs identified to date are not confirmed. This is due to sRNAs only being detected *in silico* and have not exhibited a function. Many of these functions may be missed due to the nature of their expression, as they may only be expressed under specific conditions, making them difficult to detect using standard methods (Gottesman & Storz, 2011). Since they are processed from mRNAs, it may be difficult to distinguish from the 5' and 3' untranslated regions. By improving the detection methods for sRNAs, there will undoubtably be an increase in the confirmed numbers of sRNAs and their respective targets (Gottesman & Storz, 2011).

Even though the search for sRNAs in *M. tb* is still emerging, there are many sRNAs which have been identified for other pathogens (Arnvig & Young, 2012). These can be activated during stress conditions or are required in the general functioning of the cell (Arnvig & Young, 2012). RNA polymerase sigma S (rpoS) is an alternate sigma factor of RNA polymerase which is responsible for promoter recognition and transcription of mRNA and is controlled by the ARCZ sRNA in both *E. coli* and *Salmonella* (T. Dong & Schellhorn, 2010; Mandin & Gottesman, 2010; Monteiro *et al.*, 2012). sRNAs can also play a quorum sensing role, which has been observed in *Vibrio cholerae* (Svenningsen *et al.*, 2009). Quorum regulatory RNAs 1-4 (QRR1-4) are part of the quorum-sensing signal transduction pathway, which is controlled by two feedback loops involving the sRNA-activator LuxO and the hapR sRNA target (Svenningsen *et al.*, 2009). These feedback loops allow for gene dosage compensation between the four *orr* genes to prevent the unnecessary production of sRNA (Svenningsen *et al.*, 2009). The maintenance of these levels are required for quorum-sensing genes to be properly expressed (Svenningsen *et al.*, 2009).

Stressful conditions can cause the microorganism to adapt to its surroundings resulting in increased pathogenesis, which, in some part, can lead to virulence (Arnvig & Young, 2012). This has been observed with the RybB and MicA sRNAs, both found in *Salmonella*, which target the outer membrane protein mRNAs, cause these mRNAs to degrade when under acid stress resulting

in the subsequent rearrangement of the membrane as well as the host-pathogen boundary (Papenfort *et al.*, 2006). Very few sRNAs have a direct relationship with virulence, whereby virulence is affected by the deletion of specific sRNAs. A few examples include SprD from *S. aureus*, Rli38 from *L. monocytogenes*, ssrS from *Legionella pneumophilia* and the three sRNAs IsrM, IstR and SroA from *Salmonella* (Arnvig & Young, 2012).

1.9 sRNAs in mycobacteria

The *M. tb* genome is approximately 4.4 Mbp long, which encodes 13 sigma factors and more than 100 transcriptional regulators (Arnvig & Young, 2012). This acid-fast, facultative anaerobic bacillus has a large component of its cell wall being high-molecular weight lipids. It takes approximately 3-8 weeks for visible growth on solid media and the only reservoir are humans (Fitzgerald *et al.*, 2015). Approximately 20 putative sRNAs have been identified for *M. tb* with a few being restricted to the *M. tb* species while others also belong to the MTBC (Arnvig & Young, 2012). Others are found within the pathogenic mycobacteria family such as in *M. leprae*, which is the most distant relative, while some are conserved to the *Mycobacteria* family as well as in some actinomycetes (Arnvig & Young, 2012).

The sRNAs that have been previously identified in *M. tb* range between 50 to more than 300 nucleotides (Arnvig & Young, 2012). These have also been identified to be differentially expressed during changing physical conditions and some have been lethal to the bacillus when over expressed (Arnvig & Young, 2012).

Arnvig and Young (2009) identified *M. tb* sRNAs, MTS194 (F6), MTS497 (B55) and MTS2822 (B11) under different stress conditions. These conditions were used to identify expression changes that may occur during infection. One of the conditions included oxidative stress which was induced by hydrogen peroxide. This type of condition is commonly found within the host and the upregulation of all 3 sRNAs is an indication that they may be linked to the survival of the bacterium during the early stages of infection (Arnvig & Young, 2009). The overexpression of MTS2822 and G2 (MTS1310) have been shown to be lethal and MTS194 had exhibited slow growth (Arnvig & Young, 2009). The MTS194 homolog has also been found in *M. smegmatis* with no display of phenotype when overexpressed (Arnvig & Young, 2009). MTS194 is encoded between two genes, *Rv0234* and *Rv0244c*, which are responsible for lipid degradation (Hartkoorn *et al.*, 2012). This sRNA is transcribed by SIGF (sigma factor), which is a starvation-associated alternative sigma factor and could indicate that this sRNA is required under nutrient- depleted conditions (Hartkoorn *et al.*, 2012).

MTS2822 is approximately 95 nucleotides long and encoded between the *Rv3660c* and *Rv366l* genes which are suggested to play a role in cellular differentiation. This sRNA contains a

6C motif which comprises two C loops, suggesting that it may play a structural or protein binding role rather than a regulatory role (Weinberg *et al.*, 2007). The presence and conservation of both loops suggest that it is crucial for the functioning of the sRNA (Mai *et al.*, 2019). Fifteen potential targets were identified and experimentally confirmed, with MTS2822 inhibiting translation of each target in *M. smegatis* (Mai *et al.*, 2019). There is an sRNA that is more than 90% identical to MTS2822 in *M. smegmatis*, with both sRNA's having different outcomes due to overexpression in both species (Arnvig & Young, 2009). In *M. tb*, overexpression of MTS2822 is lethal to the bacillus, however in *M. smegmatis* overexpression results in poor growth and a change in cell morphology such as irregular shape and elongation which can suggest that the sRNA is associated with cell wall synthesis and/or cell division (Arnvig & Young, 2009).

Arnvig and Young (2009) elucidated variation of expression levels of sRNA transcripts in the *M. tb* growth at exponential and stationary phases. This can be seen in the sRNAs MTS997, MTS1338 and MTS2823 which become elevated during transition from the exponential and stationary phases and increases further during the infection state, indicating that it could play an important role in pathogenesis (Arnvig & Young, 2012). Expression of MTS997 and MTS2823 was observed during the exponential phases, however, less than 10% of MTS1338 was present, suggesting that the few cells that were expressing MTS1338 may have been in a different metabolic state, and may have been possible persister cells (Arnvig *et al.*, 2011). Strains overexpressing MTS1338 were found to be resistant to unfavourable conditions, in particular, a lower pH and when grown in normal conditions, expression profiles mimic cells grown in hypoxic conditions (Salina *et al.*, 2019).

During stationary phase it was found that MTS1338 was not expressed after the deletion of DOSR which is a dormancy regulator (Arnvig & Young, 2012). It included genes that are required for energy metabolism, lipid and protein remodelling and were upregulated during the infection stage in mice as well as patients' sputum samples (Arnvig & Young, 2012). Arnvig and Young (2012) investigated the MTS1338 involvement with the generation of persister cells and as a potential marker of this subpopulation (Arnvig & Young, 2012). MTS2823 is approximately 300 nucleotides, encoded between *Rv3661* and *Rv3662c* on the plus strand and is the most well studied sRNA. Conserved in most mycobacteria as well as some actinomycetes, it is also the most highly expressed non-rRNA during the exponential phase and a 10-fold increase was observed during the stationary phase, slowly approaching a 1:1 ratio with rRNA within infected mice tissues (Arnvig *et al.*, 2011). Overexpression of this sRNA during the exponential phase can also result in a reduced growth and downregulation of energy metabolic genes, which is analogous to that observed during the transition between exponential and stationary growth phases. The largest reduction was found with *prpc* and *prpd* genes which are associated with the methyl citrate cycle,

responsible for the removal of toxins that accumulate due to catabolism of odd-numbered fatty acids and cholesterol, an important carbon source during *M. tb* growth (Arnvig & Young, 2012).

Interestingly, a homolog of the *M. smegmatis* MTS2823 was identified as a potential 6S RNA candidate during a search for energetically suboptimal structures (Pánek *et al.*, 2011). The 6S RNA molecule is crucial during the transition of the stationary phase, whereby the active growth functions are downregulated. This was accomplished by the 6S RNA molecule binding to RNA polymerase resulting in the inhibition of the principal sigma factor controlling gene transcription (Barrick *et al.*, 2005; Wassarman, 2007a, 2007b). However, since there was a lack of binding of MTS2823 to RNA polymerase, it was disregarded as potential 6S RNA for *M. smegmatis* (Pánek *et al.*, 2011). MTS2823 still remains to be investigated as an sRNA which base pairs with its mRNA target or protein or if it has a completely different function and mode of function (Arnvig & Young, 2012).

A study completed by Gerrick *et al.* (2018) revealed a mycobacterial regulatory sRNA in iron (MrsI) through exposure of bacterial cells through stressful environments. Iron-limiting environments caused the sRNA to be upregulated slightly however, a significant upregulation was observed during bacterial membrane damage (Gerrick *et al.*, 2018). One of the MrsI targets includes the *bfrA* mRNA which encodes bacterioferritin, an iron-deficiency protein. The sRNA acts by suppressing the translation of iron-binding proteins causing a more economical route for the cell to utilize iron (Gerrick *et al.*, 2018).

sRNA mcr11 was identified by RNA-sequencing and confirmed through northern blotting, with increased expression noted when transitioning to the stationary phase (DiChiara *et al.*, 2010). This increased transcription was also noted during nutrient starvation while expression decreased when exposed to a low pH (Pelly *et al.*, 2012). The overexpression of mcr11 resulted in decreased growth of the mycobacterium and was shown to be regulated by an ATP-binding Mcr11 regulator (AbmR) (Ignatov *et al.*, 2015; Girardin *et al.*, 2018).

1.10 Techniques used for the identification of sRNAs

Previously sRNAs were rarely detected through genetic screening, however, through the development of bioinformatic tools and high-throughput sequencing methods, this process has become increasingly easier. Wet laboratory methods relied on microarray transcriptome experiments for the identification of sRNAs resulting in only tens of sRNAs being discovered in the past (Van Puyvelde *et al.*, 2015). High-throughput sequencing methods with an sRNA enrichment step is another approach to experimentally identify sRNAs (Thebault *et al.*, 2015). A study completed by Tsai *et al.* (2013) employed a computational method to predict sRNAs, Deep-RACE, to identify the 5' and 3' ends, ChIP-Seq to detect regulators of the sRNAs and Northern

blot to validate and experimentally confirm the sRNA predicted. Although experimental approaches are more accurate, it is not a feasible approach for the identification of sRNAs within a genome, therefore computational methods are favoured (Sridhar & Gunasekaran, 2013).

Computational identification of sRNAs from bacteria can be divided into four categories based on their methodology, namely, comparative genomics, secondary and thermodynamic stability, "orphan" transcriptional signals and non-sequence based detection methods (Sridhar & Gunasekaran, 2013). The comparative genomics approach utilizes whole genome alignment to identify consensus sRNAs regions as well employing consensus structure analysis. Majority of the computational sRNA identification tools use this method `and these include QRNA, ERPIN (Easy RNA Profile Identification), ISI (intergenic sequence inspector), INFERNAL (Inference of RNA Alignment) and RNAz. QRNA utilizes three probabilistic models and an algorithm to detect sRNAs (Rivas & Eddy, 2001). ERPIN, predicts sRNAs using multiple sequence alignment and secondary structure consensus to calculate a log-odds score and a secondary structure profile and a subsequent e-value is calculated based on both (Gautheret & Lambert, 2001). MSARI utilizes RNAFOLD to generate the secondary structure of the sRNA and CLUSTALW to perform the multiple alignments in conjunction with a distribution-mixture method to predict its sRNAs (Coventry *et al.*, 2004). INFERNAL scores RNAs based on both sequence and structural homology. The tool performs a homology search for putative sRNAs and then a structural based multiple alignment and an E-value is assigned (Nawrocki *et al.*, 2009). The final comparative genomic tool is ISI, which predicts sRNAs through intergenic region conservation, RNA terminators and structural features (Pichon & Felden, 2003).

Secondary structure prediction is a prerequisite for many computational approaches however, this structure does not depict the functional activity of the sRNA alone. The sRNA should also have a thermodynamically favourable minimum free energy for accurate sRNA prediction (Sridhar & Gunasekaran, 2013). RNAz, a more popular tool to efficiently detect sRNAs without several input data files (Sridhar & Gunasekaran, 2013). This bioinformatic approach calculates the initial fold of the sequence using RNAfold and the consensus fold of the aligned structures using RNAalifold. A regression analysis using synthetic sequences is then applied to calculate the z-score for each to determine if the RNA is a sRNA or not (Gruber *et al.*, 2007).

Transcriptional signal-based methods entails detecting 'orphan' transcriptional signals to predict sRNAs. sRNAPredict3/SIPHT utilizes promoter signals, transcription factor binding sites, rho-independent terminator signals as predicted features of sRNAs along with sequence conservation to detect sRNAs (Livny *et al.*, 2005). The sRNAfinder employs the same method as

sRNAPredict3 when predicting sRNAs (Tjaden, 2008). For instances when details on transcriptional signals are not available, sRNAscanner is useful. This predicts intergenic signals based on family specific training datasets which in turn detects sRNAs (Sridhar *et al.*, 2010).

Based on preferential occurrence of structural components, nucleotide preferences and GC properties of a sequence, *ab initio* methods for sRNA prediction can be applied. RNAGENiE predicts sRNAs by searching for the preferential occurrence of secondary structural elements. Sequence-based descriptors are also employed to differentiate RNA genes from non-RNA genes (Klein *et al.*, 2002). Atypical GC creates a sliding window to screen and compute the G and C content of a segment to predict sRNAs (Livny *et al.*, 2008). The differential distribution of sequence motifs between noncoding RNAs and background sequences have been exploited for the detection of sRNAs by smyRNA (Salari *et al.*, 2009). NAPP (Nucleic Acids Phylogenetic Profiling) creates 50 nucleotide 'tile' segments of the intergenomic region of the reference genome and compares it to 1000 genomic sequences (Ott *et al.*, 2012). The final bioinformatic tool used is PsRNA which locates sRNA specific intergenic regions using KEGG orthology numbers of the respective flanking genes (Sridhar & Rafi, 2007). The computational approach is only a predictive method and requires further experimental validation (Sridhar & Gunasekaran, 2013).

1.11 Rationale

The investigation of sRNAs in *M. tb* has become more popular in recent years (Ostrik *et al.*, 2021). Research can be aided by high-throughput methods which will provide a large amount of data on predicted sRNAs as well as their function. Data on the regulatory information of each sRNA could also be collected to provide a wholistic understanding of the role of the sRNA in the gene regulatory networks. Network mapping could also be increased by overexpressing or knocking out specific sRNAs (Haning *et al.*, 2014b). Changes in sRNAs have been largely investigated on the laboratory H37Rv strain (Liu *et al.*, 2016), with over 200 sRNAs being discovered in *Mycobacteria* (Haning *et al.*, 2014b). There have been no reports, however, on the role of these non-coding RNA species in clinical strains of *M. tb*.

The differences in gene expression of these sRNAs, particularly under environmental stress, suggests that these transcripts may be relevant to the pathogenesis of Mycobacteria (Haning *et al.*, 2014b). The study proposes to elucidate the role of sRNAs on transcriptome changes during *in vitro* growth of genetically diverse clinical strains of *M. tb*. Understanding the role of these sRNAs might offer novel insights into crucial gene regulatory networks and pathways that are exhibited by these strains, which might ultimately contribute to strain virulence.

Furthermore, strain-specific regulatory networks and pathways identified in this study may serve as potential targets for novel vaccine and drug development.

1.12 Aims

- To perform whole genome alignment and *in silico* mine for sRNAs containing lineage-specific mutations from the eight lineages of *M. tb*.
- To investigate the role that small non-coding RNAs play in the regulation of the *M. tb* genome in genetically diverse, drug resistant, clinical strains.

1.13 Objectives

- To map and assemble sequence reads belonging to the eight *M. tb* lineages obtained from a study completed by Borrell *et al*. (2019), using *Geneious Prime* (v.2021.1.1),
- To perform multiple sequence alignment between the generated whole genome sequences and the H37Rv strain to determine genetic differences using *Geneious Prime* (v.2021.1.1),
- To identify sRNAs containing lineage-specific single nucleotide polymorphisms using the *RNAz* webserver,
- To predict potential targets of the filtered sRNAs using *IntaRNA* (v 4.9.2), and their predicted secondary structures using the *RNAfold* webserver (v 2.4.18),
- To perform functional enrichment on the targets that exhibited significant variation between the consensus and variant sequences using Uniprot, Quickgo, PANTHER, MTB Network Portal, AMIGO2 and BioCyc,
- To culture *M. tb* Beijing, F11, and F15/LAM4/KZN families as well as the H37Rv and Unique strains using standard culturing techniques,
- To extract mRNA and small RNA using Zymogen Directzol kit, from the *M. tb* strains belonging to the Beijing, F11, F15/LAM4/KZN and Unique families as wells as the H37Rv strain during *in vitro* growth in Middlebrook 7H9 media,
- To sequence the mRNAs and sRNAs extracted from the *M. tb* clinical strains using the Illumina HiSeq 2000 sequencing platform and the NEBNext Ultra II with RiboZero Plus kit for mRNA and TruSeq smRNA kit for sRNA,

- To analyse and trim the subsequent *M. tb* clinical strain RNA-Seq reads using *FastQC* (v. 0.11.9), *Trimmomatic* (v. 0.39) and *TrimGalore* (v. 0.6.5).
- To map and assemble *M. tb* sequence reads using *HiSat2* (v. 2.1.1.) and *StringTie* (v. 2.1.4.),
- To compare mRNA and small RNA expression levels of the genetically diverse, clinical strains of *M. tb* using *Ballgown* (v. 4.1.1.),
- To visualize sRNA and mRNA significantly regulated transcripts fold changes using Multiple Experiment Viewer,
- To predict the effects of significantly differentially expressed *M. tb* mRNAs and sRNAs using the MTB Network Portal (v2), UniProt and BioCyc (v. 25.0), and
- To validate the expression patterns of the mRNA and sRNAs using quantitative real-time PCR (qRT-PCR).

References

Ahmad, S. (2011). Pathogenesis, immunology, and diagnosis of latent *Mycobacterium tuberculosis* infection. *Clinical and Developmental Immunology, 2011*, 1-17.

Alahari, A., Trivelli, X., Guerardel, Y., Dover, L. G., Besra, G. S., Sacchettini, J. C., Reynolds, R. C., Coxon, G. D., & Kremer, L. (2007). Thiacetazone, an Antitubercular Drug that Inhibits Cyclopropanation of Cell Wall Mycolic Acids in Mycobacteria. *Public Library of Science One, 2*(12), 1-12.

Alderwick, L. J., Birch, H. L., Mishra, A. K., Eggeling, L., & Besra, G. S. (2007). Structure, function and biosynthesis of the *Mycobacterium tuberculosis* cell wall: arabinogalactan and lipoarabinomannan assembly with a view to discovering new drug targets. *Biochemical Society Transactions, 35*(Pt 5), 1325-1328.

Arnvig, K., Comas, I., Thomson, N. R., Houghton, J., Boshoff, H. I., Croucher, N. J., Rose, G., Perkins, T. T., Parkhill, J., Dougan, G., & Young, D. B. (2011). Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of *Mycobacterium tuberculosis*. *Public Library of Science Pathogens, 7*(11), 1-16.

Arnvig, K., & Young, D. (2012). Non-coding RNA and its potential role in *Mycobacterium tuberculosis* pathogenesis. *RNA biology, 9*(4), 427-436.

Arnvig, K., & Young, D. B. (2009). Identification of small RNAs in *Mycobacterium tuberculosis*. *Molecular Microbiology, 73*(3), 397-408.

Barrick, J. E., Sudarsan, N., Weinberg, Z., Ruzzo, W. L., & Breaker, R. R. (2005). 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. *RNA, 11*, 774-784.

Becq, J., Gutierrez, M. C., Rosas-Magallanes, V., Rauzier, J., Gicquel, B., Neyrolles, O., & Deschavanne, P. (2007). Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli. *Molecular biology and evolution, 24*(8), 1861-1871.

Belisle, J. T., Vissa, V. D., Sievert, T., Takayama, K., Brennan, P. J., & Besra, G. S. (1997). Role of the major antigen of *Mycobacterium tuberculosis* in cell wall biogenesis. *Science, 276*(5317), 1420-1422.

Bloch, S., Wegrzyn, A., Wegrzyn, G., & Nejman-Falenczyk, B. (2017). Small and Smaller-sRNAs and MicroRNAs in the Regulation of Toxin Gene Expression in Prokaryotic Cells: A Mini-Review. *Toxins, 9*(6), 1-13.

Boritsch, E. C., Supply, P., Honore, N., Seemann, T., Stinear, T. P., & Brosch, R. (2014). A glimpse into the past and predictions for the future: the molecular evolution of the tuberculosis agent. *Molecular Microbiology, 93*(5), 835-852.

Borrell, S., Trauner, A., Brites, D., Rigouts, L., Loiseau, C., Coscolla, M., Niemann, S., De Jong, B., Yeboah-Manu, D., Kato-Maeda, M., Feldmann, J., Reinhard, M., Beisel, C., & Gagneux, S. (2019). Reference set of *Mycobacterium tuberculosis* clinical strains: A tool for research and product development. *Public Library of Science One, 14*(3), 1-12.

Brown, A. K., Taylor, R. C., Bhatt, A., Futterer, K., & Besra, G. S. (2009). Platensimycin activity against mycobacterial beta-ketoacyl-ACP synthases. *Public Library of Science One, 4*(7), 1-10.

Cadena, A. M., Fortune, S. M., & Flynn, J. L. (2017). Heterogeneity in tuberculosis. *Nature Reviews Immunology, 17*(11), 691-702.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., 3rd, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S., & Barrell, B. G. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature, 393*(6685), 537- 544.

Conceicao, E. C., Refregier, G., Gomes, H. M., Olessa-Daragon, X., Coll, F., Ratovonirina, N. H., Rasolofo-Razanamparany, V., Lopes, M. L., van Soolingen, D., Rutaihwa, L., Gagneux, S., Bollela, V. R., Suffys, P. N., Duarte, R. S., Lima, K. V. B., & Sola, C. (2019). *Mycobacterium tuberculosis* lineage 1 genetic diversity in Para, Brazil, suggests

common ancestry with east-African isolates potentially linked to historical slave trade. *Infection, Genetics and Evolution, 73*, 337-341.

Coscolla, M., & Gagneux, S. (2014). Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Seminars in Immunology, 26*(6), 431-444.

Coventry, A., Kleitman, D. J., & Berger, B. (2004). MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America, 101*(33), 12102-12107.

Daffe, M., & Draper, P. (1997). The envelope layers of mycobacteria with reference to their pathogenicity. *Advances in microbial physiology, 39*, 131-203.

de Gijsel, D., & von Reyn, C. F. (2019). A Breath of Fresh Air: BCG Prevents Adult Pulmonary Tuberculosis. *International Journal of Infectious Diseases, 80S*, S6-S8.

Delogu, G., Sali, M., & Fadda, G. (2013). The biology of *mycobacterium tuberculosis* infection. *Mediterranean Journal of Hematology and Infectious Diseases, 5*(1), 1-8.

Demay, C., Liens, B., Burguiere, T., Hill, V., Couvin, D., Millet, J., Mokrousov, I., Sola, C., Zozio, T., & Rastogi, N. (2012). SITVITWEB – A publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. *Infection, Genetics and Evolution, 12*(4), 755-766.

DiChiara, J. M., Contreras-Martinez, L. M., Livny, J., Smith, D., McDonough, K. A., & Belfort, M. (2010). Multiple small RNAs identified in *Mycobacterium bovis* BCG are also expressed in *Mycobacterium tuberculosis* and *Mycobacterium smegmatis*. *Nucleic Acids Research, 38*(12), 4067-4078.

Dong, T., & Schellhorn, H. E. (2010). Role of RpoS in virulence of pathogens. *Infection and Immunity, 78*(3), 887-897.

Fitzgerald, D. W., Sterling, T. R., & Haas, D. W. (2015). *Mycobacterium tuberculosis*. In J. E. Bennett, R. Dolin, & M. J. Blaser (Eds.), *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases* (Vol. 2, pp. 2787-2818). New York: Elsevier.

Fratti, R. A., Chua, J., Vergne, I., & Deretic, V. (2003). *Mycobacterium tuberculosis* glycosylated phosphatidylinositol causes phagosome maturation arrest. *Proceedings of the National Academy of Sciences, 100*(9), 5437-5442.

Gagneux, S. (2018). Ecology and evolution of *Mycobacterium tuberculosis*. *Nature Reviews Microbiology, 16*(4), 202-213.

Gautheret, D., & Lambert, A. (2001). Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *Journal of Molecular Biology, 313*(5), 1003-1011.

Gerrick, E. R., Barbier, T., Chase, M. R., Xu, R., Francois, J., Lin, V. H., Szucs, M. J., Rock, J. M., Ahmad, R., Tjaden, B., Livny, J., & Fortune, S. M. (2018). Small RNA profiling in

Mycobacterium tuberculosis identifies MrsI as necessary for an anticipatory iron sparing response. *Proceedings of the National Academy of Sciences of the United States of America, 115*(25), 6464-6469.

Gey van Pittius, N. C., Sampson, S. L., Lee, H., Kim, Y., van Helden, P. D., & Warren, R. M. (2006). Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. *BMC Evolutionary Biology, 6*(95), 1-31.

Girardin, R. C., Bai, G., He, J., Sui, H., & McDonough, K. A. (2018). AbmR (Rv1265) is a novel transcription factor of *Mycobacterium tuberculosis* that regulates host cell association and expression of the non-coding small RNA Mcr11. *Molecular Microbiology, 110*(5), 811-830.

Goletti, D., Lindestam Arlehamn, C. S., Scriba, T. J., Anthony, R., Cirillo, D. M., Alonzi, T., Denkinger, C. M., & Cobelens, F. (2018). Can we predict tuberculosis cure? What tools are available? *European Respiratory Journal, 52*(5), 1-18.

Gottesman, S., & Storz, G. (2011). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harbor Perspectives in Biology, 3*(12), 1-16.

Gruber, A. R., Neubock, R., Hofacker, I. L., & Washietl, S. (2007). The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Research, 35*(Web Server issue), W335-338.

Guirado, E., Schlesinger, L. S., & Kaplan, G. (2013). Macrophages in tuberculosis: friend or foe. *Seminars in Immunopathology, 35*(5), 563-583.

Haning, K., Cho, S. H., & Contreras, L. M. (2014b). Small RNAs in mycobacteria: an unfolding story. *Frontiers in Cellular and Infection Microbiology, 4*, 96.

Hartkoorn, R. C., Sala, C., Uplekar, S., Busso, P., Rougemont, J., & Cole, S. T. (2012). Genome-wide definition of the SigF regulon in *Mycobacterium tuberculosis*. *Journal of Bacteriology, 194*(8), 2001-2009.

Hernandez-Pando, R., Jeyanathan, M., Mengistu, G., Aguilar, D., Orozco, H., Harboe, M., Rook, G. A., & Bjune, G. (2000). Persistence of DNA from *Mycobacterium tuberculosis* in superficially normal lung tissue during latent infection. *Lancet, 356*(9248), 2133-2138.

Hett, E. C., Chao, M. C., & Rubin, E. J. (2010). Interaction and modulation of two antagonistic cell wall enzymes of mycobacteria. *The Public Library of Science Pathogens, 6*(7), 1-14.

Houben, E. N., Bestebroer, J., Ummels, R., Wilson, L., Piersma, S. R., Jimenez, C. R., Ottenhoff, T. H., Luirink, J., & Bitter, W. (2012). Composition of the type VII secretion system membrane complex. *Molecular Microbiology, 86*(2), 472-484.

Ignatov, D. V., Salina, E. G., Fursov, M. V., Skvortsov, T. A., Azhikina, T. L., & Kaprelyants, A. S. (2015). Dormant non-culturable *Mycobacterium tuberculosis* retains stable low-abundant mRNA. *BMC Genomics, 16*, 954.

Jayaweera, M., Perera, H., Gunawardana, B., & Manatunge, J. (2020). Transmission of COVID-19 virus by droplets and aerosols: A critical review on the unresolved dichotomy. *Environmental Research, 188*, 1-19.

Kant, S., & Tyagi, R. (2021). The impact of COVID-19 on tuberculosis: challenges and opportunities. *Therapeutic Advances in Infectious Disease, 8*, 1-7.

Klein, R. J., Misulovin, Z., & Eddy, S. R. (2002). Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proceedings of the National Academy of Sciences of the United States of America, 99*(11), 7542-7547.

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.

Kumar, A., Toledo, J. C., Patel, R. P., Lancaster, J. R., Jr., & Steyn, A. J. (2007). *Mycobacterium tuberculosis* DosS is a redox sensor and DosT is a hypoxia sensor. *Proceedings of the National Academy of Sciences of the United States of America, 104*(28), 11568-11573.

Livny, J., Fogel, M. A., Davis, B. M., & Waldor, M. K. (2005). sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic Acids Research, 33*(13), 4096-4105.

Livny, J., Teonadi, H., Livny, M., & Waldor, M. K. (2008). High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. *Public Library of Science One, 3*(9), 1-12.

Mai, J., Rao, C., Watt, J., Sun, X., Lin, C., Zhang, L., & Liu, J. (2019). *Mycobacterium tuberculosis* 6C sRNA binds multiple mRNA targets via C-rich loops independent of RNA chaperones. *Nucleic Acids Research, 47*(8), 4292-4307.

Mandin, P., & Gottesman, S. (2010). Integrating anaerobic/aerobic sensing and the general stress response through the ArcZ small RNA. *European Molecular Biology Organization Journal, 29*(18), 3094-3107.

Meo, S. A., Abukhalaf, A. A., Alomar, A. A., AlMutairi, F. J., Usmani, A. M., & Klonoff, D. C. (2020). Impact of lockdown on COVID-19 prevalence and mortality during 2020 pandemic: observational analysis of 27 countries. *European Journal of Medical Research, 25*(1), 1-7

Miotto, P., Forti, F., Ambrosi, A., Pellin, D., Veiga, D. F., Balazsi, G., Gennaro, M. L., Di Serio, C., Ghisotti, D., & Cirillo, D. M. (2012b). Genome-wide discovery of small RNAs in *Mycobacterium tuberculosis*. *Public Library of Science One, 7*(12), e51950.

Monteiro, C., Papenfort, K., Hentrich, K., Ahmad, I., Le Guyon, S., Reimann, R., Grantcharova, N., & Romling, U. (2012). Hfq and Hfq-dependent small RNAs are major contributors to multicellular development in *Salmonella enterica* serovar Typhimurium. *RNA biology, 9*(4), 489-502.

Mukamolova, G. V., Turapov, O., Malkin, J., Woltmann, G., & Barer, M. R. (2010). Resuscitation-promoting factors reveal an occult population of tubercle Bacilli in Sputum. *American Journal of Respiratory and Critical Care Medicine, 181*(2), 174-180.

Mukhopadhyay, S., Nair, S., & Ghosh, S. (2012). Pathogenesis in tuberculosis: transcriptomic approaches to unraveling virulence mechanisms and finding new drug targets. *FEMS Microbiology Reviews, 36*(2), 463-485.

Nawrocki, E. P., Kolbe, D. L., & Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics, 25*(10), 1335-1337.

Neyrolles, O., Hernandez-Pando, R., Pietri-Rouxel, F., Fornes, P., Tailleux, L., Payan, J. A. B., Pivert, E., Bordat, Y., Aguilar, D., Prevost, M. C., Petit, C., & Gicquel, B. (2006). Is Adipose Tissue a Place for *Mycobacterium tuberculosis* Persistence? *Public Library of Science One, 1*(1), 1-9.

Orgeur, M., & Brosch, R. (2018). Evolution of virulence in the *Mycobacterium tuberculosis* complex. *Current Opinion in Microbiology, 41*, 68-75.

Ostrik, A. A., Azhikina, T. L., & Salina, E. G. (2021). Small Noncoding RNAs and Their Role in the Pathogenesis of *Mycobacterium tuberculosis* Infection. *Biochemistry Moscow, 86*(Suppl 1), S109-S119.

Ott, A., Idali, A., Marchais, A., & Gautheret, D. (2012). NAPP: the Nucleic Acid Phylogenetic Profile Database. *Nucleic Acids Research, 40*(Database issue), D205-209.

Pánek, J., Krásny, L., Bobek, J., Jezková, E., Korelusová, J., & Vohradsky, J. (2011). The suboptimal structures find the optimal RNAs: homology search for bacterial non-coding RNAs using suboptimal RNA structures. *Nucleic Acids Research, 39*, 3418-3426.

Papenfort, K., Pfeiffer, V., Mika, F., Lucchini, S., Hinton, J. C., & Vogel, J. (2006). SigmaE-dependent small RNAs of Salmonella respond to membrane stress by accelerating global omp mRNA decay. *Molecular Microbiology, 62*(6), 1674-1688.

Pelly, S., Bishai, W. R., & Lamichhane, G. (2012). A screen for non-coding RNA in *Mycobacterium tuberculosis* reveals a cAMP-responsive RNA that is expressed during infection. *Gene, 500*(1), 85-92.

Phetsuksiri, B., Jackson, M., Scherman, H., McNeil, M., Besra, G. S., Baulard, A. R., Slayden, R. A., DeBarber, A. E., Barry, C. E., 3rd, Baird, M. S., Crick, D. C., & Brennan, P. J. (2003). Unique mechanism of action of the thiourea drug isoxyl on *Mycobacterium tuberculosis. Journal of Biological Chemistry, 278*(52), 53123-53130.

Pichon, C., & Felden, B. (2003). Intergenic sequence inspector: searching and identifying bacterial RNAs. *Bioinformatics, 19*(13), 1707-1709.

Rachman, H., Strong, M., Ulrichs, T., Grode, L., Schuchhardt, J., Mollenkopf, H., Kosmiadi, G. A., Eisenberg, D., & Kaufmann, S. H. (2006). Unique transcriptome signature of *Mycobacterium tuberculosis* in pulmonary tuberculosis. *Infection and Immunity, 74*(2), 1233-1242.

Reed, M. B., Pichler, V. K., McIntosh, F., Mattia, A., Fallow, A., Masala, S., Domenech, P., Zwerling, A., Thibert, L., Menzies, D., Schwartzman, K., & Behr, M. A. (2009). Major *Mycobacterium tuberculosis* lineages associate with patient country of origin. *Journal of Clinical Microbiology, 47*(4), 1119-1128.

Reva, O., Korotetskiy, I., & Ilin, A. (2015). Role of the horizontal gene exchange in evolution of pathogenic Mycobacteria. *BMC Evolutionary Biology, 15* (Suppl 1), 1-8.

Rivas, E., & Eddy, S. R. (2001). Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics, 2*(8), 1-19.

Romagnoli, A., Etna, M. P., Giacomini, E., Pardini, M., Remoli, M. E., Corazzari, M., Falasca, L., Goletti, D., Gafa, V., Simeone, R., Delogu, G., Piacentini, M., Brosch, R., Fimia, G. M., & Coccia, E. M. (2012). ESX-1 dependent impairment of autophagic flux by *Mycobacterium tuberculosis* in human dendritic cells. *Autophagy, 8*(9), 1357-1370.

Roy, S., Ghatak, D., Das, P., & BoseDasgupta, S. (2020). ESX secretion system: The gatekeepers of mycobacterial survivability and pathogenesis. *European Journal of Microbiology & Immunology, 10*(4), 202-209.

Sala, A., Bordes, P., & Genevaux, P. (2014). Multiple toxin-antitoxin systems in *Mycobacterium tuberculosis. Toxins, 6*(3), 1002-1020.

Salari, R., Aksay, C., Karakoc, E., Unrau, P. J., Hajirasouliha, I., & Sahinalp, S. C. (2009). smyRNA: a novel Ab initio ncRNA gene finder. *Public Library of Science One, 4*(5), 1-6.

Salina, E. G., Grigorov, A., Skvortsova, Y., Majorov, K., Bychenko, O., Ostrik, A., Logunova, N., Ignatov, D., Kaprelyants, A., Apt, A., & Azhikina, T. (2019). MTS1338, A Small *Mycobacterium tuberculosis* RNA, Regulates Transcriptional Shifts Consistent With Bacterial Adaptation for Entering Into Dormancy and Survival Within Host Macrophages. *Frontiers in Cellular and Infection Microbiology, 9*(405), 1-11.

Sasindran, S., & Torrelles, J. (2011). *Mycobacterium Tuberculosis* Infection and Inflammation: what is Beneficial for the Host and for the Bacterium? *Frontiers in Microbiology, 2*(2), 1-16.

Senghore, M., Diarra, B., Gehre, F., Otu, J., Worwui, A., Muhammad, A. K., Kwambana-Adams, B., Kay, G. L., Sanogo, M., Baya, B., Orsega, S., Doumbia, S., Diallo, S., de Jong, B. C., Pallen, M. J., & Antonio, M. (2020). Evolution of *Mycobacterium tuberculosis* complex lineages and their role in an emerging threat of multidrug resistant tuberculosis in Bamako, Mali. *Scientific Reports, 10*(327), 1-9.

Serafini, A., Boldrin, F., Palu, G., & Manganelli, R. (2009). Characterization of a *Mycobacterium tuberculosis* ESX-3 conditional mutant: essentiality and rescue by iron and zinc. *Journal of Bacteriology, 191*(20), 6340-6344.

Shuaib, Y. A., Khalil, E. A. G., Wieler, L. H., Schaible, U. E., Bakheit, M. A., Mohamed-Noor, S. E., Abdalla, M. A., Kerubo, G., Andres, S., Hillemann, D., Richter, E., Kranzer, K., Niemann, S., & Merker, M. (2020). *Mycobacterium tuberculosis* Complex Lineage 3 as Causative Agent of Pulmonary Tuberculosis, Eastern Sudan. *Emerging Infectious Diseases, 26*(3), 427-436.

Sridhar, J., & Gunasekaran, P. (2013). Computational small RNA prediction in bacteria. *Bioinformatics and Biology Insights, 7*, 83-95.

Sridhar, J., & Rafi, Z. A. (2007). Small RNA identification in Enterobacteriaceae using synteny and genomic backbone retention. *OMICS: A Journal of Integrative Biology, 11*(1), 74-99.

Sridhar, J., Sambaturu, N., Sabarinathan, R., Ou, H. Y., Deng, Z., Sekar, K., Rafi, Z. A., & Rajakumar, K. (2010). sRNAscanner: a computational tool for intergenic small RNA detection in bacterial genomes. *Public Library of Science One, 5*(9), 1-14.

Stinear, T. P., Seemann, T., Harrison, P. F., Jenkin, G. A., Davies, J. K., Johnson, P. D., Abdellah, Z., Arrowsmith, C., Chillingworth, T., Churcher, C., Clarke, K., Cronin, A., Davis, P., Goodhead, I., Holroyd, N., Jagels, K., Lord, A., Moule, S., Mungall, K., Norbertczak, H., Quail, M. A., Rabbinowitsch, E., Walker, D., White, B., Whitehead, S., Small, P. L., Brosch, R., Ramakrishnan, L., Fischbach, M. A., Parkhill, J., & Cole, S. T. (2008). Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*. *Genome Research, 18*(5), 729-741.

Storz, G., Vogel, J., & Wassarman, K. M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. *Molecular Cell, 43*(6), 880-891.

Stucki, D., Brites, D., Jeljeli, L., Coscolla, M., Liu, Q., Trauner, A., Fenner, L., Rutaihwa, L., Borrell, S., Luo, T., Gao, Q., Kato-Maeda, M., Ballif, M., Egger, M., Macedo, R., Mardassi, H., Moreno, M., Tudo Vilanova, G., Fyfe, J., Globan, M., Thomas, J.,

Jamieson, F., Guthrie, J. L., Asante-Poku, A., Yeboah-Manu, D., Wampande, E., Ssengooba, W., Joloba, M., Henry Boom, W., Basu, I., Bower, J., Saraiva, M., Vaconcellos, S. E. G., Suffys, P., Koch, A., Wilkinson, R., Gail-Bekker, L., Malla, B., Ley, S. D., Beck, H. P., de Jong, B. C., Toit, K., Sanchez-Padilla, E., Bonnet, M., Gil-Brusola, A., Frank, M., Penlap Beng, V. N., Eisenach, K., Alani, I., Wangui Ndung'u, P., Revathi, G., Gehre, F., Akter, S., Ntoumi, F., Stewart-Isherwood, L., Ntinginya, N. E., Rachow, A., Hoelscher, M., Cirillo, D. M., Skenders, G., Hoffner, S., Bakonyte, D., Stakenas, P., Diel, R., Crudu, V., Moldovan, O., Al-Hajoj, S., Otero, L., Barletta, F., Jane Carter, E., Diero, L., Supply, P., Comas, I., Niemann, S., & Gagneux, S. (2016). *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nature Genetics, 48*(12), 1535-1543.

Svenningsen, S. L., Tu, K. C., & Bassler, B. L. (2009). Gene dosage compensation calibrates four regulatory RNAs to control Vibrio cholerae quorum sensing. *European Molecular Biology Organization Journal*, *28*(4), 429-439.

Thain, N., Le, C., Crossa, A., Ahuja, S. D., Meissner, J. S., Mathema, B., Kreiswirth, B., Kurepina, N., Cohen, T., & Chindelevitch, L. (2019). Towards better prediction of *Mycobacterium tuberculosis* lineages from MIRU-VNTR data. *Infection, Genetics and Evolution, 72*, 59-66.

Thebault, P., Bourqui, R., Benchimol, W., Gaspin, C., Sirand-Pugnet, P., Uricaru, R., & Dutour, I. (2015). Advantages of mixing bioinformatics and visualization approaches for analyzing sRNA-mediated regulatory bacterial networks. *Briefings in Bioinformatics, 16*(5), 795-805.

Tjaden, B. (2008). Prediction of small, noncoding RNAs in bacteria using heterogeneous data. *Journal of Mathematical Biology, 56*(1-2), 183-200.

Tsai, C. H., Baranowski, C., Livny, J., McDonough, K. A., Wade, J. T., & Contreras, L. M. (2013). Identification of novel sRNAs in mycobacterial species. *Public Library of Science One, 8*(11), 1-8.

van den Boogaard, J., Kibiki, G. S., Kisanga, E. R., Boeree, M. J., & Aarnoutse, R. E. (2009). New drugs against tuberculosis: problems, progress, and evaluation of agents in clinical development. *Antimicrobial Agents and Chemotherapy, 53*(3), 849-862.

van der Wel, N., Hava, D., Houben, D., Fluitsma, D., van Zon, M., Pierson, J., Brenner, M., & Peters, P. J. (2007). *M. tuberculosis* and *M. leprae* translocate from the phagolysosome to the cytosol in myeloid cells. *Cell, 129*(7), 1287-1298.

Van Puyvelde, S., Vanderleyden, J., & De Keersmaecker, S. C. (2015). Experimental approaches to identify small RNAs and their diverse roles in bacteria--what we have learnt in one decade of MicA research. *Microbiology, 4*(5), 699-711.

Velayati, A. A., & Farnia, P. (2017). Chapter 1 - The Species Concept. In A. A. Velayati & P. Farnia (Eds.), *Atlas of Mycobacterium Tuberculosis*: Academic Press.

Velayati, A. A., Farnia, P., & Masjedi, M. R. (2013). The totally drug resistant tuberculosis (TDR-TB). *International Journal of Clinical and Experimental Medicine, 6*(4), 307-309.

Veyrier, F. J., Dufort, A., & Behr, M. A. (2011). The rise and fall of the Mycobacterium tuberculosis genome. *Trends in Microbiology, 19*(4), 156-161.

Veyrier, F. J., Pletzer, D., Turenne, C., & Behr, M. A. (2009). Phylogenetic detection of horizontal gene transfer during the step-wise genesis of *Mycobacterium tuberculosis*. *BMC Evolutionary Biology, 9*(196), 1-14.

Voskuil, M. I., Visconti, K. C., & Schoolnik, G. K. (2004). *Mycobacterium tuberculosis* gene expression during adaptation to stationary phase and low-oxygen dormancy. *Tuberculosis, 84*(3-4), 218-227.

Wang, J., McIntosh, F., Radomski, N., Dewar, K., Simeone, R., Enninga, J., Brosch, R., Rocha, E. P., Veyrier, F. J., & Behr, M. A. (2015). Insights on the emergence of *Mycobacterium tuberculosis* from the analysis of Mycobacterium kansasii. *Genome Biology and Evolution, 7*(3), 856-870.

Wassarman, K. M. (2007a). 6S RNA: a regulator of transcription. *Molecular Microbiology, 65*, 1425-1431.

Wassarman, K. M. (2007b). 6S RNA: a small RNA regulator of transcription. *Current Opinion in Microbiology, 10*(2), 164-168.

Weinberg, Z., Barrick, J. E., Yao, Z., Roth, A., Kim, J. N., & Gore, J. (2007). Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Research, 35*, 4809-4819.

World Health Organization. (2019). *Global tuberculosis report 2019* (978-92-4-156571-4). Retrieved from Geneva:

World Health Organization. (2021). *Global tuberculosis report 2021* (978-92-4-003702-1). Retrieved from Geneva:

Yimer, S. A., Norheim, G., Namouchi, A., Zegeye, E. D., Kinander, W., Tonjum, T., Bekele, S., Mannsaker, T., Bjune, G., Aseffa, A., & Holm-Hansen, C. (2015). *Mycobacterium tuberculosis* lineage 7 strains are associated with prolonged patient delay in seeking treatment for pulmonary tuberculosis in Amhara Region, Ethiopia. *Journal of Clinical Microbiology, 53*(4), 1301-1309.

CHAPTER 2

This chapter has been formatted, submitted and revision invited by the International Journal of Molecular Sciences

Title: *Mycobacterium tuberculosis* complex transcriptome is regulated by lineage-specific small RNAs mutations and abundance

Divenita Govender [1], Manormoney Pillay [2] and Nontobeko Eunice Mvubu [1*]

1. Microbiology, School of Life Sciences, College of Agriculture, Engineering and Science, University of Kwa-Zulu-Natal, Westville 3630, South Africa; divenitagovender@gmail.com; mvubun@ukzn.ac.za

2. Medical Microbiology, School of Laboratory Medicine and Medical Sciences, College of Health Sciences, University of KwaZulu-Natal, 719 Umbilo Road, Private Bag 7, Congella 4013, Durban; pillayc@ukzn.ac.za

* Correspondence: mvubun@ukzn.ac.za; Tel: +2731 260 7404

Abstract

Transcriptome regulation in bacteria is essential for the adaptation to ever changing environments. Small RNAs (sRNA) have been identified to play a role in this regulation as well as in virulence and pathogenicity. Very little is known about the sRNAs within the *Mycobacterium tuberculosis* complex (MTBC) and even less with respect to the lineage-specific differences. This study aims to investigate the role of sRNA in transcriptome regulation of the MTBC lineages and their abundance in clinical strains of *M. tuberculosis*. An *in silico* analysis was performed to identify sRNAs with lineage-specific mutations and their respective targets. The sRNA and mRNA extracted from clinical Beijing and F15/LAM4/KZN strains were sequenced and the data was analyzed using a Hisat, Stringtie and Ballgown Bioinformatics pipeline. RT-qPCR was performed for clinical strains belonging to the F11 and Unique families to determine novel sRNA expression within those strains. Here, a total of twenty-eight sRNAs were predicted containing lineage-specific mutations and their respective targets from the *in silico* analysis and four using sRNA sequencing. The *in silico* analysis revealed sRNAs that may potentially implicate macrophage entry, lipid biosynthesis and environment adaptation mechanisms in a lineage-specific manner. All sRNAs presented were predicted to be novel with different targets identified for the consensus and mutated transcripts. Of the four sRNAs identified through sRNA sequencing, MSTRG.40.1 had a varying fold change of 2.1 and 1.27 for the Beijing and F15/LAM4/KZN strains, respectively, inferring that this sRNA may play a regulatory role within these strains. The predicted sRNA targets were absent from the mRNA sequencing data, indicating a potential inhibitory role of these sRNAs. Identification and characterization of sRNAs within the MTBC lineages may provide a novel perspective in control strategies against this human-adapted, globally prevalent infectious pathogen.

Keywords: small RNAs; *Mycobacterium tuberculosis*; transcriptome; regulation; lineages; RNA sequencing; bioinformatics

2.1 Introduction

Tuberculosis (TB) is one of the 10 causes of death across the world and is a threat to health with an incidence rate of 127 new cases per 100 000 population and an estimated death toll of 1.5 million in 2020 (World Health Organization, 2021). South Africa, one of the 30 high TB burden countries, contributed to 3.3% of the total global incidence and a TB mortality of 61 000 in 2020 (World Health Organization, 2021). The most widespread TB infections in South Africa were caused by the W/Beijing and the F15/LAM4/KZN families, of the East-Asian (EA) and Euro-American (EAM) lineages, respectively, with the latter family being the most common in the KwaZulu-Natal province in 2014 (Gandhi *et al.*, 2014). With the COVID-19 pandemic plaguing the globe, a 1.3 million drop in the number of newly diagnosed people has been reported (World Health Organization, 2021). This is a result of reduced access to TB treatment and testing facilities causing a drastic increase in TB related deaths (World Health Organization, 2021). It is believed that the important role of small RNAs (sRNAs) in the regulation of the mycobacterial transcriptome contributes to their success as pathogens (Ostrik *et al.*, 2020). With over 200 sRNAs being identified within the *Mycobacterium tuberculosis* complex (MTBC), the majority are predicted while a few have been confirmed and functionally characterized (Haning *et al.*, 2014a).

Transcriptional regulation is a critical cellular process required for the survival of bacteria, particularly those found in constantly changing environmental conditions (Casamassimi & Ciccodicola, 2019). Transcriptional regulation is controlled by various mechanisms, one of which is sRNAs that are found to be the most abundant category of post-transcriptional regulators in bacteria (Papenfort & Vogel, 2009). Over recent years sRNAs have emerged as key players in the regulation of gene expression, gene silencing, DNA maintenance as well as mRNA stability (Jost *et al.*, 2011; Storz *et al.*, 2011). These RNA molecules range between 50-500 nucleotides in length and perform regulation through a variety of mechanisms. Many execute control by either extensively or partially binding to their mRNA target through base pairing. Others mimic secondary structures of nucleic acids resulting in control at the protein level by affecting their activity (Gottesman & Storz, 2011).

The Mcr7 was one of the first sRNAs to be classified in *M. tb*. This transcript forms part of the PhoPR two-component system, which is essential for virulence in *M. tb*. The Mcr7 has been found to modulate translation of the tatC mRNA affecting the Twin Arginine Translocation protein secretion complex causing a cascade of protein concentrations being affected (Solans *et al.*, 2014). The MTBC exclusive MTS1338 sRNA has been identified to have a direct interaction with macrophages (Salina *et al.*, 2019). The expression of this transcript allows for adaptation of

the bacterium within the macrophage environment. MTS1338 overexpression improved survival at low pH conditions, however, numerous transcriptome changes were also triggered, such as the reduction in the translational activity and decreased bacterial growth. This was indicative of the bacterium transitioning to a dormant state (Salina *et al.*, 2019). The mycobacterial regulatory sRNA in iron (Mrsl) directly binds to one of the targets *bfrA*, which is a response that is activated during iron-limiting conditions resulting in iron-sparing, thus allowing for survival of the *Mycobacterium*. This is activated as an anticipatory response toward iron deprivation that is typical within a macrophage environment (Gerrick *et al.*, 2018). These are just a few of the small transcripts identified in MTBC, which mainly target several mRNAs, in an antisense manner, and are activated as a stress, virulence, pathogenicity and adaptive response (Storz *et al.*, 2011; Michaux *et al.*, 2014).

Although the genomic sequences are highly similar within the MTBC, even more so among the *M. tb* strains, there is still significant phenotypic variation among them, which can be attributed to their transcriptome regulation (Homolka *et al.*, 2010; Haning *et al.*, 2014a). Understanding the underlying molecular mechanisms controlling the MTBC strains, particularly those involved in virulence and pathogenicity, would contribute to the development of novel strategies that can be applied for the treatment of the TB disease (Ami *et al.*, 2020). To date, there is limited information and understanding of the MTBC sRNAs and even fewer have been functionally characterized, with none being investigated within clinical strains and in a lineage-specific manner. Due to the significant role sRNAs play in the modulation of pathogen's transcriptome, these transcripts are ideal to explore as novel therapeutic targets as they may greatly contribute to the lineage-specific transcriptome regulation. The current study investigated lineage-specific mutations in the sRNA genes, *in silico* and their effect on transcriptome regulation through sRNA and mRNA sequencing. Furthermore, this study reported on sRNA and mRNA sequencing of the KwaZulu-Natal province (South Africa), prevalent F15/LAM4/KZN and Beijing strains of the Euro-American and East-Asian lineages, respectively, to uncover strain-specific novel sRNAs and their respective transcript abundances.

2.2. Materials and Methods

2.2.1 *In silico* data analysis

2.2.1.1 Data acquisition and multiple sequence alignment

Sixteen strains, with two per lineage (except for one) were selected from the reference set of *Mycobacterium tuberculosis* clinical strains Borrell *et al.* (2019). Publicly available whole

genome sequence reads were obtained for each of the selected strains, including a reference, from the European Nucleotide Archive (ENA) generated by Borrell *et al*. (2019) as well as whole genome sequences from the Genbank database (https://www.ncbi.nlm.nih.gov/genbank/) (Table 2.1).

Table 2.1: MTBC clinical strains that have been investigated, *in silico*, in the study and the corresponding lineages and source.

Sequence ID	**Lineage**	**Sequence reads/Whole genome**	**Source**
ERR2704679	1 (Indo-Oceanic)	Sequence reads	(Borrell *et al.*, 2019)
ERR2704680	1	Sequence reads	(Borrell *et al.*, 2019)
CM001043.1	2 (East Asia)	Whole genome	(Ioerger *et al.*, 2010)
CP011510.1	2	Whole genome	(W. Li & Lv, 2015)
ERR2704693	3 (East-African-Indian)	Sequence reads	(Borrell *et al.*, 2019)
ERR2704678	3	Sequence reads	(Borrell *et al.*, 2019)
ERR2704705	4 (Euro America)	Sequence reads	(Borrell *et al.*, 2019)
CP001976.1	4	Whole genome	(Galagan *et al.*, 2010)
ERR2704686	5 (West Africa 1)	Sequence reads	(Borrell *et al.*, 2019)
ERR2704708	5	Sequence reads	(Borrell *et al.*, 2019)
ERR2704687	6 (West Africa 2)	Sequence reads	(Borrell *et al.*, 2019)
ERR2704681	6	Sequence reads	(Borrell *et al.*, 2019)
ERR2704711	7 (Ethiopia)	Sequence reads	(Borrell *et al.*, 2019)
ERR756347	7	Sequence reads	(Borrell *et al.*, 2019)
CP048071.1	8	Whole genome	(Ngabonziza *et al.*, 2020)
NC_000962	Reference	Whole genome	(Lew *et al.*, 2011)

Sequence reads were trimmed and assembled to the *M. tb* H37Rv reference genome using Geneious prime (v.2021.1.1) (https://www.geneious.com) to produce whole genome sequences

under default parameters. Whole genome multiple alignment was performed using the subsequent sequences as well as the whole genome sequences obtained from GenBank. This was accomplished using the Geneious prime 2021.1.1 and the Mauve (Darling *et al.*, 2010) plugin under default parameters. Three multiple alignment FASTA files containing 6 strains each were obtained.

2.2.1.2 Identification of putative sRNAs containing single nucleotide polymorphisms

The individual FASTA alignment files were uploaded to the RNAz WebServer (http://rna.tbi.univie.ac.at/cgi-bin/RNAz/RNAz.cgi). Default parameters were used with the following modifications: the maximum fraction of gaps: 0, the minimum mean pairwise identity: 80, the target mean pairwise identity: 80 and the P value was set to 0.99 to increase the stringency of the sRNA SNP detection. The output was further filtered to include those sRNAs with a sequence similarity < 100% and a *p*-value > 0.9995. These parameters were selected to ensure that the sRNAs investigated are highly likely to be confirmed sRNAs and have significant differences to identify changes among the MTBC lineages. The sRNAs were further filtered to include those that contained lineage-specific mutations (Gruber *et al.*, 2007).

2.2.1.3 Target prediction

The sRNA sequences were uploaded to IntaRNA (v 4.9.2) (http://rna.informatik.uni-freiburg.de/IntaRNA/Input.jsp) using default parameters with the following modification: Target NCBI RefSeq ID: NC_000962. The sRNA target outputs were compared to detect significant differences among the predicted targets. The sRNAs containing MTBC lineage-specific mutations and different putative targets were selected. Hypothetical protein targets were excluded and remaining sRNAs and subsequent targets were listed (Table 2.2) (Mann *et al.*, 2017). The subsequent sRNAs were subjected to the RNAfold webserver (v 2.4.18) to identify their predicted structures. The predicted structures were coloured to denote the probability of the bases being paired with red being the highest probability (1) and violet being the lowest (0) (Gruber *et al.*, 2008).

2.2.1.4 Functional enrichment

The remaining sRNAs were explored using UniProt ((https://www.uniprot.org/), to view protein functional information (UniProt, 2019); Quickgo (https://www.ebi.ac.uk/QuickGO/) to view and filter gene ontology annotations (Ashburner *et al.*, 2000); PANTHER (http://www.pantherdb.org/) to further classify proteins by families and subfamilies, molecular function, biological processes and pathways (Mi *et al.*, 2019); MTB Network Portal (v2) (http://networks.systemsbiology.net/mtb/?tour=bicluster_exploration&step=1) to view integrated, predicted gene regulatory networks and host pathogen interactions (Turkarslan *et al.*,

2015), AMIGO2 (v 2.5.15) (http://amigo.geneontology.org/amigo/landing) to search gene ontology data for annotation, terms and gene products (Carbon *et al.*, 2009) and BioCyc (v 25.0) (https://biocyc.org/) to explore regulatory and metabolic networks and gene essentiality (Karp *et al.*, 2019).

2.2.2 *In vitro* RNA sequencing analysis

2.2.2.1 Ethical clearance

The study was approved by the University of KwaZulu-Natal Biomedical Research Ethics Committee (BREC/00001272/2020) (Figure S1).

2.2.2.2 RNA isolation

The F15/LAM4/KZN and Beijing strains, previously isolated from patients (Gandhi *et al.*, 2006; Chihota *et al.*, 2012) were selected for sRNA and mRNA sequencing. Clinical strains belonging to the F11 and Unique families were selected for quantitative real-time PCR analysis (RT-qPCR). The strains have been isolated from clinical samples from patients in KwaZulu-Natal and characterized at the University of KwaZulu-Natal (Medical School). The virulent control being used is the laboratory strain *M. tb* H37Rv (ATCC 27294). All strains were subjected to drug susceptibility testing to confirm drug susceptibility profiles. The H37Rv, Beijing and F15/LAM4/KZN strains were subjected to restriction fragment length polymorphism (RFLP) testing to confirm strain families. Bacterial culture was collected during the stationary phase (OD: 0.8-1) and centrifuged for 1 minute and 30 seconds at 10 000 g. The supernatant was discarded and 1 mL of TRI reagent (Zymo Research, United States of America) was added and the samples were kept on ice. One millilitre of solution was added to an O-ring tube containing 300 µL of zirconia beads of 0.1 mm diameter. The solution was vortexed with a bead beater (Precellys 24, USA) at 10 °C for 30 second cycles and kept on ice for 2 minutes and repeated three times. The Zymogen DirectZol RNA Miniprep kit (Zymo Research, USA) was used to enrich for *sRNAs* following the manufacturer's instructions. The RNA was quantified by the Nanodrop 2000c Spectrophotometer (ThermoScientific, USA) and the quality and integrity were evaluated by 3-(N-morpholino) propanesulfonic acid (MOPS) (Sigma-Aldrich, USA) gel electrophoresis (Sambrook & Russell, 2001)and Bioanalzyer (Admera, USA). The RNA isolated from laboratory H37Rv reference strain was used as a control. The RNA was stored at -80°C (Sambrook & Russell, 2001).

2.2.2.3 Library preparation and RNA sequencing

The RNA-Seq library for the 3 biological replicates was prepared for the Illumina HiSeq 2000 sequencing platform using the NEBNext Ultra II with RiboZero Plus kit with a sequencing

depth of 60 million total pair-end reads per sample for mRNA and the TruSeq smRNA kit with a sequencing depth of 20 M pair- end reads per sample for sRNA following the manufacturer's instructions. In summary, 200 ng of total RNA in 50 µL was mixed with equal volumes of RNA purification magnetic beads. The resulting fragmented RNA was converted to double stranded cDNA that was repaired, tailed, and ligated with indexed adaptors. The subsequent adaptor linked cDNA library was amplified by PCR, purified and electrophoresis was performed on an Agilent high sensitivity DNA chip for quality control. Each sample lane for the 100 bp sequencing was pooled in equimolar concentration and sequenced for single reads of 100 cycles.

2.2.2.4 Read quality analysis and trimming

The quality of the reads was assessed by *FastQC* (v. 0.11.9)(Andrews, 2010) and reads were trimmed and processed using *Trimmomatic* (v. 0.39)(Bolger *et al.*, 2014) and *TrimGalore* (v. 0.6.5) to remove adapters and low-quality bases (Krueger, 2015).

2.2.2.5 Mapping and transcript assembly

The *HiSat2* (v. 2.1.1.) software was used to map sequence reads against the *M. tb* H37Rv reference genome (ASM19595v2) with alignment statistics depicted in Table S4 and S5 (Kim *et al.*, 2015). *Samtools* (v. 1.13) was used to sort and convert the mapped files from the sam format to the bam format (H. Li *et al.*, 2009). *StringTie* (v. 2.1.4) was used to assemble and quantify the transcripts (Pertea *et al.*, 2015). The assembled transcripts were then merged to allow for partially covered reads to be completed for further analysis of consistent transcripts among the different strains. The merged transcripts were then re-entered into *StringTie* so that transcript abundances can be re-estimated (Pertea *et al.*, 2016). The gffcompare function on *StringTie* was then used to determine the number of transcripts which match annotated genes, either partially or completely and the number of novel transcripts (Pertea *et al.*, 2016).

2.2.2.6 Ballgown analysis and data visualization by Integrative Genomics Viewer

Ballgown (v. 2.24.0) was used to calculate differential expression and fold changes among the clinical and reference strains (Fu *et al.*, 2021). Global statistics of the data as well as transcript abundance between the strains was visualized using the *Ballgown* software within R:Bioconductor (v. 4.1.1.) analysis interphase (Smyth, 2005). In the R workspace, each strain was compared with the H37Rv control strain and filtered (Pertea *et al.*, 2016) while the heat maps to display fold changes were plotted using Multiple Experiment Viewer (MeV) (Howe *et al.*, 2011).

2.2.2.7 Enrichment and functional analysis

Enrichment and functional classification of the differentially expressed transcripts with a greater than 2-fold increase among the clinical strains and H37Rv was performed using the MTB

Network Portal (v2) (http://networks.systemsbiology.net/mtb/?tour=bicluster_exploration&step=1) (Turkarslan *et al.*, 2015), UniProt ((https://www.uniprot.org/) (UniProt, 2019) and BioCyc (v 25.0).

2.2.2.8 sRNA confirmation using quantitative real time PCR (RT-qPCR)

Qualitative confirmation of the presence of the identified sRNA in F15/LAM4/KZN and Beijing strains, with other clinical strains belonging to F11 and Unique strain families, was carried out through quantitative real time PCR (RT-qPCR) using the H37Rv strain as a virulent control. Extracted RNA was standardized to 1 µg and converted to cDNA using high capacity cDNA Reverse Transcription kit (ThermoScientific, USA). FASTA files for sRNAs were extracted from the small RNA sequencing data and used to synthesize transcript-specific primers (Supplementary Table S6) using Primer Blast (Ye *et al.*, 2012), followed by optimization through gradient PCR. Primer efficiencies were calculated (Supplementary Table S7 and S8) and RT-qPCR amplification and transcript-specific melt curves were observed in the CFX-96 (Biorad, South Africa) (Supplementary Figures S5-S20). Data was presented as fold changes between clinical strains and the laboratory strain, H37Rv using a $2^{-\Delta\Delta Ct}$ relative quantification method.

2.3 Results and Discussion

2.3.1 Prediction of potential sRNAs with lineage-specific sRNAs mutations

The mapped and completed whole genomes of the respective sequences belonging to the eight MTBC lineages were aligned, and the resulting multiple sequence alignment file was used to identify putative sRNAs using RNAz. After filtering the putative sRNAs, SNP mutations were detected in 438 predicted sRNAs. Twenty-eight sRNAs containing the following characteristics: lineage-specific mutations; different consensus and mutated predicted targets; and non-hypothetical targets, were identified and investigated further (Table 2.2). Each sRNA included a consensus target, that was identified using the sRNA sequence common to all strains except for the lineage of interest. The mutated target was detected using the mutated sequence present in all strains belonging to one lineage. Both the consensus and variant targets varied in *p*-value (1.00e-07 - 0.00181) and energy of the secondary structure (-23.72 - -14.61), that is the way in which the RNA folds into a specific conformation.

Table 2.2: sRNAs containing lineage specific mutations and the corresponding IntaRNA consensus and mutated targets and *p*-values.

Name of sRNA	Mutated lineage	Consensus IntaRNA target	Gene	*p*-value	Mutated IntaRNA target	Gene	*p*-value
38649	L1	Rv3102c	*ftsE*	4.43e-05	Rv2332	*mez*	4.6e-06
16881	L2	Rv1336	*cysM*	1.01e-05	Rv0315		4.3e-06
16883	L2	Rv0575c		1.6e-05	Rv0456c	*echA2*	0.0002268
29983	L2	Rv2776c		0.0003423	Rv0383c		0.0003069
15864	L3	Rv1473A		7.07e-05	Rv0260c		0.0001408
12070	REF, L4	Rv0239	*vapB24*	8.52e-05	Rv0636	*hadB*	2.58e-05
26139	REF, L4	Rv0604	*IpqO*	1.00e-07	Rv0527	*ccdA*	1.53e-05
11617	L5	Rv2155c	*murD*	8.49e-05	Rv0280	*PPE3*	0.0001844
11618	L5	Rv0242c	*fabG4*	2.88e-05	Rv1722		4.2e-06
14874	L5	Rv2436	*rbsK*	0.0005826	Rv0514		0.0011203
14908	L5	Rv2192c	*trpD*	6.3e-06	Rv2940c	*mas*	7.3e-06
34611	L5	Rv2989		7.32e-05	Rv1677	*dsbF*	0.0002423
5145	L6	Rv1143	*mcr*	0.0004084	Rv0083		0.0005395
19358	L6	Rv2277c		0.0003157	Rv1064c	*IpqV*	0.0009681
26173	L6	Rv2457c	*clpx*	3.1e-06	Rv3087		5.3e-06
1080	L7	Rv1862	*adhA*	0.0016401	Rv1085c		0.0002485
1082	L7	Rv2332	*mez*	2.35e-05	Rv3828c		1.07e-05
1224	L7	Rv0666		2.09e-05	Rv1391	*dfp*	2.02e-05
23747	L7	Rv0806c	*cpsY*	0.0007797	Rv0771		0.0002927
24026	L7	Rv2775		1.54e-05	Rv0858c	*dapC*	8.8e-06
38691	L7	Rv1001	*arcA*	4.35e-05	Rv2732c		1.02e-05

Name	Mutated lineage	Consensus IntaRNA target	Gene	*p*-value	Mutated IntaRNA target	Gene	*p*-value
37867	L7	Rv3270	*ctpC*	0.0010524	Rv1365c	*rsfA*	0.000704
11527	L8	Rv1937		1.22e-05	Rv2399c	*cysT*	1,00E-05
21006	L8	Rv3019c	*esxR*	3.39e-05	Rv0737		1.02e-05
21945	L8	Rv2202c	*adoK*	9.36e-05	Rv1381	*pyrC*	3.3e-06
27226	L8	Rv1332		0.0001216	Rv1124	*ephC*	0.0002727
27232	L8	Rv0194		2.09e-05	Rv2946c	*pks1*	0.0001115
35296	L8	Rv2535c	*pepQ*	0.0001801	Rv0734	*mapA*	0.00181

2.3.1.1 Lineage 1 and 3

Lineage 1 (Figure 2.1) included only one sRNA with a varying consensus and mutated target, *ftsE* and *mez*, respectively. The *ftsE* is characterised as a cell division ATP-binding protein essential for the growth of H37Rv while *mez* (malic enzyme) encodes a NAD-dependent malate oxidoreductase (Tyagi *et al.*, 1996; Basu *et al.*, 2018). It has been identified that *ftsE* is not an essential gene required for optimal growth and may not exhibit a stress response function in *M. tb* (Roy *et al.*, 2011). Basu *et al.* (2018) demonstrated that the malic enzyme is one of the four enzymes forming part of the anaplerotic (ANA) node (Basu *et al.*, 2018). This node connects the main pathways of central metabolism which are glycolysis, gluconeogenesis and the tricarboxylic acid cycle (Basu *et al.*, 2018). The ANA node is essential to the intracellular growth of *M. tb*. In addition to its role in central metabolism, *mez*, is involved in lipid biosynthesis. The knockout *mez* strains formed cell walls that were altered and had decreased efficiency following entry into macrophages (Basu *et al.*, 2018). A clear structural change could be seen in the sRNA when mutated as compared to the consensus sequence. The mutated sequence causes the structure to be closed, thereby blocking potential biding sites. *mez* is also the consensus target for sRNA 1082 (lineage 7) with Rv3828c as the variant target. This mRNA encodes a possible resolvase, preventing the cointegration of unknown DNA prior to integration into the chromosome (Minato *et al.*, 2019). Devasundaram *et al.* (2015) reported a fold change increase of 3.18 and 1.72 in Rv3823c in the H37Rv strain and a clinical strain belonging to the S7 family within the Euro-American lineage (Lineage 4) respectively, grown in an oxygen-deficient environment. The resulting upregulation suggests that this gene plays a regulatory role in the adaptation of the

organism to an unfavourable environment (Devasundaram *et al.*, 2015). With sRNA 1082 favouring Rv3823c as a target, it may be possible that this sRNA may also play an important role within lineage 7 assisting in hypoxic environments. Lineage 3 also contains only one sRNA, 15864, with the consensus and variant targets identified to be Rv1473A and Rv0260c (DeJesus *et al.*, 2017). Rv1473 is a novel ATP binding cassette efflux pump that is associated with antibiotic resistance, in particular macrolides, via the efflux mechanism Duan *et al.* (2019). Rv0260c encodes a transcriptional regulatory protein/response regulator (Gautam *et al.*, 2019). The interaction of the consensus sRNA with the Rv1473 target may lead to changes in the antibiotic resistance of the strains. This could be a potential upregulation or downregulation depending on the position and conformation change resulting from the interaction. As seen in the structural figure in Supplementary Figure S3, the mutation does not have a drastic effect on the structure of the sRNA, suggesting that the consensus target is the more likely candidate.

	Consensus		**Mutated**	
Name of sRNA	**Target**	**Gene**	**Target**	**Gene**
38649	Rv3102c	*ftsE*	Rv2332	*mez*

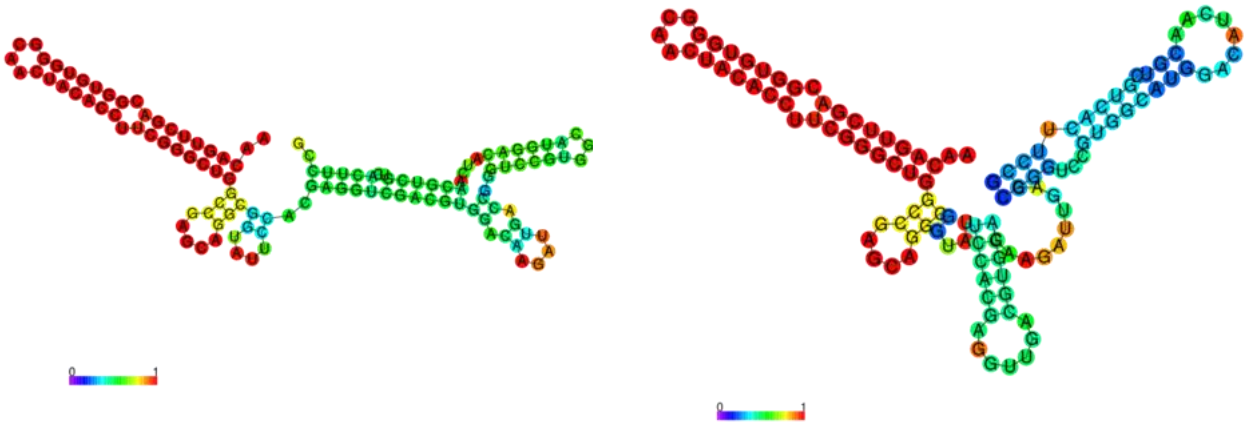


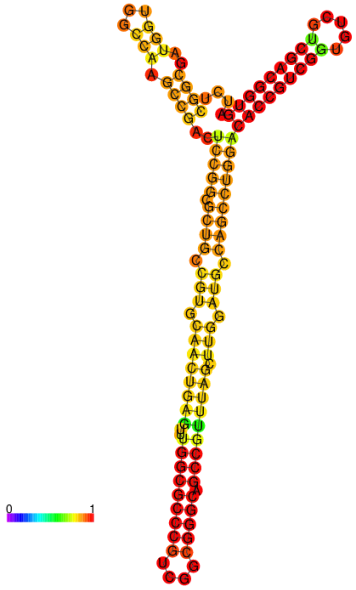
Figure 2.1: sRNAs containing lineage specific mutations and the corresponding IntaRNA consensus and mutated targets and structure predictions for lineage 1. A significant structural and base-pairing probability change was observed for sRNA 38649.

2.3.1.2 Lineage 2

Three sRNA, namely 16881, 16883 and 29983, were identified for lineage 2 (Figure 2.2 and Supplementary Figure S3). The consensus target for 16881 is *cysM*, encoding cysteine

synthase B, a requirement for cysteine biosynthesis, whereas the variant target encodes an immunostimulatory antigen. Both have been identified as non-essential for growth of the H37Rv strain, however, the disruption of these genes proves to be advantageous for the bacterium (Minato *et al.*, 2019). Burns-Huang and Mundhra (2019) showed that the *mec*+-*cysO*-*cysM* gene cluster formed part of a novel cysteine biosynthesis pathway, though was not identified as essential. This gene cluster however, was found to be essential for the resistance to the drug clofazimine suggesting that it may play a defence role, particularly during oxidative stress (Burns-Huang & Mundhra, 2019). Rv0315 showed no hydrolytic activity and has become an inactive β-1,3-glucanase and its role in virulence and pathogenesis needs to be explored further (W. Dong *et al.*, 2015). Due to the higher probability of base pairing in the consensus sRNA structure, the mutated sRNA may be more likely to interact with its target. The association between mutated sRNA and target may restore hydrolytic activity as well as glucanase activity. The sRNA 16883 targets a possible oxidoreductase, however the lineage 2 strains target the *echA2* transcript which encodes an enoyl-CoA hydratase which is also non-essential for growth. Rv0575c was found to have a 3.05 fold increase in expression in low-oxygen conditions, suggesting an environmental adaptation role (Bacon *et al.*, 2004). Due to the high base-pairing probabilities of the residues, it is unlikely that this sRNA will have an effect on either variant or consensus target. The sRNA 29983, targets a probable oxidoreductase with an unknown function, that is non-essential and a TMM transport factor A when the consensus and variant sRNAs are present (Fay *et al.*, 2019; Minato *et al.*, 2019). Rv0383c encodes a protein that forms part of the MmpL3 (mycobacterial membrane protein, large) transport machinery. This machinery is responsible for the transport of mycolic acid to the cell wall. This gene has been found to be essential for the growth of the bacterium and its product is anchored to the cytoplasmic membrane thereby interacting with the MmpL3 (Fay *et al.*, 2019). Mycolic acids play an important role in defence of the pathogen by making the bacterium less susceptible to antibiotics through its integration into the cell wall (Zhao *et al.*, 2015). The targeting of a component of the mycolic acid transport system by the variant may allow for faster transport of the molecule, therefore, a higher concentration of mycolic acids present in the cell wall. This sRNA may confer a selective advantage to the Beijing strain, which forms part of Lineage 2, as well as being at the forefront of drug resistance development (Ebrahimi-Rad *et al.*, 2003). No significant changes were observed for sRNA 15864 mutations identified for lineage 3 (Supplementary Figure S3).

	Consensus		Mutated	
Name of sRNA	**Target**	**Gene**	**Target**	**Gene**
16881	Rv1336	*cysM*	Rv0315	



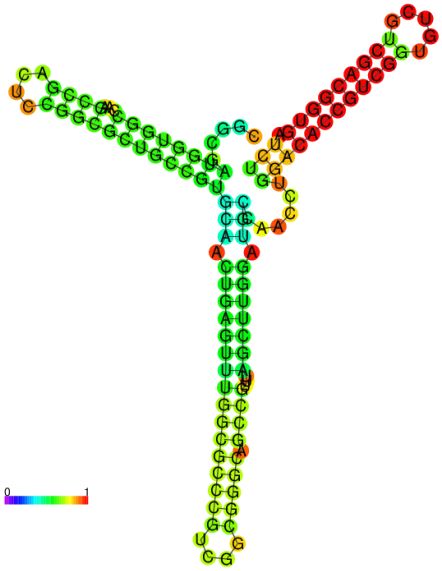
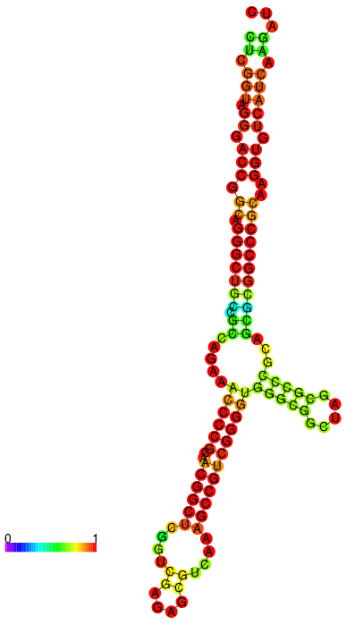
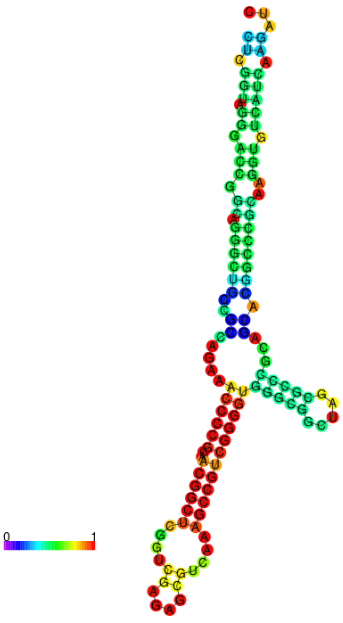


Figure 2.2: sRNAs containing lineage specific mutations and the corresponding IntaRNA consensus and mutated targets and structure predictions for lineage 2. sRNA 16881 was the only sRNA to exhibiting both structural and base-pairing changes.

2.3.1.3 Lineage 4

Due to the reference H37Rv strain belonging to lineage 4, only sRNAs, which include 12070 and 26139 belonging to the selected lineage 4 strains and H37Rv, are presented (Figure 2.3). The consensus target for 12070 is Rv0239, which encodes the possible antitoxin VapB24, while the mutated target encodes the (3R)-hydroxylacyl-ACP dehydratase subunit HadB. The VapB24 target had been identified as an antitoxin by Solano-Gutierrez *et al.* (2019) and has been shown to interact with two toxins namely MazF3 and VapC25 (Zhu *et al.*, 2010). The amino acid sequence of this protein was identical in all lineages except lineage 7, whereby a stop codon was inserted in amino acid 10 (Solano-Gutierrez *et al.*, 2019). The gene *hadB* encodes a protein that is involved in type II fatty acid synthesis and is essential for growth (Brown *et al.*, 2007; Sacco

et al., 2007). A disruption in the expression of a toxin-antitoxin system can prove detrimental to a bacterium as they play a stabilizing role for the organism's chromosomes. With sRNA 12070 preferentially binding to the hadB mRNA, instead of the VapB24 mRNA, the resulting over or under expression of the gene may have a direct effect on the pathogenicity of the lineage. The sRNA, 26139, lineage 4, targets a probable conserved lipoprotein LpqO and a possible cytochrome C-type. Due to the gaps in research, the exact function of LpqO is yet to be uncovered. However, it was discovered that the gene responsible is important for survival and viability of the bacterium (Gautam *et al.*, 2019). Rv0527 is a member of the CcdA family of electron transporters and provides electrons during the reduction of c-type cytochromes, contributing to the type II cytochrome- maturation system (Goldstone *et al.*, 2016). Han and Wilson (2013) investigated the effects of two copies of the *ccdA* gene in *Bacillus anthracis* and discovered that the loss of both genes resulted in reduced cytochrome c production, increased virulence factor expression and reduced sporulation efficiency. It is possible that sRNA 26139 may bind to the Rv0527 mRNA within lineage 4, thereby resulting in the transcript being non-functional and a similar result may occur by increasing the virulence factor expression.

	Consensus		**Mutated**	
Name of sRNA	**Target**	**Gene**	**Target**	**Gene**
12070	Rv0239	*vapB24*	Rv0636	*hadB*
				

26139	Rv0604	*IpqO*	Rv0527	*ccdA*

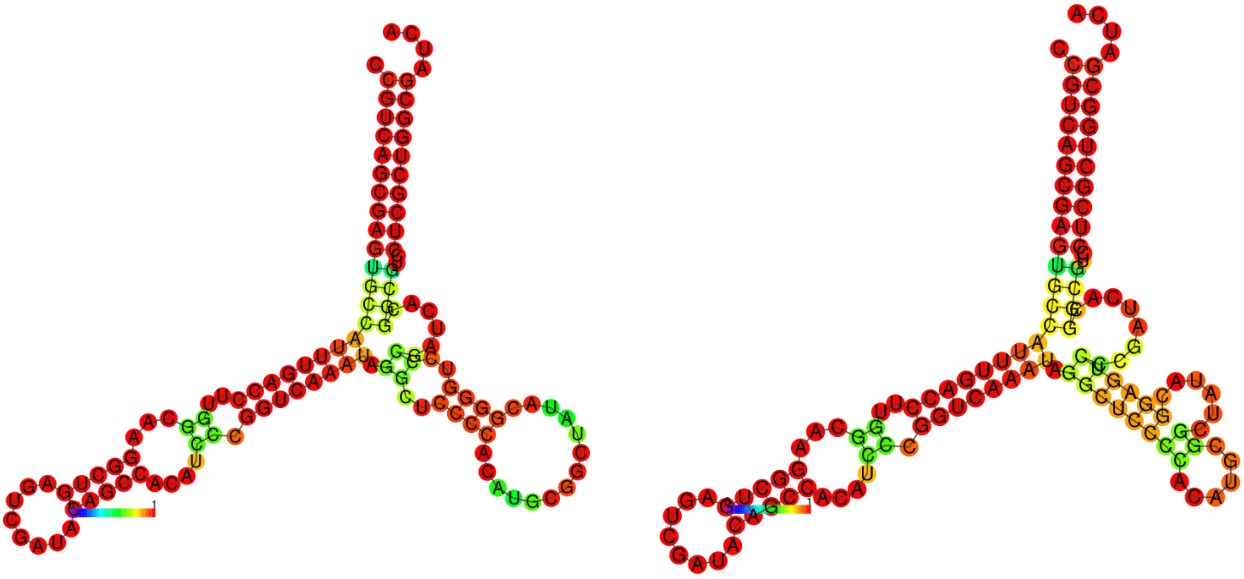


Figure 2.3: sRNAs containing lineage specific mutations and the corresponding IntaRNA consensus and mutated targets and structure predictions for lineage 4. No significant changes were observed for sRNA 26139 however, sRNA 12070 had a slight base-pair probability change decreasing the likelihood of the sRNA to be bound to itself.

2.3.1.4 Lineage 5

Lineage 5 presented five sRNAs, that is, 11617, 11618, 14874, 14908 and 34611, that met the respective criteria (Figure 2.4 and Supplementary Figure S3). Consensus target *murD*, encodes the UDP-N-acetylmuramoylalanine-D-glutamate ligase, which is essential for growth and is involved in peptidoglycan biosynthesis and cell wall formation (Thakur & Chakraborti, 2008). The opposing mutated target for sRNA 11617 is *PPE3*, a gene encoding the PPE family protein PPE3, with this protein found to be specifically expressed in MDR and XDR strains (Ullah *et al.*, 2021). Sun *et al.* (2019) found that Rv2155c was one of the six genes that contained polymorphisms after ethambutol exposure. This suggests that the *murD* gene may play a role in antibiotic resistance within lineages and sRNA 11617 may play an activation or repressive role in this case. From the secondary structure prediction, the base change had no significant effect on the sRNA structure, therefore it is highly likely that this sRNA would preferentially target the transcript encoded for *murD* in both the consensus and mutated form as it is an essential gene. Similar results may be applicable to the sRNA 14874, as no change in base pairing probability or structural change were observed between the consensus and variant target. This sRNA is most likely to target the consensus target, which is Rv2436, encoding a ribokinase which catalyzes the first step of ribose phosphorylation (Yang *et al.*, 2011). A similar effect can be seen in the

consensus and mutated sRNAs 11618 and 14908. However, the upper node of the sRNA has a higher binding probability in the mutated form than the consensus form in the case of 11618. The right node of the sRNA 14908 has a higher base pairing probability in the consensus form than the mutated form. Due to these changes, it is highly probable that sRNA 11618 will target the consensus target, *fabG4* transcript encoding a NADH-dependent β-ketoacyl CoA reductase, involved in the fatty acid biosynthesis pathway. The mutated target is a possible carboxylase potentially involved with lipid metabolism that has yet to be experimentally validated (Banerjee *et al.*, 2015). The targets for 14908 are involved in tryptophan biosynthesis and lipid metabolism. The consensus target transcript encodes an anthranilate phosphoribosyl transferase and the mutated transcript encodes a mycocerosic acid synthase (Lee *et al.*, 2006; Herbst *et al.*, 2016). The *trypD* gene has been found to be essential to cause disease in mice and is therefore, required for the progression of TB infection (Lott, 2020). It is, therefore, likely that the consensus target, *trypD*, will be the preferred target for both the consensus and variant sRNA. It may be possible that the degree of control that the sRNA has on the transcript may cause an increase in the TB disease progression. The sRNA 34611, targets a member of the lclR transcriptional factors, whereas the mutated sRNA targets the gene encoding a putative disulphide bond isomerase (Chim *et al.*, 2010; Q. Li *et al.*, 2016). There was a major change present in the base-pairing probabilities between the consensus and mutated sRNAs. The blue residues depict a much lower base-paring probability, close to 0, indicating that the left half of the sRNA may be single-stranded, allowing for the target mRNA to attach fully. The mutated sRNA in lineage 5 is enclosed and lacking a large attachment area, which may lead to very little to no regulatory role of the sRNA on Rv0280 transcript.

2.3.1.5 Lineage 6

Within lineage 6, only three sRNAs were identified (Figure 2.5 and Supplementary Figure S3). The consensus target for sRNA 5145 targets the gene encoding an alpha-methylacyl-CoA racemase required for the catabolism of branched-chain fatty acids and bile acid synthesis. The mutated counterpart of this sRNA targets a probable oxidoreductase that still needs to be further studied (Savolainen *et al.*, 2005). Base-pairing probabilities were only affected slightly with the consensus sequence having a higher base-pair probability than the mutated structure. Lu *et al.* (2015) discovered that *mcr* encodes a protein that facilitates catabolization of cholesterol esters which may be activated during times of cholesterol scarcity. The regulation of this transcript by sRNA 5145 has the potential to provide other lineages a selective advantage over lineage 6, particularly in a stressed environment. A glycerophosphoryl diester phosphodiesterase and a lipoprotein LpqV are the respective consensus and variant targets for sRNA 19358 (Tsolaki *et al.*, 2004). The significant structural and base-pair probability differences observed between the

consensus and mutated sRNA could result in improved binding of the consensus sequence to Rv2277c, leading to regulation of this gene. The sRNA 26173 mutation caused major structural and base-pair probability changes. The consensus target is *clpX*, which encodes the ATP-dependent Clp protease ATP-binding subunit (Ollinger *et al.*, 2012). The Clp protease, in which the subunit is involved, is a major protease playing a role in maintaining and ensuring functionality of all proteins, particularly under stress. This can contribute to increased virulence within the organism (Frees *et al.*, 2003). The variant target is a possible triacylglycerol synthase that was upregulated within the hypoxic environment in macrophages, when compared to growth on laboratory media (Rajaram *et al.*, 2011). The open structure and low base-pairing probability of the consensus sequences suggests that it is highly likely that sRNA 26173 binds to the ClpX mRNA. This may offer increased activity and efficiency of Clp protease, thus conferring an increase in subsequent virulence in other lineages except lineage 6.

	Consensus		**Mutated**	
Name of sRNA	**Target**	**Gene**	**Target**	**Gene**
34611	Rv2989		Rv0280	*PPE3*

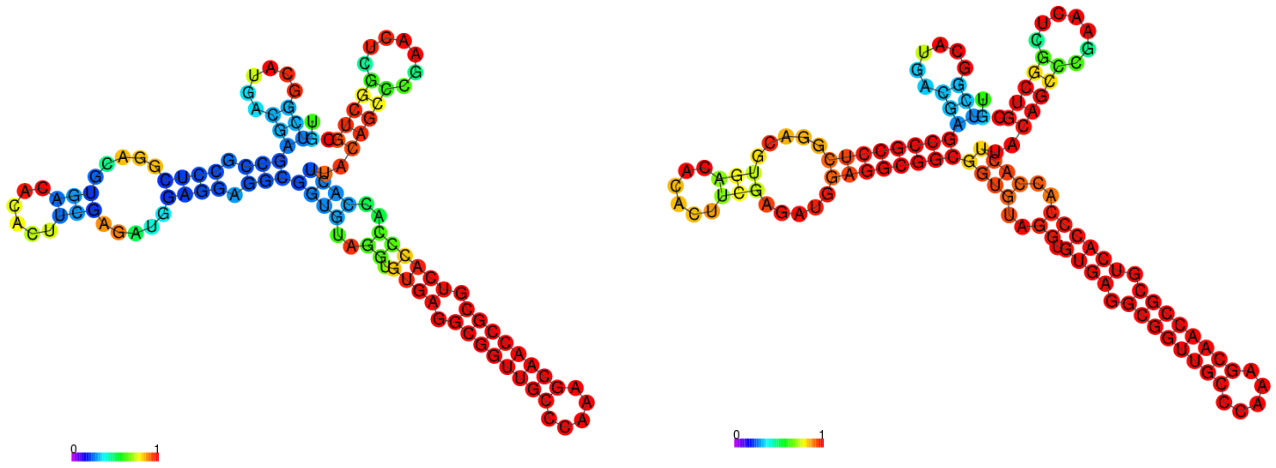


Figure 2.4: sRNAs containing lineage specific mutations and the corresponding IntaRNA consensus and mutated targets and structure predictions for lineage 5. Only one sRNA identified in lineage 5 showed a significant difference in their base-pairing probability, being sRNA 34611. The mutation resulted in an increase in the probability of the sRNA to become a closed structure.

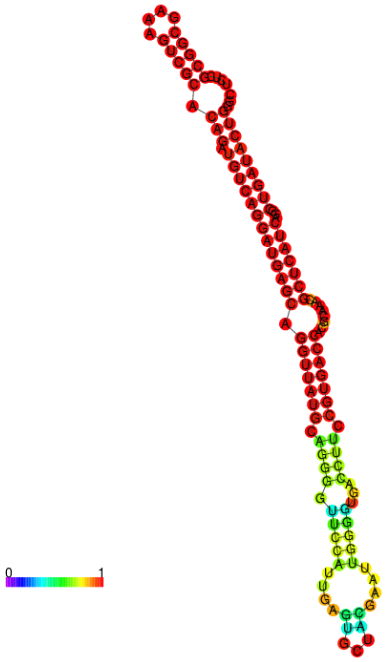
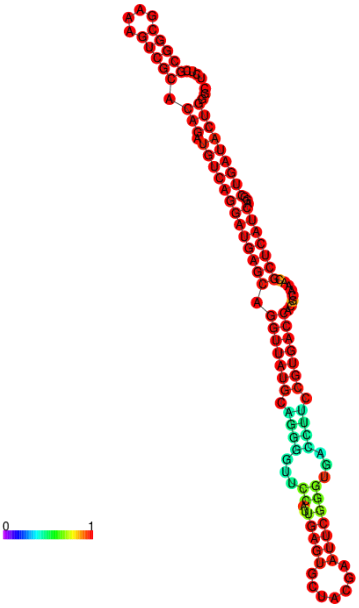
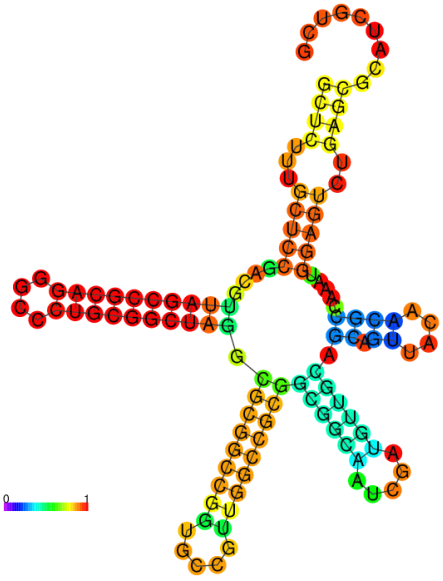
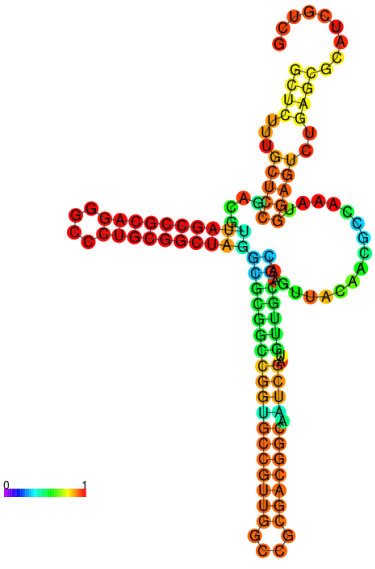
	Consensus		Mutated	
Name of sRNA	**Target**	**Gene**	**Target**	**Gene**
19358	Rv2277c		Rv1064c	*IpqV*
				
26173	Rv2457c	*clpx*	Rv3087	
				

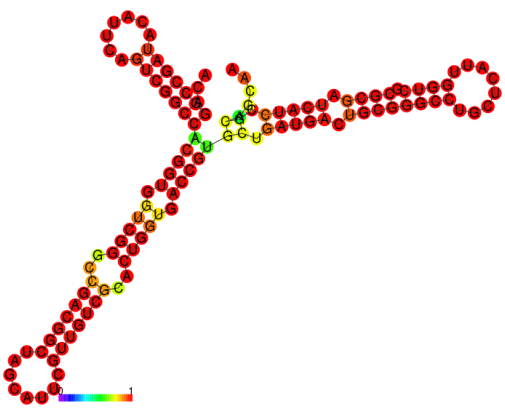
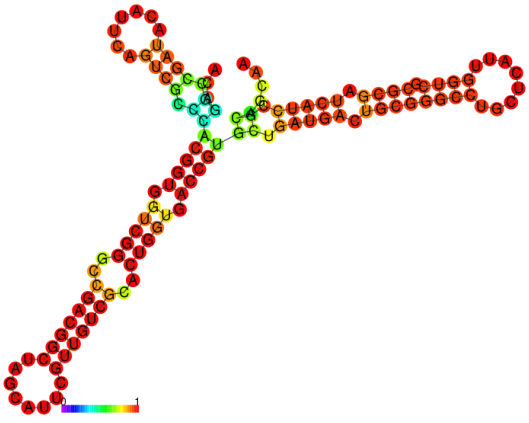
Figure 2.5: sRNAs containing lineage specific mutations and the corresponding IntaRNA

consensus and mutated targets and structure predictions for lineage 6. A significant conformational change was observed between the mutated and consensus sRNAs 26173. The other sRNAs identified in lineage 6 showed slight changes in the base-pairing probabilities.

2.3.1.6 Lineage 7

Lineage 7 exhibited nine mutated, lineage-specific sRNAs and was the most abundant lineage, in terms of identified sRNAs (Figure 2.6). The consensus sRNA target for 1080 encodes a probable alcohol dehydrogenase that catalyses the reversible oxidation of ethanol to acetaldehyde, particularly under anaerobic environments to assist with anaerobic glycolysis (Ogata *et al.*, 1999). The mutated target encodes a possible hemolysin-like protein that is supposedly involved in virulence (Deshpande *et al.*, 1997; L. Zhang *et al.*, 2020). A minor structural change was observed between the mutated and consensus sRNAs, decreasing the base-pairing probabilities of the mutated target. This allowed for more accessibility, leading to easier attachment, resulting in regulation of this hemolysin-like encoding gene. This can confer a selective advantage to lineage 7 by increasing the bacterium's virulence. The consensus and mutated target for sRNA 1224 is a possible virulence associated membrane protein, and a phosphopantothenoylcysteine decarboxylase/Phosphopantothenoylcysteine synthetase, required for pantothenate and CoA biosynthesis (Sassetti *et al.*, 2003). Evans *et al.* (2016) showed that the silencing of the *dfp* gene caused a bactericidal effect to *M. tb.* A large decrease in base-pairing caused by the mutation may confer an increased activation role for this sRNA, preventing this gene from becoming inactive. The *cpsY* gene is the target for the consensus 23747 sRNA, which encodes a possible UDP-glucose-4-epimerase, thought to be involved in the exopolysaccharide and lipopolysaccharide biosynthetic pathway (Sassetti *et al.*, 2003). Its mutated counterpart encodes a possible 4-carboxymuconolactone decarboxylase, involved in aromatic hydrocarbon metabolism (Wolfe *et al.*, 2010). Interestingly, aromatic hydrocarbon metabolism could have both a beneficial and harmful effect. While increased carbon and energy sources may be beneficial, compounds that could cause oxidative damage may also be produced (Kweon *et al.*, 2015). The bacteria from lineage 7 may be benefitted by activating of the genes that increase availability of additional carbon and energy sources or preventing the expression of the gene that would normally cause oxidative damage (Kweon *et al.*, 2015). The exact mechanism will need to be further investigated to determine the pathway that is activated. The sRNA 24026 has a consensus target which includes a GCN5 acetyltransferase while the variant target includes a N-succinyldiaminopimelate aminotransferase, which is involved in biosynthesis of diaminopimelate and lysine from aspartate semialdehyde (Nizolenko *et al.*, 2005; Weyand *et al.*, 2006). The *arcA* gene encodes an arginine deaminase allowing for arginine degradation within the bacterium and is the consensus target for sRNA 38691 (Surken *et al.*, 2008). Surken *et al.* (2008) demonstrated

that *arcA* was non-essential for the aerogenic infection in mice. The variant target identified for this sRNA encodes a membrane anchoring protein (Sreenu *et al.*, 2006). The sRNA 37867 has a consensus target that is *ctpC* and a variant target, *rsfA*. The gene *ctpC* encodes a probable metal cation-transporting P-type ATPase, involved in the transport of metal cations out of the cell conferring resistance to the same metals while *rsfA* encodes an anti-sigma factor antagonist. The *rsfA* gene also negatively regulates Rv3287c, another anti-sigma factor (Parida *et al.*, 2005; Salina *et al.*, 2018). The sRNA 37867 exhibits a major difference between the consensus and mutated structure as well as base-pair probabilities. This large difference may allow for preferential pairing of the mutated sRNA, eliciting control of the anti-sigma factor antagonist. This will in turn regulate the expression of sigma factor SigF when exposed to stress (Parida *et al.*, 2005), thus assisting in adaptability of the bacterium to changing environments.

	Consensus		**Mutated**	
Name of sRNA	**Target**	**Gene**	**Target**	**Gene**
1080	Rv1862	*adhA*	Rv1085c	
				
1082	Rv2332	*mez*	Rv3828c	

1224	Rv0666		Rv1391	*dfp*
23747	Rv0806c	*cpsY*	Rv0771	

24026	Rv2775		Rv0858c	*dapC*
38691	Rv1001	*arcA*	Rv2732c	

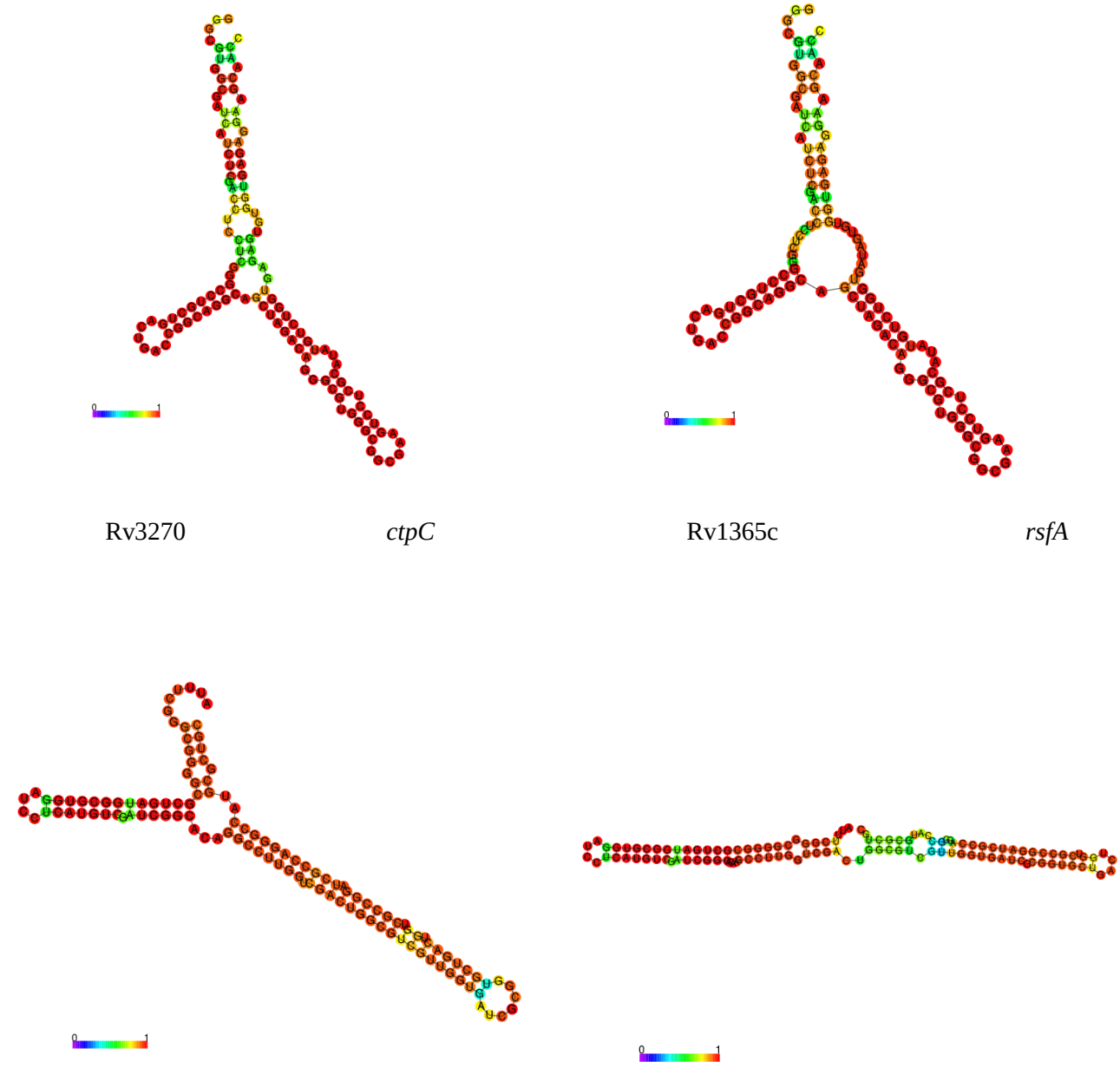
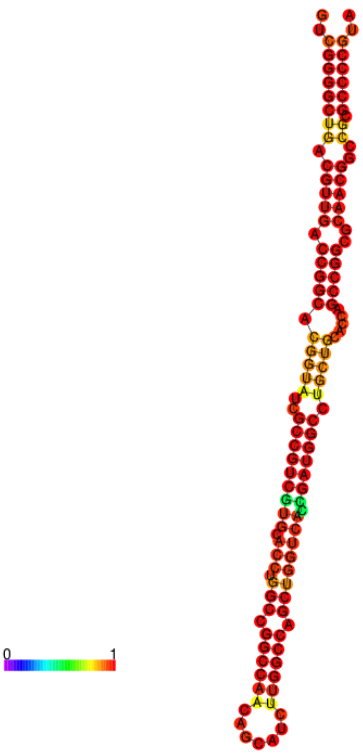


Figure 2.6: sRNAs containing lineage specific mutations and the corresponding IntaRNA consensus and mutated targets and structure predictions for lineage 7. The only significant change observed in lineage 7 was the structural change observed within sRNA 37867, causing the mutated sRNA to become linear. The other sRNAs within the lineage discovered slight base-pair probability changes when mutated.

2.3.1.7 Lineage 8

The most recent lineage 8 included in the MTBC demonstrated a total of six sRNAs (Figure 2.7 and Supplementary Figure S3). The sRNA, 11527 targets a possible monooxygenase while the mutated version targets the transcript encoding a sulphate transport system permease

protein. Rv1937 has been previously identified as a target for another sRNA, ncRv11875C, that showed increased expression in iron-limiting conditions, which proved positive regulation by ncR11875C (Ami *et al.*, 2020). It is possible that the other lineages within the MTBC may have a selective advantage in iron-limiting advantages in contrast to lineage 8. The gene encoding the secreted ESAT-6 like protein EsxR is the consensus target for sRNA 21006, while the mutated sRNA targets the gene encoding a transcriptional regulator from the MarR family (Arbing *et al.*, 2010; Gong *et al.*, 2019). The EsxR protein forms part of a complex that plays a mediatory role in the interactions between bacterium and host. Due to the lower base-pair probability found in the consensus strain, it may be possible that the association between the consensus sRNA and target may grant the other lineages apart from lineage 8, an advantage for survival within the host due to the EsxR protein. The sRNAs 21945, 27232 and 35296 all have very slight changes in their structure and base-pair probability, which may result in very small, insignificant changes between the targets for the consensus and mutated sRNAs. The sRNA 21945 has a consensus target involved in adenosine phosphorylation while the variant is involved in pyrimidine biosynthesis. The former target encodes adenosine kinase, while the latter encodes a probable dihyroorotase (Long *et al.*, 2003; Krawczyk *et al.*, 2009). Rv0194 encodes a transmembrane multidrug efflux pump, involved in the active transport of drugs across the membrane and was found to be partially or completely deleted in several clinical isolates (Tsolaki *et al.*, 2004). It is a nonessential gene, the disruption of which results in increased bacterial growth, and it is the consensus target for sRNA 27232 (DeJesus *et al.*, 2017). The variant target was identified to be *pks1* which encodes a polyketide synthase involved in lipid synthesis (Pang *et al.*, 2012). The genes *pepQ* and *mapA* are the consensus and variant targets for sRNA 35296. The *pepQ* encodes a cytoplasmic peptidase while *mapA* encodes methionine aminopeptidase, both included in the intermediary metabolism and respiration functional category (X. Zhang *et al.*, 2009; Yew *et al.*, 2017). The sRNA 27226 targets a transcription factor in all lineages except lineage 8, where it targets the gene encoding epoxide hydrolase, an enzyme involved in detoxification reactions during oxidative damage to lipids (Schulz *et al.*, 2020). The transcription factor also plays a part in the removal of toxic compounds created during oxidative stress (Khan *et al.*, 2021). There is a significant difference between the consensus and variant sRNAs. Even though both targets play a role in correcting oxidative damage, it is possible that one pathway may be more efficient that the other, resulting in a selective advantage.

	Consensus		Mutated	
Name of sRNA	**Target**	**Gene**	**Target**	**Gene**
21945	Rv2202c	*adoK*	Rv1381	*pyrC*
				
27226	Rv1332		Rv1124	*ephC*

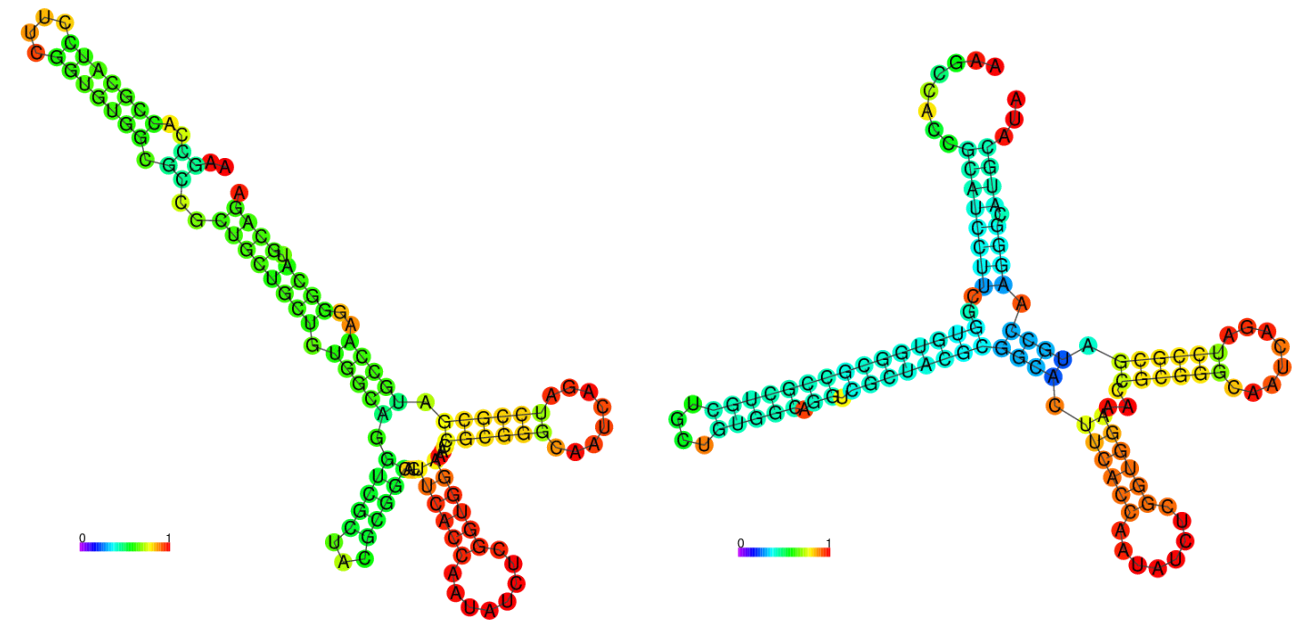


Figure 2.7: sRNAs containing lineage specific mutations and the corresponding IntaRNA consensus and mutated targets and structure predictions for lineage 8. A significant change observed in lineage 8 was between the mutated and consensus structure and base-pairing probability of sRNA 27226. Slight base-pair probability changes were observed for the other sRNAs in the lineage when mutated.

2.3.2 Small RNA sequencing and RT-qPCR reveals novel transcripts implicated in growth of Beijing, F15/LAM4/KZN, F11 and Unique clinical strains of *M. tb*

The drug susceptibility profiles of the clinical strains revealed the Beijing strain as MDR and the F15/LAM4/KZN strain as XDR while the remaining clinical strains were susceptible (Supplementary Table S1 and S2). The RFLP results presented in Supplementary Figure S2 confirmed the families the clinical strains belonged to. The extracted RNA remained intact, as seen in Supplementary Figure S4, with concentration ranging from 535,2 ng/µL- 2255,1 ng/µL (Supplementary Table S3). The sRNA and mRNA sequences were mapped, assembled and annotated (Supplementary File 1 and 2; Supplementary Table S4 and S5).

The RNA sequencing data was filtered to include transcripts with a p-value less than 0.05, a fold change greater than 1 and a sequence length less than 500bp. These parameters were selected to ensure that statistically significant sRNAs were selected with fold changes, which could have resulted through differences between the clinical strains. Due to sRNAs not being

completely characterized within the *M. tb* species, it is expected to have novel transcripts identified, which is evident in Table 2.3. MSTRG tags were assigned to those sequences that were not identified as genes by *Stringtie*. The fold changes were measured against the reference *M. tb* H37Rv strain. The proposed novel sRNAs were identified in all virulent *M. tb* (H37Rv, Beijing, F15/LAM4/KZN, Unique, F11) strains through RT-qPCR (Supplementary Table S8), however, their level of expression significantly varies between strains (Figure 2.10), which might be indicative of specific virulence traits in strain families.

The Beijing strain exhibited an approximate fold change of 3 and 2 in sRNA MSTRG.34.1 and MSTRG.40.1, respectively when compared to the H37Rv strain (Table 2.3). MSTRG.34.1 had been identified to target Rv1543 which encodes a fatty acyl-CoA reductase. This enzyme is reported to be involved in the biosynthesis of wax esters (Sirakova *et al.*, 2012). Wax esters have been shown to accumulate within *M. tb* during states of dormancy in stressful environments, resulting in an alternate form of lipid storage (Deb *et al.*, 2009). Target Rv2304c was identified as a hypothetical protein and was also identified in the F15/LAM4/KZN strain with a fold change of 1.27. The Beijing strain exhibited double the number of transcripts than the F15/LAM4/KZN strain indicating a regulatory requirement for this sRNA.

Table 2.3: sRNAs identified from small RNA sequencing from the Beijing and F15/LAM4/KZN clinical strains with a *p*-value < 0.05, a fold change > 1 and a sequence length < 500bp.

Strain	**sRNAs**	**Size**	**Fold change**	***p*-value**	**IntaRNA target**	**Function**
Beijing						
	MSTRG.34.1	276	3.34	0.0053	Rv1543, *fcr2*	fatty acyl-CoA reductase
	MSTRG.40.1	372	2.10	0.012	Rv2304c	hypothetical protein
F15/LAM4/KZN						
	MSTRG.53.1	242	2.57	0.00017	Rv2639c	Integral membrane protein
	MSTRG.26.1	234	2.27	0.013	Rv0233, *nrdB*	ribonucleoside-diphosphate reductase subunit beta
	MSTRG.40.1	372	1.27	0.028	Rv2304c	hypothetical protein

Within the F15/LAM4/KZN, two other sRNAs, namely Rv2639c and Rv0233 were identified to be significantly higher than the laboratory strain. These targets encode an integral membrane protein and the ribonucleoside-diphosphate reductase subunit beta respectively (Kapopoulou *et al.*, 2011). Since sRNA MSTRG.26.1 was identified to target *nrdB*, the potential base-pairing mechanism would cause a blockage in the transcriptional machinery, preventing in the NrdB protein from being translated. The resulting effect would cause a cascading effect on the deoxyribonucleotide biosynthetic process (Figure 2.8).

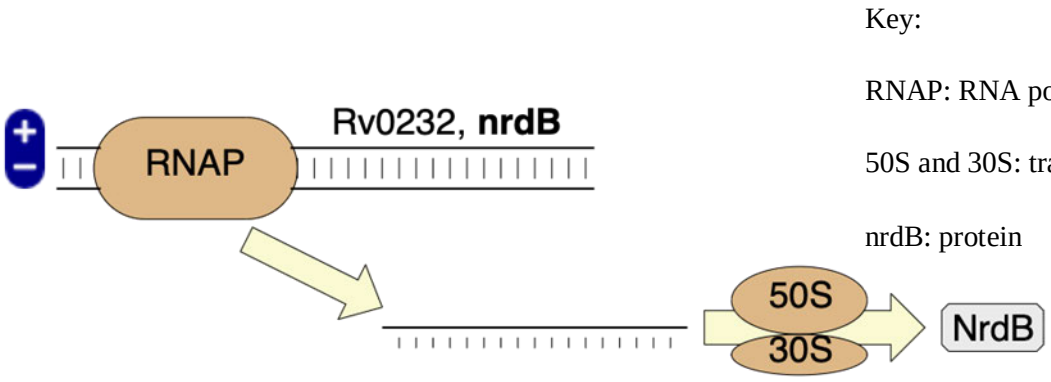


Figure 2.8: A graphical representation of the proposed regulation of *nrdB* by sRNA MSTRG.26.1 identified in F15/LAM4/KZN strain (Karp *et al.*, 2019).

Following mRNA sequencing analysis, F15/LAM4/KZN and Beijing strains induced 19 and 14 significantly regulated transcripts, respectively, compared to the laboratory H37Rv strain. All four sRNA transcripts targets were not identified in F15/LAM4/KZN (Figure 2.9A) and Beijing (Figure 2.9B) strains mRNA data. It can then be deduced that the identified sRNAs have an inhibitory role in the expression of the selected mRNA targets. The majority of the sRNA and mRNA transcripts identified in this study did not overlap with known and existing genes, hence, future studies will involve functional characterization of these MSTRG transcripts.

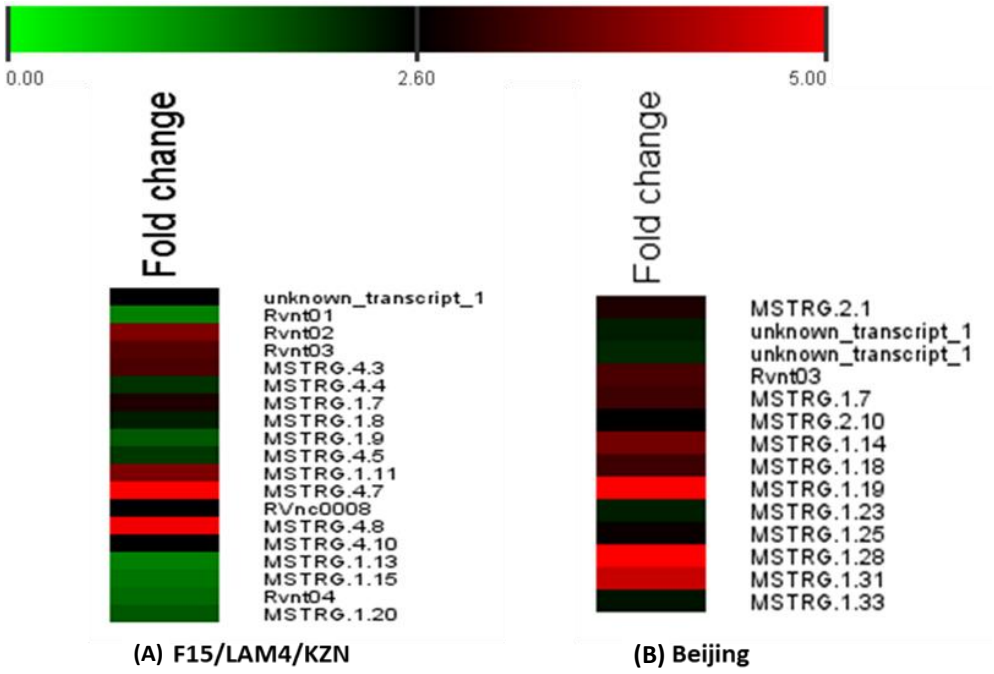


Figure 2.9: Significant mRNA transcript expression in (A) F15/LAM4/KZN and (B) Beijing strains compared to the laboratory H37Rv strain. mRNA sequencing reads were analyzed following the Hisat, Stringtie and Ballgown Bioinformatics pipeline. Only significantly ($p < 0.05$) were selected as fold changes and plotted for visualization using MeV. F15/LAM4/KZN induced 19 significantly regulated mRNA compared to the 14 induced by the Beijing strains against H37Rv laboratory control.

From the mRNA sequencing data, three genes that were present in both the F15/LAM4/KZN and Beijing strains, Rvnt01, Rvnt02 and MSTRG.1.7, were selected for confirmation of the fold change data from the sequencing analysis (Figure 2.10). These genes along with MSTRG.26.1, MSTRG.34.1, MSTRG.40.1 and MSTRG.53.1 were subjected to real time PCR amplification (qPCR) and fold changes are presented in Figure 2.10. Apart from MSTRG.40.1, MSTRG.53.1 and Rvnt01, the remaining mRNA and sRNA sequences are consistent between the sequencing and qPCR data. Differences observed in fold changes between sequencing and qPCR may be due to the lower efficiencies of the primers used in RT-qPCR. *M. tb* clinical strains, Unique and F11 were included in the qPCR analysis and significant variation was observed between all four strains, indicating changes in regulation of these transcripts.

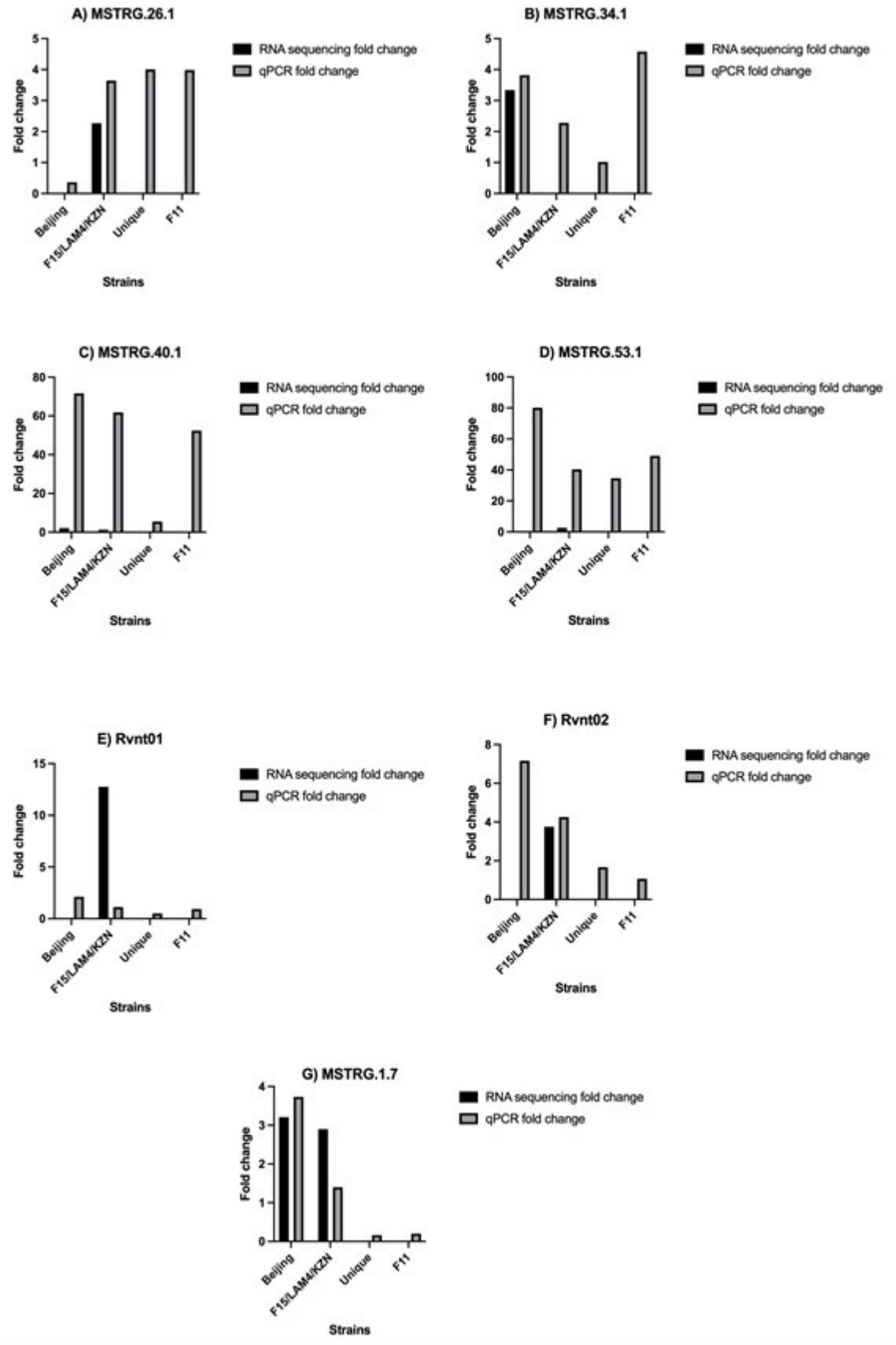


Figure 2.10: RNA sequencing and real time PCR quantification fold changes of (A) MSTRG.26.1,

(B) MSTRG.34.1, (C) MSTRG.40.1, (D) MSTRG.53.1, (E) Rvnot01, (F) Rvnt02 and (G) MSTRG.1.7.

The majority of the sRNA and mRNA transcripts from the clinical strains were found to be upregulated when compared to the H37Rv strain. MSTRG.26.1 was down-regulated by a fold change of 0.36 in the Beijing strain when compared to the H37Rv strain according to the qPCR data. Since the sequencing data did not indicate a similar result, it can be concluded that this down-regulation is insignificant as the primer efficiency may have caused inconsistencies in the RT-qPCR data. Both Unique and F11 *M. tb* clinical strains exhibited down-regulation with respect to the Rvnt01 and MSTRG.1.7 mRNA indicating a similar role may be shared between these transcripts.

Even though two out of the four targets play important roles, only one, fcr2, was identified as an essential gene. Mowa *et al.* (2009) showed that the *nrdB* gene was not essential for growth or survival of the bacterium and a stressful environment did not trigger expression. The predicted sRNAs obtained from the *in silico* analysis provides a starting point to begin investigating sRNAs in a lineage-specific manner, offering novel insight to the phenotypical characteristics that are observed between the lineages due to transcriptional changes. It was hypothesized that the sRNAs identified from the RNA sequencing data would be present in the *in silico* analysis, however, this was not the case. This may be due to strain-specific differences between the F15/LAM4/KZN and Beijing strains isolated in South Africa, compared to other globally distributed strains. Therefore, the sRNAs identified from RNA sequencing may be strain- or family-specific and may not necessarily represent the entire lineage. This can only be verified by sequencing multiple strains from the different lineages.

Due to the hypothetical nature of many of the sRNA targets, both mutated and consensus as well as the targets identified during sRNA sequencing, it is therefore critical to functionally characterize these as they may provide a novel perspective in the *M. tb* virulence and effective control measures. It is necessary to experimentally validate each of the sRNAs identified. Since sRNAs play a crucial role in transcriptome regulation, SNPs may cause variations in the way they interact with targets within different *M. tb* strain families and lineages. Differences in sRNA targets may be responsible for the phenotypic differences observed among the various strain families that may cause variations in their virulence and transmissibility.

The limitations of the study include the high-cost involved with the sRNA sequencing of each strain, therefore only three strains (F15/LAM4/KZN, Beijing, H37Rv) were sequenced. Although this study presents an effective method of screening sRNAs, there are still many that may be novel and unaccounted for without sequencing among the lineages of the MTBC. A larger

number of strains belonging to the same lineage is required in order to make inferences on whole lineage regulation. The novel sRNAs presented in the current study remain to be functionally characterized and validated, through Northern blotting and gene knockout experiments.

2.4 Conclusion

The MTBC lineage *in silico* analysis elucidated sRNAs and their corresponding targets that may play a significant role in the pathogenesis of the MTBC lineages. Several pathways involved in the metabolism of compounds for carbon, energy as well as lipid sources were implicated. The small RNA sequencing data demonstrated that sRNAs can play a significant role on the expression of genes by the absence of the putative targets identified. The sRNA MSTRG.40.1 exhibited differential expression between the Beijing and F15/LAM4/KZN strains, indicating a fascinating role that these sRNAs may play in transcriptome regulation of genetically diverse strains of MTBC. The findings in this study revealed MTBC transcriptome regulation through mutations of sRNAs, resulting in regulation of mRNA targets in lineage-specific manner. The proposed novel sRNA in clinical strains remains to be functionally characterized in future studies. The use of the laboratory H37Rv strain in TB research could hinder our understanding of crucial regulatory mechanisms that are critical in clinically relevant strains of MTBC, as shown by lineage-specific mutations and abundances. Understanding MTBC sRNAs regulatory mechanisms and their respective pathways can provide novel strategies to control TB.

References

Ami, V. K. G., Balasubramanian, R., & Hegde, S. R. (2020). Genome-wide identification of the context-dependent sRNA expression in *Mycobacterium tuberculosis*. *BMC Genomics, 21*(1), 1-12.

Andrews, S. (2010). FASTQC. A quality control tool for high throughput sequence data.

Arbing, M. A., Kaufmann, M., Phan, T., Chan, S., Cascio, D., & Eisenberg, D. (2010). The crystal structure of the *Mycobacterium tuberculosis* Rv3019c-Rv3020c ESX complex reveals a domain-swapped heterotetramer. *Protein Science, 19*(9), 1692-1703.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics, 25*(1), 25-29.

Bacon, J., James, B. W., Wernisch, L., Williams, A., Morley, K. A., Hatch, G. J., Mangan, J. A., Hinds, J., Stoker, N. G., Butcher, P. D., & Marsh, P. D. (2004). The influence of reduced

oxygen availability on pathogenicity and gene expression in *Mycobacterium tuberculosis*. *Tuberculosis, 84*(3-4), 205-217.

Banerjee, D. R., Senapati, K., Biswas, R., Das, A. K., & Basak, A. (2015). Inhibition of *M. tuberculosis* beta-ketoacyl CoA reductase FabG4 (Rv0242c) by triazole linked polyphenol-aminobenzene hybrids: comparison with the corresponding gallate counterparts. *Bioorganic and Medicinal Chemistry Letters, 25*(6), 1343-1347.

Basu, P., Sandhu, N., Bhatt, A., Singh, A., Balhana, R., Gobe, I., Crowhurst, N. A., Mendum, T. A., Gao, L., Ward, J. L., Beale, M. H., McFadden, J., & Beste, D. J. V. (2018). The anaplerotic node is essential for the intracellular survival of *Mycobacterium tuberculosis*. *Journal of Biological Chemistry, 293*(15), 5695-5704.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120.

Borrell, S., Trauner, A., Brites, D., Rigouts, L., Loiseau, C., Coscolla, M., Niemann, S., De Jong, B., Yeboah-Manu, D., Kato-Maeda, M., Feldmann, J., Reinhard, M., Beisel, C., & Gagneux, S. (2019). Reference set of *Mycobacterium tuberculosis* clinical strains: A tool for research and product development. *Public Library of Science One, 14*(3), 1-12.

Brown, A. K., Bhatt, A., Singh, A., Saparia, E., Evans, A. F., & Besra, G. S. (2007). Identification of the dehydratase component of the mycobacterial mycolic acid-synthesizing fatty acid synthase-II complex. *Microbiology, 153*(12), 4166-4173.

Burns-Huang, K., & Mundhra, S. (2019). *Mycobacterium tuberculosis* cysteine biosynthesis genes mec+-cysO-cysM confer resistance to clofazimine. *Tuberculosis, 115*, 63-66.

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., Ami, G. O. H., & Web Presence Working, G. (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics, 25*(2), 288-289.

Casamassimi, A., & Ciccodicola, A. (2019). Transcriptional Regulation: Molecules, Involved Mechanisms, and Misregulation. *International Journal of Molecular Sciences, 20*(6), 1-5.

Chihota, V. N., Muller, B., Mlambo, C. K., Pillay, M., Tait, M., Streicher, E. M., Marais, E., van der Spuy, G. D., Hanekom, M., Coetzee, G., Trollip, A., Hayes, C., Bosman, M. E., Gey van Pittius, N. C., Victor, T. C., van Helden, P. D., & Warren, R. M. (2012). Population structure of multi- and extensively drug-resistant *Mycobacterium tuberculosis* strains in South Africa. *Journal of Clinical Microbiology, 50*(3), 995-1002.

Chim, N., Riley, R., The, J., Im, S., Segelke, B., Lekin, T., Yu, M., Hung, L. W., Terwilliger, T., Whitelegge, J. P., & Goulding, C. W. (2010). An extracellular disulfide bond forming protein (DsbF) from *Mycobacterium tuberculosis*: structural, biochemical, and gene expression analysis. *Journal of Molecular Biology, 396*(5), 1211-1226.

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *Public Library of Science One, 5*(6), 1-17.

Deb, C., Lee, C. M., Dubey, V. S., Daniel, J., Abomoelak, B., Sirakova, T. D., Pawar, S., Rogers, L., & Kolattukudy, P. E. (2009). A novel in vitro multiple-stress dormancy model for *Mycobacterium tuberculosis* generates a lipid-loaded, drug-tolerant, dormant pathogen. *Public Library of Science One, 4*(6), 1-15.

DeJesus, M. A., Gerrick, E. R., Xu, W., Park, S. W., Long, J. E., Boutte, C. C., Rubin, E. J., Schnappinger, D., Ehrt, S., Fortune, S. M., Sassetti, C. M., & Ioerger, T. R. (2017). Comprehensive Essentiality Analysis of the *Mycobacterium tuberculosis* Genome via Saturating Transposon Mutagenesis. *MBio, 8*(1), 1-17.

Deshpande, R. G., Khan, M. B., Bhat, D. A., & Navalkar, R. G. (1997). Isolation of a contact-dependent haemolysin from *Mycobacterium tuberculosis*. *Journal of Medical Microbiology, 46*(3), 233-238.

Devasundaram, S., Khan, I., Kumar, N., Das, S., & Raja, A. (2015). The influence of reduced oxygen availability on gene expression in laboratory (H37Rv) and clinical strains (S7 and S10) of *Mycobacterium tuberculosis*. *Journal of Biotechnology, 210*, 70-80.

Dong, W., Huang, J., Li, Y., Tan, Y., Shen, Z., Song, Y., Wang, D., Xiao, S., Chen, H., Fu, Z. F., & Peng, G. (2015). Crystal structural basis for Rv0315, an immunostimulatory antigen and inactive beta-1,3-glucanase of *Mycobacterium tuberculosis*. *Scientific Reports, 5*, 1-14.

Duan, W., Li, X., Ge, Y., Yu, Z., Li, P., Li, J., Qin, L., & Xie, J. (2019). *Mycobacterium tuberculosis* Rv1473 is a novel macrolides ABC Efflux Pump regulated by WhiB7. *Future Microbiology, 14*, 47-59.

Ebrahimi-Rad, M., Bifani, P., Martin, C., Kremer, K., Samper, S., Rauzier, J., Kreiswirth, B., Blazquez, J., Jouan, M., van Soolingen, D., & Gicquel, B. (2003). Mutations in putative mutator genes of *Mycobacterium tuberculosis* strains of the W-Beijing family. *Emerging Infectious Diseases, 9*(7), 838-845.

Evans, J. C., Trujillo, C., Wang, Z., Eoh, H., Ehrt, S., Schnappinger, D., Boshoff, H. I., Rhee, K. Y., Barry, C. E., 3rd, & Mizrahi, V. (2016). Validation of CoaBC as a Bactericidal Target in the Coenzyme A Pathway of *Mycobacterium tuberculosis*. *American Chemical Society Infectious Diseases, 2*(12), 958-968.

Fay, A., Czudnochowski, N., Rock, J. M., Johnson, J. R., Krogan, N. J., Rosenberg, O., & Glickman, M. S. (2019). Two Accessory Proteins Govern MmpL3 Mycolic Acid Transport in Mycobacteria. *MBio, 10*(3), 1-17.

Frees, D., Qazi, S. N., Hill, P. J., & Ingmer, H. (2003). Alternative roles of ClpX and ClpP in *Staphylococcus aureus* stress tolerance and virulence. *Molecular Microbiology, 48*(6), 1565-1578.

Fu, J., Frazee, A. C., Collado-Torres, L. J., A.E., & Leek, J. T. (2021). Ballgown: Flexible, isoform-level differential expression analysis. R package version 2.26.0.

Galagan, J., Murray, M., Pillay, M., Borowsky, M. L., Young, S. K., Zeng, Q., Koehrsen, M., Fitzgerald, M., Abouelleil, A., Alvarado, L., Arachchi, H. M., Berlin, A. M., Borenstein, D., Brown, A., Chapman, S. B., Chen, Z., Dunbar, C., Engels, R., Freedman, E., Gearin, G., Gellesch, M., Goldberg, J., Griggs, A., Gujja, S., Heilman, E. R., Heiman, D. I., Howarth, C., Jen, D., Larson, L., Lui, A., MacDonald, J. P., Mehta, T., Montmayeur, A., Neiman, D., Park, D., Pearson, M., Priest, M., Richards, J., Roberts, A., Saif, S., Shea, T. D., Shenoy, N., Sisk, P., Stolte, C., Sykes, S. N., Walk, T., White, J., Yandava, C., Haas, B., Nusbaum, C., & Birren, B. (2010). *The Genome Sequence of Mycobacterium tuberculosis strain KZN 605*. The Broad Institute Genome Sequencing Platform, The Broad Institute Genome Sequencing Center for Infectious Disease. Cambridge.

Gandhi, N. R., Brust, J. C., Moodley, P., Weissman, D., Heo, M., Ning, Y., Moll, A. P., Friedland, G. H., Sturm, A. W., & Shah, N. S. (2014). Minimal diversity of drug-resistant *Mycobacterium tuberculosis* strains, South Africa. *Emerging Infectious Diseases, 20*(3), 426-433.

Gandhi, N. R., Moll, A., Sturm, A. W., Pawinski, R., Govender, T., Lalloo, U., Zeller, K., Andrews, J., & Friedland, G. (2006). Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. *The Lancet, 368*(9547), 1575-1580.

Gautam, U. S., Mehra, S., Kumari, P., Alvarez, X., Niu, T., Tyagi, J. S., & Kaushal, D. (2019). *Mycobacterium tuberculosis* sensor kinase DosS modulates the autophagosome in a DosR-independent manner. *Communications Biology, 2*, 1-12.

Gerrick, E. R., Barbier, T., Chase, M. R., Xu, R., Francois, J., Lin, V. H., Szucs, M. J., Rock, J. M., Ahmad, R., Tjaden, B., Livny, J., & Fortune, S. M. (2018). Small RNA profiling in *Mycobacterium tuberculosis* identifies MrsI as necessary for an anticipatory iron sparing response. *Proceedings of the National Academy of Sciences of the United States of America, 115*(25), 6464-6469.

Goldstone, D. C., Metcalf, P., & Baker, E. N. (2016). Structure of the ectodomain of the electron transporter Rv2874 from *Mycobacterium tuberculosis* reveals a thioredoxin-like domain combined with a carbohydrate-binding module. *Acta Crystallographica Section D: Structural Biology, 72*(Pt 1), 40-48.

Gong, Z., Li, H., Cai, Y., Stojkoska, A., & Xie, J. (2019). Biology of MarR family transcription factors and implications for targets of antibiotics against tuberculosis. *Journal of Cellular Physiology, 234*(11), 19237-19248.

Gottesman, S., & Storz, G. (2011). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harbor Perspectives in Biology, 3*(12), 1-16.

Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R., & Hofacker, I. L. (2008). The Vienna RNA websuite. *Nucleic Acids Research, 36*(Web Server issue), W70-74.

Gruber, A. R., Neubock, R., Hofacker, I. L., & Washietl, S. (2007). The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Research, 35*(Web Server issue), W335-338.

Han, H., & Wilson, A. C. (2013). The two CcdA proteins of *Bacillus anthracis* differentially affect virulence gene expression and sporulation. *Journal of Bacteriology, 195*(23), 5242-5249.

Haning, K., Cho, S. H., & Contreras, L. M. (2014a). Small RNAs in mycobacteria: an unfolding story. *Frontiers in Cellular and Infection Microbiology, 4*, 1-11.

Herbst, D. A., Jakob, R. P., Zahringer, F., & Maier, T. (2016). Mycocerosic acid synthase exemplifies the architecture of reducing polyketide synthases. *Nature, 531*(7595), 533-537.

Homolka, S., Niemann, S., Russell, D. G., & Rohde, K. H. (2010). Functional genetic diversity among *Mycobacterium tuberculosis* complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. *The Public Library of Science Pathogens, 6*(7), 1-17.

Howe, E. A., Sinha, R., Schlauch, D., & Quackenbush, J. (2011). RNA-Seq analysis in MeV. *Bioinformatics, 27*(22), 3209-3210.

Ioerger, T. R., Feng, Y., Chen, X., Dobos, K. M., Victor, T. C., Streicher, E. M., Warren, R. M., Gey van Pittius, N. C., Van Helden, P. D., & Sacchettini, J. C. (2010). The non-clonality of drug resistance in Beijing-genotype isolates of *Mycobacterium tuberculosis* from the Western Cape of South Africa. *BMC Genomics, 11*, 1-14.

Jost, D., Nowojewski, A., & Levine, E. (2011). Small RNA biology is systems biology. *Biochemistry and Molecular Biology Reports, 44*(1), 11-21.

Kapopoulou, A., Lew, J. M., & Cole, S. T. (2011). The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis, 91*(1), 8-13.

Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S. M., & Subhraveti, P. (2019). The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics, 20*(4), 1085-1093.

Khan, M. Z., Singha, B., Ali, M. F., Taunk, K., Rapole, S., Gourinath, S., & Nandicoori, V. K. (2021). Redox homeostasis in *Mycobacterium tuberculosis* is modulated by a novel actinomycete-specific transcription factor. *European Molecular Biology Organization Journal, 40*(14), 1-23.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods, 12*(4), 357-360.

Krawczyk, J., Kohl, T. A., Goesmann, A., Kalinowski, J., & Baumbach, J. (2009). From *Corynebacterium glutamicum* to *Mycobacterium tuberculosis*--towards transfers of gene regulatory networks and integrated data analyses with MycoRegNet. *Nucleic Acids Research, 37*(14), 1-15.

Krueger, F. (2015). Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, 516*, 517.

Kweon, O., Kim, S. J., Blom, J., Kim, S. K., Kim, B. S., Baek, D. H., Park, S. I., Sutherland, J. B., & Cerniglia, C. E. (2015). Comparative functional pan-genome analyses to build connections between genomic dynamics and phenotypic evolution in polycyclic aromatic hydrocarbon metabolism in the genus *Mycobacterium*. *BMC Evolutionary Biology, 15*(21), 1-23.

Lee, C. E., Goodfellow, C., Javid-Majd, F., Baker, E. N., & Shaun Lott, J. (2006). The crystal structure of TrpD, a metabolic enzyme essential for lung colonization by *Mycobacterium tuberculosis*, in complex with its substrate phosphoribosylpyrophosphate. *Journal of Molecular Biology, 355*(4), 784-797.

Lew, J. M., Kapopoulou, A., Jones, L. M., & Cole, S. T. (2011). TubercuList - 10 years after. *Tuberculosis, 91*(1), 1-7.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.

Li, Q., Fu, T., Li, C., Fan, X., & Xie, J. (2016). Mycobacterial IclR family transcriptional factor Rv2989 is specifically involved in isoniazid tolerance by regulating the expression of catalase encoding gene katG. *RSC Advances, 6*(60), 54661-54667.

Li, W., & Lv, L. (2015). *Identification of the Mycobacterium tuberculosis reference strain of China: Whole genome sequencing and characterization*. Beijing Tuberculosis and Thoracic Tumor Research Institute. Beijing.

Long, M. C., Escuyer, V., & Parker, W. B. (2003). Identification and Characterization of a Unique Adenosine Kinase from *Mycobacterium tuberculosis*. *Journal of Bacteriology, 185*(22), 6548-6555.

Lott, J. S. (2020). The tryptophan biosynthetic pathway is essential for *Mycobacterium tuberculosis* to cause disease. *Biochemical Society Transactions, 48*(5), 2029-2037.

Lu, R., Schmitz, W., & Sampson, N. S. (2015). alpha-Methyl Acyl CoA Racemase Provides *Mycobacterium tuberculosis* Catabolic Access to Cholesterol Esters. *Biochemistry, 54*(37), 5669-5672.

Mann, M., Wright, P. R., & Backofen, R. (2017). IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions. *Nucleic Acids Research, 45*(W1), W435–W439.

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research, 47*(D1), D419-D426.

Michaux, C., Verneuil, N., Hartke, A., & Giard, J. C. (2014). Physiological roles of small RNA molecules. *Microbiology (Reading), 160*(Pt 6), 1007-1019.

Minato, Y., Gohl, D. M., Thiede, J. M., Chacon, J. M., Harcombe, W. R., Maruyama, F., & Baughn, A. D. (2019). Genomewide Assessment of *Mycobacterium tuberculosis* Conditionally Essential Metabolic Pathways. *mSystems, 4*(4), 1-13.

Mowa, M. B., Warner, D. F., Kaplan, G., Kana, B. D., & Mizrahi, V. (2009). Function and regulation of class I ribonucleotide reductase-encoding genes in mycobacteria. *Journal of Bacteriology, 191*(3), 985-995.

Ngabonziza, J.-C., Loiseau, C.-M., Marceau, M., Jouet, A., Menardo, F., Tzfadia, O., Niyigena, E. B., Mulders, W., Fissette, K., Diels, M., Ssengooba, W., Kaswa, M. K., Habimana, Y. M., Brites, D., Affolabi, D., Mazarati, J. B., de Jong, B. C., Rigouts, L., Gagneux, S., Meehan, C. J., & Supply, P. (2020). *A new lineage and primary branch of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region*. Genome. Department of Biomedical Sciences. Unit of Mycobacteriology, Institute of Tropical Medicine Antwer. Belgium.

Nizolenko, L., Kozhina, E., Yarygin, A., & Bachinsky, A. (2005). Study of the Amino Acid Sequences of Open Reading Frames of the Complete Genome of *Mycobacterium tuberculosis* Using the Protein Family Pattern Bank Prof_Pat. *Biophysics, 50*(6), 19-23.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research, 27*(1), 29-34.

Ollinger, J., O'Malley, T., Kesicki, E. A., Odingo, J., & Parish, T. (2012). Validation of the essential ClpP protease in *Mycobacterium tuberculosis* as a novel drug target. *Journal of Bacteriology, 194*(3), 663-668.

Ostrik, A. A., Salina, E. G., Skvortsova, Y. V., Grigorov, A. S., Bychenko, O. S., Kaprelyants, A. S., & Azhikina, T. L. (2020). Small RNAs of *Mycobacterium tuberculosis* in Adaptation

to Host-Like Stress Conditions *in vitro*. *Applied Biochemistry and Microbiology, 56*(4), 381-386.

Pang, J. M., Layre, E., Sweet, L., Sherrid, A., Moody, D. B., Ojha, A., & Sherman, D. R. (2012). The Polyketide Pks1 Contributes to Biofilm Formation in *Mycobacterium tuberculosis*. *Journal of Bacteriology, 194*(3), 715-721.

Papenfort, K., & Vogel, J. (2009). Multiple target regulation by small noncoding RNAs rewires gene expression at the post-transcriptional level. *Research in Microbiology, 160*(4), 278-287.

Parida, B. K., Douglas, T., Nino, C., & Dhandayuthapani, S. (2005). Interactions of anti-sigma factor antagonists of *Mycobacterium tuberculosis* in the yeast two-hybrid system. *Tuberculosis, 85*(5-6), 347-355.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols, 11*(9), 1650-1667.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology, 33*(3), 290-295.

Rajaram, M. V., Ni, B., Morris, J. D., Brooks, M. N., Carlson, T. K., Bakthavachalu, B., Schoenberg, D. R., Torrelles, J. B., & Schlesinger, L. S. (2011). *Mycobacterium tuberculosis* lipomannan blocks TNF biosynthesis by regulating macrophage MAPK-activated protein kinase 2 (MK2) and microRNA miR-125b. *Proceedings of the National Academy of Sciences, 108*(42), 17408-17413.

Roy, S., Vijay, S., Arumugam, M., Anand, D., Mir, M., & Ajitkumar, P. (2011). *Mycobacterium tuberculosis* expresses ftsE gene through multiple transcripts. *Current Microbiology, 62*(5), 1581-1589.

Sacco, E., Covarrubias, A. S., O'Hare, H. M., Carroll, P., Eynard, N., Jones, T. A., Parish, T., Daffe, M., Backbro, K., & Quemard, A. (2007). The missing piece of the type II fatty acid synthase system from *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America, 104*(37), 14628-14633.

Salina, E. G., Grigorov, A., Skvortsova, Y., Majorov, K., Bychenko, O., Ostrik, A., Logunova, N., Ignatov, D., Kaprelyants, A., Apt, A., & Azhikina, T. (2019). MTS1338, A Small *Mycobacterium tuberculosis* RNA, Regulates Transcriptional Shifts Consistent With Bacterial Adaptation for Entering Into Dormancy and Survival Within Host Macrophages. *Frontiers in Cellular and Infection Microbiology, 9*(405), 1-11.

Salina, E. G., Huszar, S., Zemanova, J., Keruchenko, J., Riabova, O., Kazakova, E., Grigorov, A., Azhikina, T., Kaprelyants, A., Mikusova, K., & Makarov, V. (2018). Copper-related

toxicity in replicating and dormant *Mycobacterium tuberculosis* caused by 1-hydroxy-5-R-pyridine-2(1H)-thiones. *Metallomics, 10*(7), 992-1002.

Sambrook, J., & Russell, D. W. (2001). *Molecular cloning : a laboratory manual* (3rd ed. ed. Vol. 3). Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Sassetti, C. M., Boyd, D. H., & Rubin, E. J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. *Molecular Microbiology, 48*(1), 77-84.

Savolainen, K., Bhaumik, P., Schmitz, W., Kotti, T. J., Conzelmann, E., Wierenga, R. K., & Hiltunen, J. K. (2005). Alpha-methylacyl-CoA racemase from *Mycobacterium tuberculosis*. Mutational and structural characterization of the active site and the fold. *Journal of Biological Chemistry, 280*(13), 12611-12620.

Schulz, E. C., Henderson, S. R., Illarionov, B., Crosskey, T., Southall, S. M., Krichel, B., Uetrecht, C., Fischer, M., & Wilmanns, M. (2020). The crystal structure of mycobacterial epoxide hydrolase A. *Scientific Reports, 10*(16539), 1-14.

Sirakova, T. D., Deb, C., Daniel, J., Singh, H. D., Maamar, H., Dubey, V. S., & Kolattukudy, P. E. (2012). Wax ester synthesis is required for *Mycobacterium tuberculosis* to enter in vitro dormancy. *Public Library of Science One, 7*(12), 1-15.

Smyth, G. K. (2005). limma: Linear Models for Microarray Data. In R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, & S. Dudoit (Eds.), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (pp. 397-420). New York, NY: Springer New York.

Solano-Gutierrez, J. S., Pino, C., & Robledo, J. (2019). Toxin-antitoxin systems shows variability among *Mycobacterium tuberculosis* lineages. *FEMS Microbiology Letters, 366*(1), 1-8.

Solans, L., Gonzalo-Asensio, J., Sala, C., Benjak, A., Uplekar, S., Rougemont, J., Guilhot, C., Malaga, W., Martin, C., & Cole, S. T. (2014). The PhoP-dependent ncRNA Mcr7 modulates the TAT secretion system in *Mycobacterium tuberculosis*. *The Public Library of Science Pathogens, 10*(5), 1-17.

Sreenu, V. B., Kumar, P., Nagaraju, J., & Nagarajaram, H. A. (2006). Microsatellite polymorphism across the *M. tuberculosis* and *M. bovis* genomes: implications on genome evolution and plasticity. *BMC Genomics, 7*(78), 1-10.

Storz, G., Vogel, J., & Wassarman, K. M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. *Molecular Cell, 43*(6), 880-891.

Sun, L., Zhang, L., Wang, T., Jiao, W., Li, Q., Yin, Q., Li, J., Qi, H., Xu, F., Shen, C., Xiao, J., Liu, S., Mokrousov, I., Huang, H., & Shen, A. (2019). Mutations of *Mycobacterium tuberculosis* induced by anti-tuberculosis treatment result in metabolism changes and elevation of ethambutol resistance. *Infection, Genetics and Evolution, 72*, 151-158.

Surken, M., Keller, C., Rohker, C., Ehlers, S., & Bange, F. C. (2008). Anaerobic arginine metabolism of *Mycobacterium tuberculosis* is mediated by arginine deiminase (arcA),

but is not essential for chronic persistence in an aerogenic mouse model of infection. *International Journal of Medical Microbiology, 298*(7-8), 657-661.

Thakur, M., & Chakraborti, P. K. (2008). Ability of PknA, a mycobacterial eukaryotic-type serine/threonine kinase, to transphosphorylate MurD, a ligase involved in the process of peptidoglycan biosynthesis. *Biochemical Journal, 415*(1), 27-33.

Tsolaki, A. G., Hirsh, A. E., DeRiemer, K., Enciso, J. A., Wong, M. Z., Hannan, M., Goguet de la Salmoniere, Y. O., Aman, K., Kato-Maeda, M., & Small, P. M. (2004). Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. *Proceedings of the National Academy of Sciences of the United States of America, 101*(14), 4865-4870.

Turkarslan, S., Peterson, E. J., Rustad, T. R., Minch, K. J., Reiss, D. J., Morrison, R., Ma, S., Price, N. D., Sherman, D. R., & Baliga, N. S. (2015). A comprehensive map of genome-wide gene regulation in *Mycobacterium tuberculosis*. *Scientific Data, 2*(150010), 1-10.

Tyagi, J. S., Das, T. K., & Kinger, A. K. (1996). An *M. tuberculosis* DNA fragment contains genes encoding cell division proteins ftsX and ftsE, a basic protein and homologues of PemK and Small protein B. *Gene, 177*(1-2), 59-67.

Ullah, N., Hao, L., Banga Ndzouboukou, J. L., Chen, S., Wu, Y., Li, L., Borham Mohamed, E., Hu, Y., & Fan, X. (2021). Label-Free Comparative Proteomics of Differentially Expressed *Mycobacterium tuberculosis* Protein in Rifampicin-Related Drug-Resistant Strains. *Pathogens, 10*(5), 1-24.

UniProt, C. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research, 47*(D1), D506-D515.

Weyand, S., Kefala, G., & Weiss, M. S. (2006). Cloning, expression, purification, crystallization and preliminary X-ray diffraction analysis of DapC (Rv0858c) from *Mycobacterium tuberculosis*. *Acta Crystallographica. Section F, Structural Biology Communications, 62*(Pt 8), 794-797.

Wolfe, L. M., Mahaffey, S. B., Kruh, N. A., & Dobos, K. M. (2010). Proteomic definition of the cell wall of *Mycobacterium tuberculosis*. *Journal of Proteome Research, 9*(11), 5816-5826.

World Health Organization. (2021). *Global tuberculosis report 2021* (978-92-4-003702-1). Retrieved from Geneva:

Yang, Q., Liu, Y., Huang, F., & He, Z. G. (2011). Physical and functional interaction between D-ribokinase and topoisomerase I has opposite effects on their respective activity in *Mycobacterium smegmatis* and *Mycobacterium tuberculosis*. *Archives of Biochemistry and Biophysics, 512*(2), 135-142.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., & Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC bioinformatics*, *13*, 134.

Yew, W. W., Liang, D., Chan, D. P., Shi, W., & Zhang, Y. (2017). Molecular mechanisms of clofazimine resistance in *Mycobacterium tuberculosis*. *Journal of Antimicrobial Chemotherapy*, *72*(10), 2943-2944.

Zhang, L., Hendrickson, R. C., Meikle, V., Lefkowitz, E. J., Ioerger, T. R., & Niederweis, M. (2020). Comprehensive analysis of iron utilization by *Mycobacterium tuberculosis*. *The Public Library of Science Pathogens*, *16*(2), 1-27.

Zhang, X., Chen, S., Hu, Z., Zhang, L., & Wang, H. (2009). Expression and characterization of two functional methionine aminopeptidases from *Mycobacterium tuberculosis* H37Rv. *Current Microbiology, 59*(5), 520-525.

Zhao, J., Siddiqui, S., Shang, S., Bian, Y., Bagchi, S., He, Y., & Wang, C. R. (2015). Mycolic acid-specific T cells protect against *Mycobacterium tuberculosis* infection in a humanized transgenic mouse model. *elife, 4*, 1-18.

Zhu, L., Sharp, J. D., Kobayashi, H., Woychik, N. A., & Inouye, M. (2010). Noncognate *Mycobacterium tuberculosis* toxin-antitoxins can physically and functionally interact. *Journal of Biological Chemistry*, *285*(51), 39732-39738.

CHAPTER 3

3.1 General Discussion and Recommendations

Tuberculosis (TB) is a global threat to health and until the COVID-19 pandemic, was the leading cause of death from an infectious agent (World Health Organization, 2021). With South Africa being part of the eight countries contributing to two thirds of the people who are estimated to have developed TB in 2020, the eradication of TB is of particular importance. The causative agent of TB belongs to well-known human-adapted members of the *Mycobacterium tuberculosis* complex (MTBC) with 8 well defined lineages, distributed as genetically diverse strains in different geographic locations. Despite the current understanding of molecular mechanisms governing pathogenesis of clinical strains of *Mycobacterium tuberculosis* (*M. tb),* there is still a great need to investigate alternative methods and pathways that can be used to target, treat or prevent the development of this disease (Miotto *et al.*, 2012a). An approach to this would be studying the regulatory processes of *M. tb*, with sRNAs being excellent candidates. Their fast synthesis and low-cost to the bacterium makes them an efficient choice for regulation induced by stress (Rose *et al.*, 2013). Gene expression can be regulated by sRNAs by direct base-pairing to one or more target mRNA, either inhibiting translation or increasing translational efficiency and mRNA stability (Storz *et al.*, 2011).

Due to the high sequence similarity of the strains belonging to the MTBC, there are underlying regulatory mechanisms at play to elicit significant changes that are observed phenotypically, in particular the pathogenicity and virulence of the clinical strains (Homolka *et al.*, 2010). Very little is known about the underlying mechanisms contributing to transcriptome regulation within the strains of the MTBC, in particular the functional characterization of the sRNAs identified and none being investigated within clinical strains. This study investigated sRNAs containing lineage-specific mutations and their subsequent effect on the MTBC lineages. Furthermore, transcriptome regulation by sRNAs in clinical strains of *M. tb* was elucidated through RNA sequencing.

The *in silico* analysis of sRNAs uncovered 28 sRNAs containing lineage-specific mutations. The highest number of sRNAs identified per lineage belonged to lineage 7 suggesting a higher amount of variation within that lineage, which may cause significant transcriptional changes. The sRNAs with lineage-specific mutations were selected to better understand the notable variation in their virulence and pathogenicity. The mutations within several sRNAs caused significant structural changes to the proposed conformation of the sRNA as well as changes in the base-pairing probabilities. This would have a direct relationship to the targets which they regulate and in turn have a significant effect on the bacterium itself.

Lineages 1 and 3 presented with only one sRNA each, however no significant changes were observed in the lineage 3 variant, indicating that the sRNA would not have a large enough impact on the lineage. The lineage 1 sRNA, however, exhibited a clear structural change in the variant sRNA, with the variant targeting the *mez* gene. With its involvement in lipid biosynthesis, the absence of *mez* was found to cause changes in the cell wall formation and decreased efficiency during macrophage entry (Basu *et al.*, 2018). The variant sRNA may bind to the *mez* mRNA ensuring the gene is translated, supporting macrophage entry of the strains belonging to this lineage.

Some regulatory RNAs may be able to restore activity of its target, such as in the case of the lineage 2 sRNA, 16881. The variant sRNA had a lower base-pairing probability, which may cause the sRNA to preferentially bind to the variant target which is a β-1,3-glucanase. The binding of the sRNA to this protein may restore hydrolytic and glucanse activity as the current findings suggest that β-1,3-glucanase is inactive (W. Dong *et al.*, 2015).

The sRNA 29983 may provide lineage 2 with a selective advantage associated with *M. tb* defence. The variant sRNA targets a TMM transport factor A, a protein involved the MmpL3 transport machinery responsible for the transport of mycolic acid (Fay *et al.*, 2019). If this sRNA induces overexpression of this target mRNA, it may provide faster transportation of mycolic acids, causing the bacterium to become less susceptible to antibiotics. Virulence may also be affected in a lineage-specific manner. This is seen in the sRNA 26139 identified in lineage 4. The variant sRNA targets the gene encoding a member of the CcdA family of electron transporters (Goldstone *et al.*, 2016). A previous study discovered that the loss of two copies of the *ccdA* gene resulted in increased virulence factor expression in *B. anthracis* (Han & Wilson, 2013). By sRNA 26139 binding to the target, the resulting effect may cause the *ccdA* gene to become inactive and have a similar effect on virulence factor expression in the strains belonging to lineage 4 of MTBC. The variant sRNA 1090 identified in lineage 7 targets a hemolysin-like protein that is possibly involved in virulence (Deshpande *et al.*, 1997; L. Zhang *et al.*, 2020). The structural change observed between the consensus and variant sRNAs favours the attachment of the variant target, which may lead to an increase in virulence of this lineage.

Virulence can also be decreased by sRNAs. Within lineage 6, the consensus target for sRNA 26173 is the *ClpX* transcript. This encodes the ATP-dependent Clp protease ATP-binding subunit which maintains protein function under stress and can contribute to virulence of the *Mycobacterium* (Frees *et al.*, 2003). Due to the lower base pairing probability of this target and the open structure, sRNA 26173 may preferentially bind to this target. Since the mutated sRNAs target differed, lineage 6 may lack this increase virulence. The sRNA identified in lineage 7 may

potentially play an adaptative role for the strains belonging to this lineage. The variant target for sRNA 37867 targets the *rsfA* gene, encoding an anti-sigma factor antagonist, which in turn negatively regulates another anti-sigma factor (Parida *et al.*, 2005; Salina *et al.*, 2018). Due to the significant difference between the variant and consensus sRNA structures, the mutation may cause the variant target to be preferentially selected. This may in turn regulate the expression of sigma factor SigF when exposed to stress, thereby increasing the adaptability of *M. tb* to changing environments, such as iron-limiting conditions (Parida *et al.*, 2005).

Mutations within several sRNAs belonging to lineage 8 suggest that the remaining lineages may have a selective advantage over lineage 8. In the case of sRNA 11527, the consensus target is a possible monooxygenase that has been identified to be upregulated during iron-limiting conditions (Ami *et al.*, 2020). This may assist other lineages during this environmental stress, however due to the mutations, strains belonging to lineage 8 may not benefit. The EsxR protein, that plays a mediatory role during host and *M. tb* interactions, was the consensus target for sRNA 21006 (Arbing *et al.*, 2010). This sRNA may confer a selective advantage for the other lineages as the consensus sRNA has a lower base-pairing probability increasing the likeliness of this target pairing with the sRNA. This may be advantageous for *M .tb* survival within the host.

A disruption of a gene may also confer a selective advantage and can be seen in sRNA 27232. The consensus target encodes a transmembrane multidrug efflux pump, that is involved in the transport of drugs across the membrane. A study revealed that disruption of this gene increases bacterial growth (DeJesus *et al.*, 2017). The mutated sRNA found in lineage 8 may not be able to inhibit the expression of this gene, causing reduced bacterial growth.

In one instance (sRNA 27226), both consensus and mutant targets were found to play a role in reducing damage caused by oxidative stress. Even though these two targets control the same pathway, one sRNA may result in the corresponding target to become more efficient providing either lineage 8 or the remaining lineages with a selective advantage.

The mutation identified in lineage 7 for sRNA 23747, may play an interesting role for the strains belonging to that lineage. The mutated target for this sRNA encodes a possible 4-carboxymuconolactone decarboxylase, a protein involved in aromatic hydrocarbon metabolism (Wolfe *et al.*, 2010). The exact role of this sRNA would need to be further investigated as it may allow for increased carbon sources to be utilized, which is beneficial to the bacterium or produce compounds that could cause oxidative damage. The regulation of the target by sRNA 23747 could change to increase utilization of carbon sources or be repressed to prevent oxidative damage.

Small RNA and mRNA sequencing was performed on two clinical strains belonging to the Beijing and F15/LAM4/KZN families with the H37Rv strain used as the reference. From the

selected parameters, 4 novel sRNAs were identified from the sRNA sequencing. Following IntaRNA target prediction, one target was predicted as a hypothetical protein, one as a fatty acyl-CoA reductase, an integral membrane protein and a ribonucleoside-diphosphate reductase subunit beta. The sRNA MSTRG.40.1 was present in both Beijing and KZN/LAM4/KZN strains, with varying fold changes of 2.1 and 1.27, respectively. This is an indication of a variation in strain expression of this sRNA. There may be underlying mechanisms involving the hypothetical protein that requires further investigation. MSTRG.34.1 identified in the Beijing strain, targets the fatty acyl-CoA reductase, that is involved in wax ester synthesis and which may have an effect on the storage of wax esters as an alternate form of lipids during dormancy of the *Mycobacterium* (Deb *et al.*, 2009). This may provide Beijing strains with a selective advantage during dormancy.

Two other sRNAs were identified in the F15/LAM4/KZN strain. The first, MSTRG.53.1, targets an integral membrane protein, while MSTRG.26.1 is thought to target the ribonucleoside-diphosphate reductase subunit beta, encoded by the *nrdB* gene. The fold changes of the sRNAs and mRNAs were confirmed by real-time PCR. A comparison between the fold changes generated by the sequencing data and the real time qPCR data revealed variation between them. This may be owed to the low PCR efficiencies calculated during real- time qPCR. RNA-sequencing is more sensitive however, as it quantifies individual reads instead of calculation relative expression, allowing for subtle changes in expression to be detected.

All novel sRNAs were observed in the clinical strains (Beijing, F15/LAM4/KZN, F11 and Unique) with significant variation observed between the strains. This suggests regulation from strains family level within the same lineage. None of the sRNA targets predicted were identified from the mRNA sequencing data suggesting that these transcripts were not transcribed and the sRNAs act in an inhibitory manner on these transcripts. This can be achieved through binding of the sRNA to its respective target, blocking the ribosomal binding site and can result in cleavage of the RNA-RNA interaction by RNases (Bossi *et al.*, 2012).

3.2 Limitations

Due to the expensive nature of sequencing, only two clinical strains belonging to the highly prevalent South African *M. tb* families and one laboratory strain was sequenced and is a subsequent limitation to the study. It is not possible to base current findings from the sequencing data as a representation of the entire lineage. Many novel and uncharacterized sRNAs could also be overlooked without sequencing. The fold changes calculated through sRNA sequencing remains to be validated with Northern blotting.

3.3 Future recommendations

Due to the hypothetical nature of the sRNA analysis, laboratory confirmation is required. The methodology presented is ideal for screening large sets of data that would be tedious and time-consuming to perform experimentally. Real- time qPCR can be used as a cheaper alternative to RNA sequencing to investigate the expression of specific sRNAs belonging to many representatives of several lineages, provided that the primer efficiency remains in the range of 95% -100%. The role of hypothetical protein Rv2304c, that interact with the sRNA MSTRG.40.1, should be categorized in order to uncover potential role that may affect the strains in a lineage-specific manner causing the targeting sRNA to be differentially expressed. Future studies will confirm the predicted mRNA targets identified in the *in silico* analysis, together with functional annotation of the novel sRNAs.

3.4 Conclusion

The *in silico* analysis of the MBTC revealed several sRNAs and their targets that may play a role in lineage-specific traits observed, with functions such as macrophage entry, lipid biosynthesis and environment adaptation mechanisms being potentially implicated. Novel sRNAs were identified from sRNA sequencing analysis belonging to the Beijing and F15/LAM4/KZN families with varying fold changes indicative of an underlying regulation of these transcripts. Although the sRNAs would need to be confirmed, this study presents cost-effective alternative for screening potential sRNAs from whole genomes. By understanding the underlying regulatory mechanisms of the bacterium, it brings us one step closer to uncovering a potential control option for the fight against TB.

References

Ami, V. K. G., Balasubramanian, R., & Hegde, S. R. (2020). Genome-wide identification of the context-dependent sRNA expression in *Mycobacterium tuberculosis*. *BMC Genomics, 21*(1), 1-12.

Arbing, M. A., Kaufmann, M., Phan, T., Chan, S., Cascio, D., & Eisenberg, D. (2010). The crystal structure of the *Mycobacterium tuberculosis* Rv3019c-Rv3020c ESX complex reveals a domain-swapped heterotetramer. *Protein Science, 19*(9), 1692-1703.

Basu, P., Sandhu, N., Bhatt, A., Singh, A., Balhana, R., Gobe, I., Crowhurst, N. A., Mendum, T. A., Gao, L., Ward, J. L., Beale, M. H., McFadden, J., & Beste, D. J. V. (2018). The anaplerotic node is essential for the intracellular survival of *Mycobacterium tuberculosis*. *Journal of Biological Chemistry, 293*(15), 5695-5704.

Bossi, L., Schwartz, A., Guillemardet, B., Boudvillain, M., & Figueroa-Bossi, N. (2012). A role for Rho-dependent polarity in gene regulation by a noncoding small RNA. *Genes & Development, 26*(16), 1864-1873.

Deb, C., Lee, C. M., Dubey, V. S., Daniel, J., Abomoelak, B., Sirakova, T. D., Pawar, S., Rogers, L., & Kolattukudy, P. E. (2009). A novel in vitro multiple-stress dormancy model for *Mycobacterium tuberculosis* generates a lipid-loaded, drug-tolerant, dormant pathogen. *Public Library of Science One, 4*(6), 1-15.

DeJesus, M. A., Gerrick, E. R., Xu, W., Park, S. W., Long, J. E., Boutte, C. C., Rubin, E. J., Schnappinger, D., Ehrt, S., Fortune, S. M., Sassetti, C. M., & Ioerger, T. R. (2017). Comprehensive Essentiality Analysis of the *Mycobacterium tuberculosis* Genome via Saturating Transposon Mutagenesis. *MBio, 8*(1), 1-17.

Deshpande, R. G., Khan, M. B., Bhat, D. A., & Navalkar, R. G. (1997). Isolation of a contact-dependent haemolysin from *Mycobacterium tuberculosis*. *Journal of Medical Microbiology, 46*(3), 233-238.

Dong, W., Huang, J., Li, Y., Tan, Y., Shen, Z., Song, Y., Wang, D., Xiao, S., Chen, H., Fu, Z. F., & Peng, G. (2015). Crystal structural basis for Rv0315, an immunostimulatory antigen and inactive beta-1,3-glucanase of *Mycobacterium tuberculosis*. *Scientific Reports, 5*, 1-14.

Fay, A., Czudnochowski, N., Rock, J. M., Johnson, J. R., Krogan, N. J., Rosenberg, O., & Glickman, M. S. (2019). Two Accessory Proteins Govern MmpL3 Mycolic Acid Transport in Mycobacteria. *MBio, 10*(3), 1-17.

Frees, D., Qazi, S. N., Hill, P. J., & Ingmer, H. (2003). Alternative roles of ClpX and ClpP in *Staphylococcus aureus* stress tolerance and virulence. *Molecular Microbiology, 48*(6), 1565-1578.

Goldstone, D. C., Metcalf, P., & Baker, E. N. (2016). Structure of the ectodomain of the electron transporter Rv2874 from *Mycobacterium tuberculosis* reveals a thioredoxin-like domain combined with a carbohydrate-binding module. *Acta Crystallographica Section D: Structural Biology, 72*(Pt 1), 40-48.

Han, H., & Wilson, A. C. (2013). The two CcdA proteins of *Bacillus anthracis* differentially affect virulence gene expression and sporulation. *Journal of Bacteriology, 195*(23), 5242-5249.

Homolka, S., Niemann, S., Russell, D. G., & Rohde, K. H. (2010). Functional genetic diversity among *Mycobacterium tuberculosis* complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. *The Public Library of Science Pathogens, 6*(7), 1-17.

Miotto, P., Forti, F., Ambrosi, A., Pellin, D., Veiga, D. F., Balazsi, G., Gennaro, M. L., Di Serio, C., Ghisotti, D., & Cirillo, D. M. (2012a). Genome-wide discovery of small RNAs in *Mycobacterium tuberculosis*. *Public Library of Science One, 7*(12), 1-11.

Parida, B. K., Douglas, T., Nino, C., & Dhandayuthapani, S. (2005). Interactions of anti-sigma factor antagonists of *Mycobacterium tuberculosis* in the yeast two-hybrid system. *Tuberculosis, 85*(5-6), 347-355.

Rose, G., Cortes, T., Comas, I., Coscolla, M., Gagneux, S., & Young, D. B. (2013). Mapping of genotype-phenotype diversity among clinical isolates of *Mycobacterium tuberculosis* by sequence-based transcriptional profiling. *Genome Biology and Evolution, 5*(10), 1849-1862.

Salina, E. G., Huszar, S., Zemanova, J., Keruchenko, J., Riabova, O., Kazakova, E., Grigorov, A., Azhikina, T., Kaprelyants, A., Mikusova, K., & Makarov, V. (2018). Copper-related toxicity in replicating and dormant *Mycobacterium tuberculosis* caused by 1-hydroxy-5-R-pyridine-2(1H)-thiones. *Metallomics, 10*(7), 992-1002.

Storz, G., Vogel, J., & Wassarman, K. M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. *Molecular Cell, 43*(6), 880-891.

Wolfe, L. M., Mahaffey, S. B., Kruh, N. A., & Dobos, K. M. (2010). Proteomic definition of the cell wall of *Mycobacterium tuberculosis*. *Journal of Proteome Research, 9*(11), 5816-5826.

World Health Organization. (2021). *Global tuberculosis report 2021* (978-92-4-003702-1). Retrieved from Geneva:

Zhang, L., Hendrickson, R. C., Meikle, V., Lefkowitz, E. J., Ioerger, T. R., & Niederweis, M. (2020). Comprehensive analysis of iron utilization by *Mycobacterium tuberculosis*. *The Public Library of Science Pathogens, 16*(2), 1-27.

APPENDICES

29 April 2020

Miss Divenita Govender (215023332)
School of Life Sciences
Westville

Dear Miss Govender,

Protocol reference number: BREC/00001272/2020
Project title: Investigating the Role of Small RNAs in Transcriptome Regulation of Genetically Diverse Clinical Strains of Mycobacterium tuberculosis
Degree Purposes: Masters

EXPEDITED APPLICATION: APPROVAL LETTER

A sub-committee of the Biomedical Research Ethics Committee has considered and noted your application.

The conditions have been met and the study is given full ethics approval and may begin as from 29 April 2020. Please ensure that outstanding site permissions are obtained and forwarded to BREC for approval before commencing research at a site.

This approval is subject to national and UKZN lockdown regulations and the general BREC circular emailed by the Research Office on 23rd March 2020 and repeatedly since.

This approval is valid for one year from 29 April 2020. To ensure uninterrupted approval of this study beyond the approval expiry date, an application for recertification must be submitted to BREC on the appropriate BREC form 2-3 months before the expiry date.

Any amendments to this study, unless urgently required to ensure safety of participants, must be approved by BREC prior to implementation.

Your acceptance of this approval denotes your compliance with South African National Research Ethics Guidelines (2015), South African National Good Clinical Practice Guidelines (2006) (if applicable) and with UKZN BREC ethics requirements as contained in the UKZN BREC Terms of Reference and Standard Operating Procedures, all available at http://research.ukzn.ac.za/Research-Ethics/Biomedical-Research-Ethics.aspx.

BREC is registered with the South African National Health Research Ethics Council (REC-290408-009). BREC has US Office for Human Research Protections (OHRP) Federal-wide Assurance (FWA 678).

The sub-committee's decision will be noted by a full Committee at its next meeting taking place on 09 June 2020.

Yours sincerely

Prof D Wassenaar
Chair: Biomedical Research Ethics Committee

Biomedical Research Ethics Committee
Chair: Professor D R Wassenaar
UKZN Research Ethics Office Westville Campus, Govan Mbeki Building
Postal Address: Private Bag X54001, Durban 4000
Email: BREC@ukzn.ac.za
Website: http://research.ukzn.ac.za/Research-Ethics/Biomedical-Research-Ethics.aspx

Figure S1: Letter of ethics approval

Table S1: Average colony forming units (cfu) for each clinical strain when exposed to first line and second line anti-tuberculosis drugs and the control plate containing no drugs.

Sample	First line drugs (concentration)				Second line drugs (concentration)			Control
	Isoniazid 0.2	**Isoniazid 1.0**	**Etham -butol 7.5**	**Rifampicin 1.0**	**Kanamycin 5.0**	**Ofloxacin 2.0**	**Strepto-mycin 2.0**	
H37Rv	-	-	-	-	-	-	-	72
Beijing	6	18	18	3	-	-	-	13
F15/LAM 4/KZN	9	10	11	14	12	9	4	32
F11	-	-	-	-	-	-	-	46
Unique	-	-	-	-	-	-	-	57

TMTC: Too many to count; - : no growth

Table S2: Drug resistance profile of clinical strains belonging to the Beijing, F15/LAM4/KZN, F11 and Unique families as well as the H37Rv strain.

Strain	Current resistance profile
H37Rv	Susceptible
Beijing	IRE (MDR)
F15/LAM4/KZN	IRSEOK (XDR)
F11	Susceptible
Unique	Susceptible

I-Isoniazid, R-Rifampicin, O-Ofloxacin, K-Kanamycin, E-Ethambutol, S-Streptomycin, MDR-Multi-drug resistant, XDR- Extensively drug-resistant

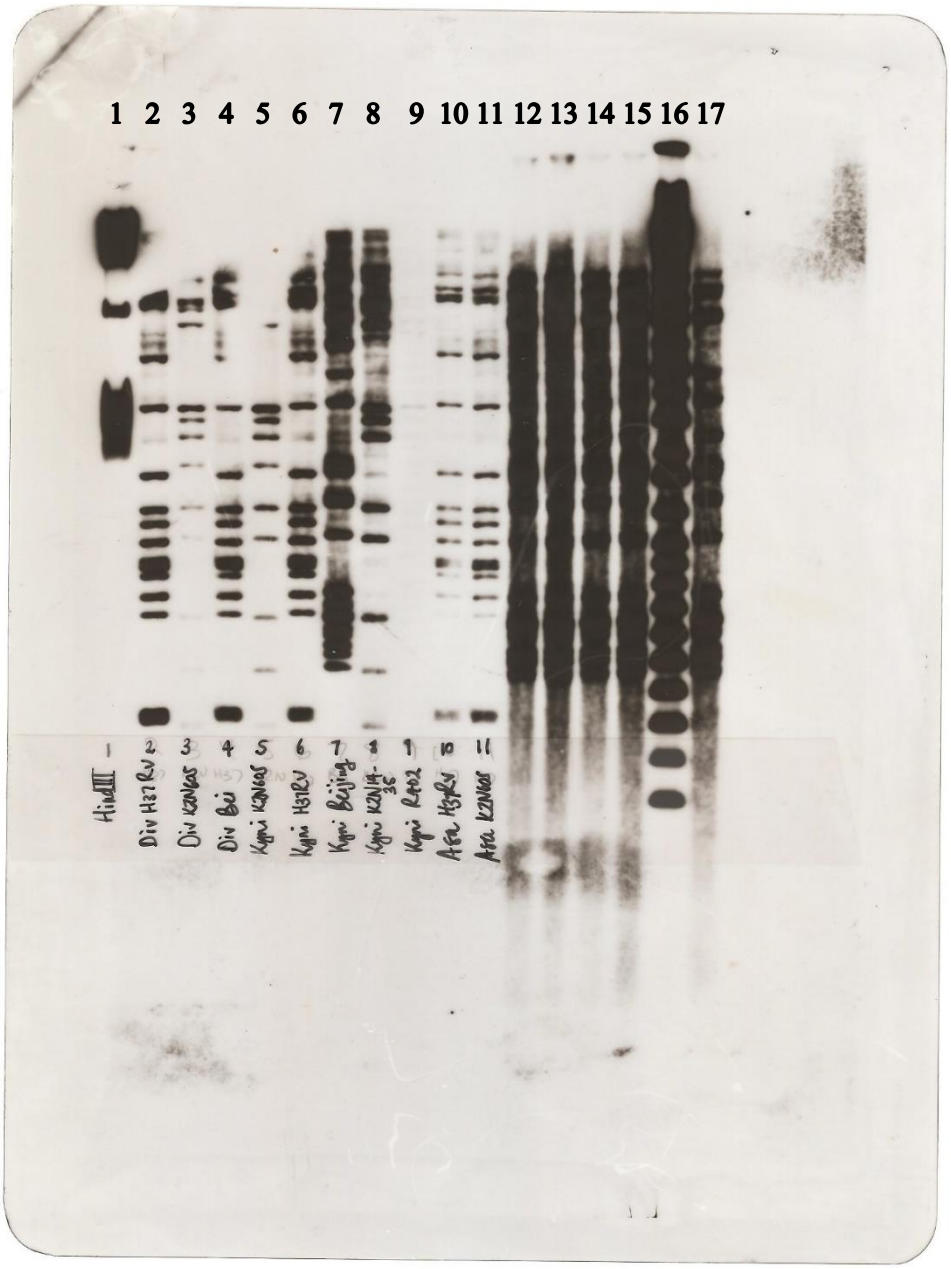
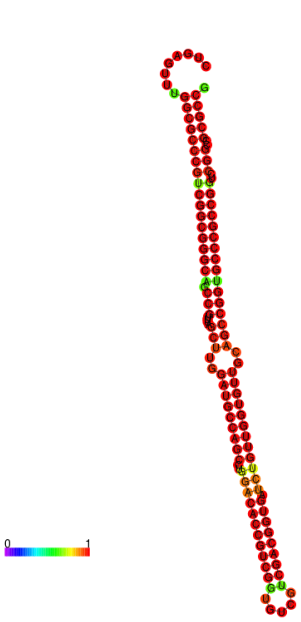
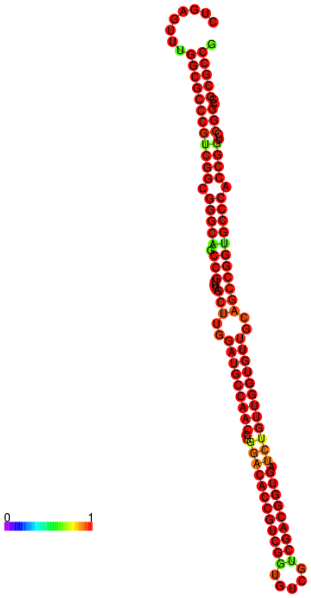
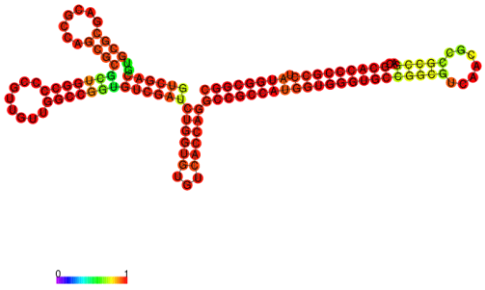
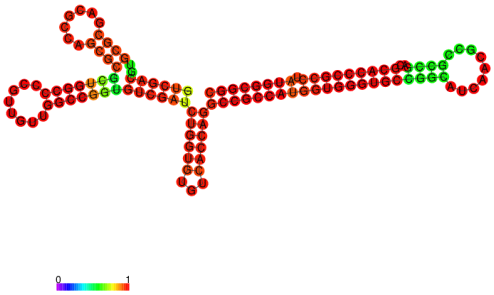
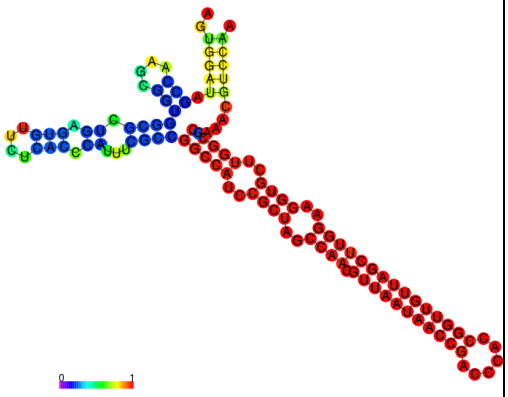
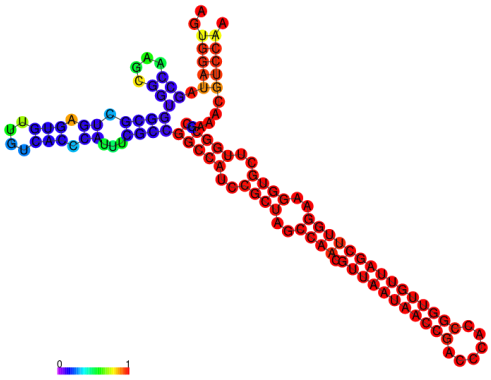
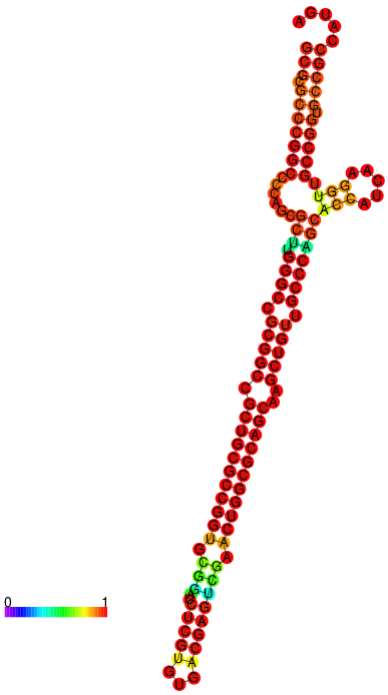
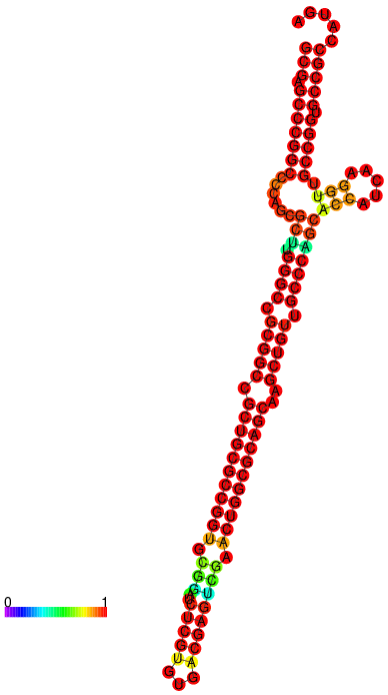


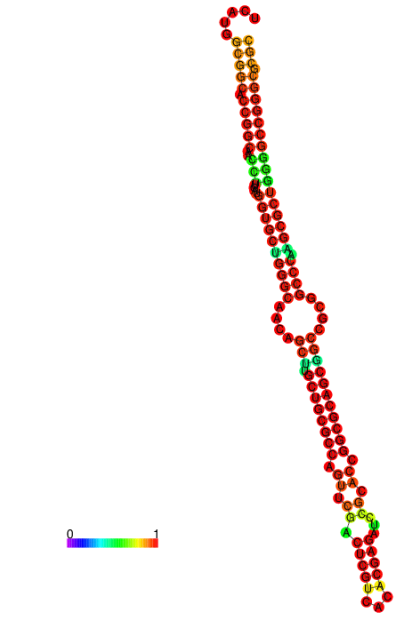
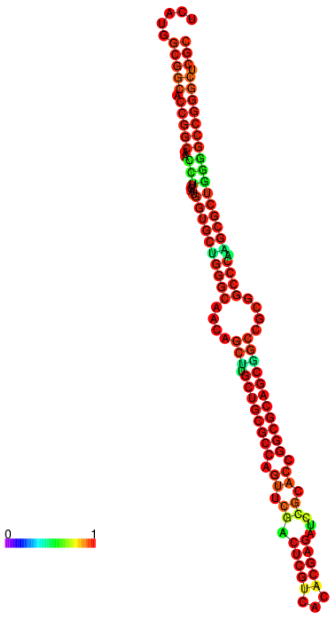
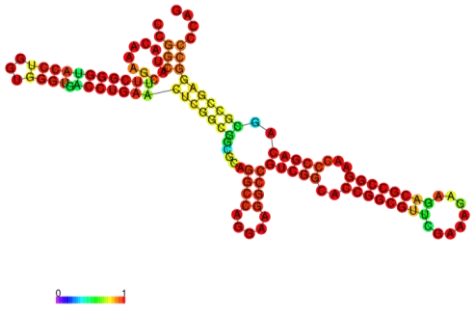
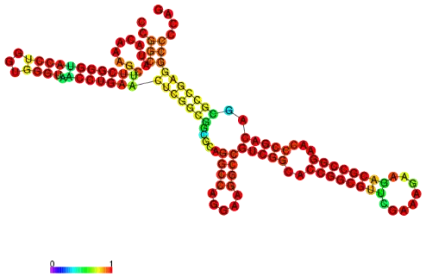
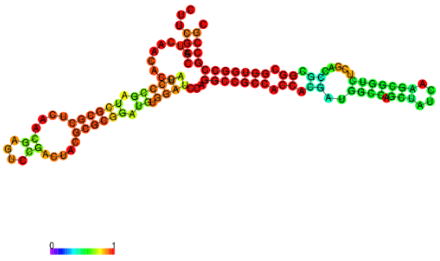
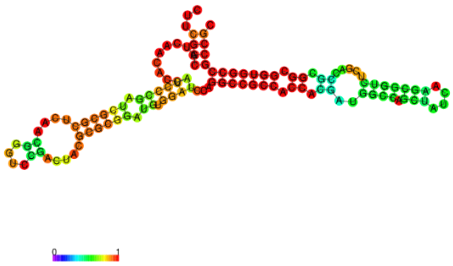
Figure S2: RFLP result for the H37Rv, Beijing and F15/LZM4/KZN. Lane 5 represents F15/LAM4/KZN, lane 6 represents H37Rv, lane 7 represents Beijing and lane 16 represents the molecular marker. All clinical strains presented the correct banding pattern indicating the family classification was correct. Lanes 1, 2, 3, 4, 8, 9, 10-15 and 17 were samples belonging to other studies and were not involved in this research.

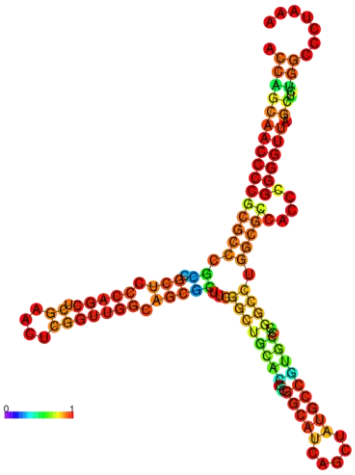
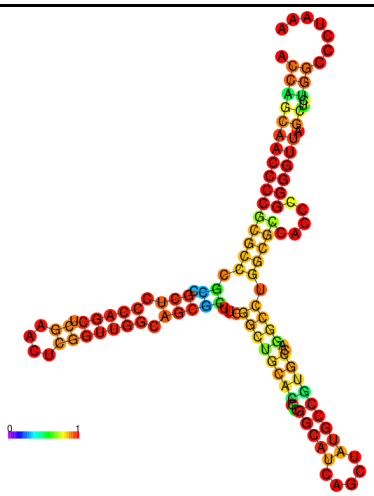
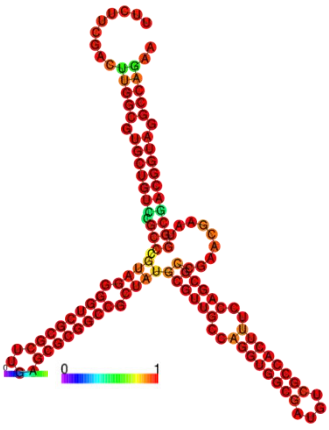
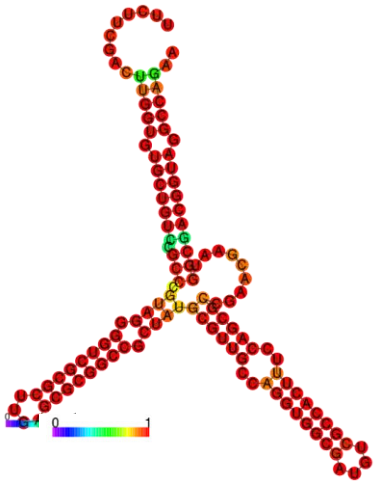
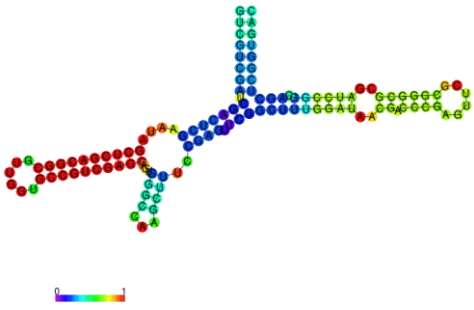
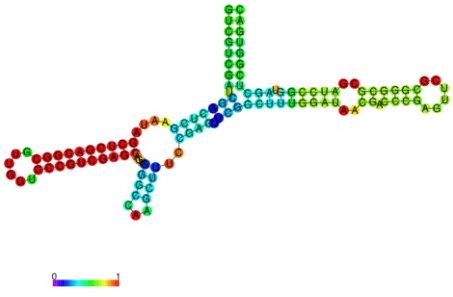
Table S3: RNA concentrations and purity ratios of extracted RNA from clinical strains

Sample	Concentration (ng/µL)	260/280	260/230
H37Rv 1	850	2.01	1.97
H37Rv 2	2027.8	1.98	2.16
H37Rv 3	1341.5	1.97	2.09
F15/LAM4/KZN 1	535,2	1,98	2,01
F15/LAM4/KZN 2	1412,1	2,01	2,15
F15/LAM4/KZN 3	2011	1,98	2,14
Beijing 1	949,8	1,89	2,03
Beijing 2	936,9	1,87	1,98
Beijing 3	2255,1	1,95	2,14
F11 1	1296,1	1,93	2
F11 2	2150,2	1,99	2,01
F11 3	1827,1	1,91	2
Unique 1	1650,2	1,92	2,04
Unique 2	1092,4	2,06	2,12
Unique 3	2135,1	1,91	2

Lineage	Name of sRNA	Control Target	Control Gene	Mutated Target	Mutated Gene
2	16883	Rv0575c		Rv0456c	*echA2*
					
2	29983	Rv2776c		Rv0383c	
					
3	15864	Rv1473A		Rv0260c	

5	11617	Rv2155c	*murD*		Rv0280	*PPE3* 
5	11618	Rv0242c	*fabG4*		Rv1722	

5	14874		Rv2436	*rbsK*	 Rv0514
5	14908		Rv2192c	*trpD*	 Rv2940c *mas*
6	5145		Rv1143	*mcr*	 Rv0083

					
8	11527	Rv1937		Rv2399c	*cysT*
					
8	21006	Rv3019c	*esxR*	Rv0737	
					
	27232	Rv0194		Rv2946c	*pks1*

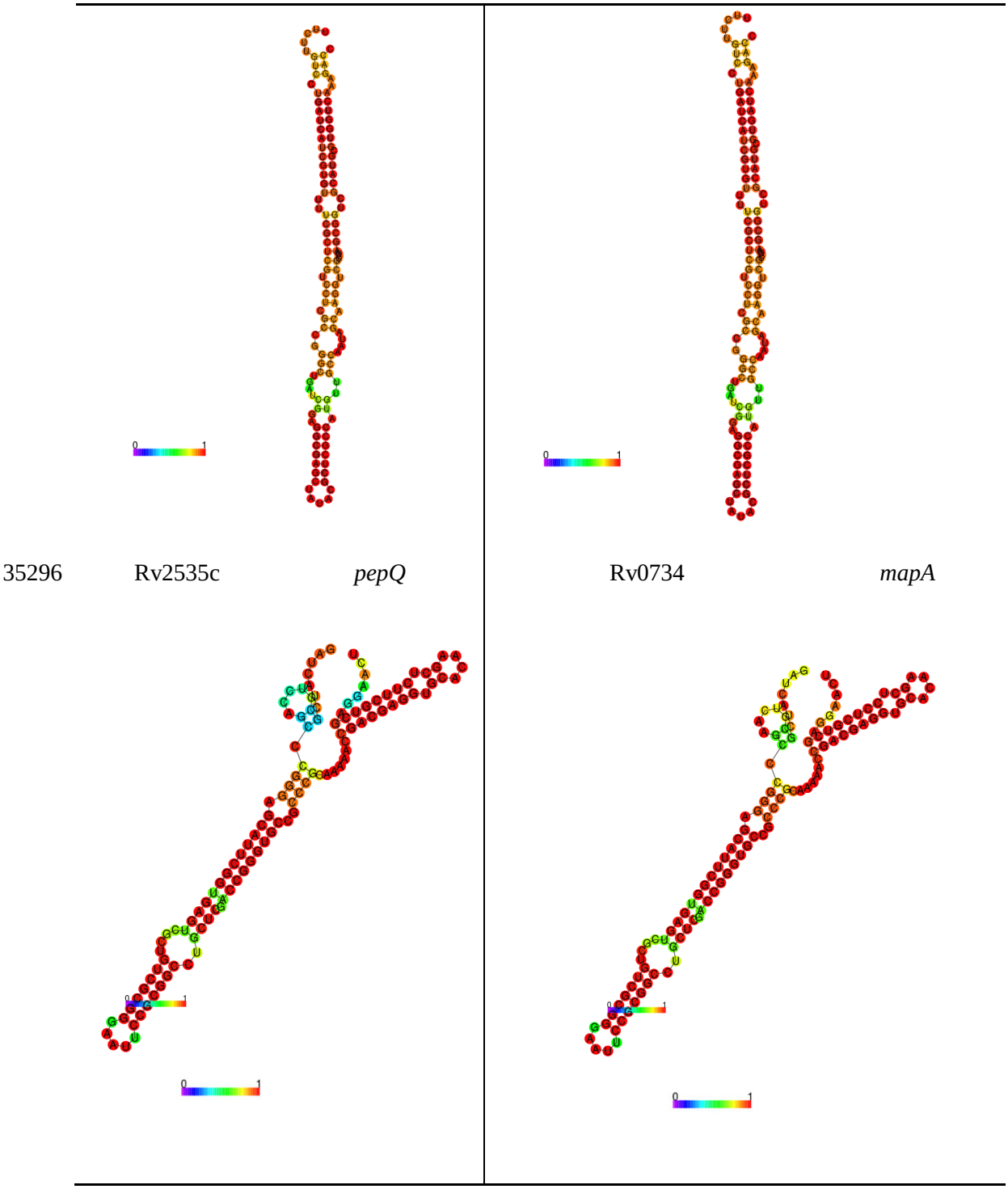


Figure S3: sRNAs containing insignificant structural, lineage specific mutations and the corresponding IntaRNA A) Control and B) Mutated targets and structure predictions for lineages 1- 8.

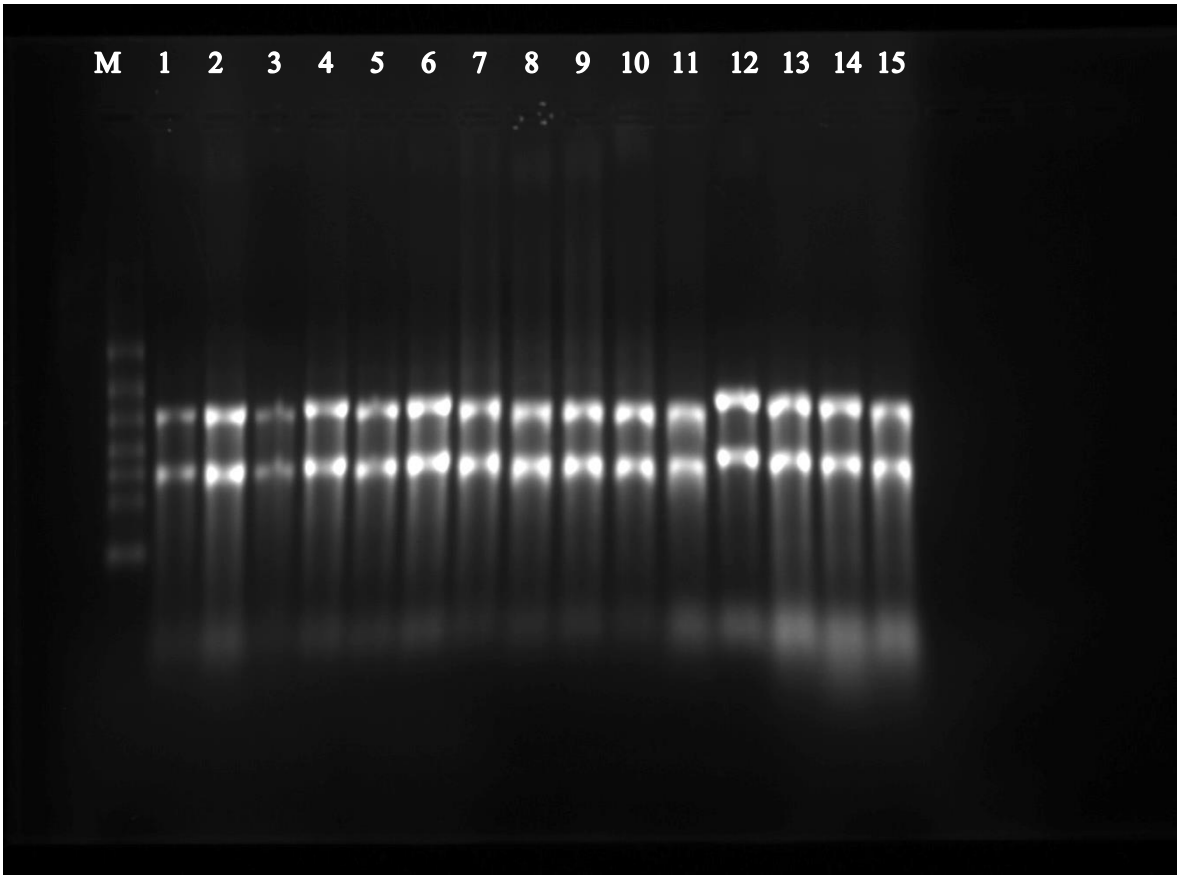


Figure S4: MOPS gel image of extracted RNA from *M. tb* clinical strains run on a 1% agarose gel for 3 hours at 50V. M represents the molecular weight marker, lanes 1,2 and 4 represents the H37Rv samples, lanes 3, 5 and 6 represent the F15/LAM4/KZN samples, lanes 7-9 represents the Beijing samples, lanes 10-12 represents the F11 samples and lanes 13-15 represents the Unique samples

Supplementary File 1: Source code used for sRNA and mRNA sequencing analysis.

Trimming of sRNA sequencing reads:

$ java -jar /<path to trimmomatic>/trimmomatic-0.39.jar PE /home/<path to sample sequence file Read 1>/S1R1.fastqsanger.gz /home/<path to sample sequence file Read 2>/S1R2.fastqsanger.gz /home/<path to output file Read 1>/S1R1_trimmed.fastqsanger.gz /home/<path to output file Read 2>/S1R2_trimmed.fastqsanger.gz ILLUMINACLIP:smRNA_TruSes3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:40

Conversion and sorting of .sam to .bam files:

$ samtools view -bS <sample name>.sam > <sample name>_unsorted.bam

$ samtools sort <sample name>_unsorted.bam -o <sample name>.bam

Assembly, quantification and merging of transcripts:

$ stringtie -G /location of reference gtf file/GCF_000195955.2_ASM19595v2_genomic.gtf -o <sample name>.gtf -l <sample name> <sample name>.bam

$ stringtie --merge -G /location of reference gtf file/GCF_000195955.2_ASM19595v2_genomic.gtf -o merged.gtf mergelist.txt

$ gffcompare -r /location of reference gtf file/GCF_000195955.2_ASM19595v2_genomic.gtf -G -o merged merged.gtf

$ stringtie -e -B -G merged.gtf -o ballgown/<sample name>/<sample name>.gtf <sample name>.bam

Calculation of differential expression and fold changes:

In R,

>library(ballgown)

> library(RSkittleBrewer)

> library(genefilter)

> library(dplyr)

> library(devtools)

>Pheno_data <- read.table("/location of strain comparison file/ballgown_strain.csv", header=TRUE, sep=",", colClasses= rep("character",2))

> bg_strain = ballgown (dataDir = "/location of strain and reference gtf files/ballgown_strain/", samplePattern = "D", pData=Pheno_data)

> bg_strain_filt = subset (bg_strain,"rowVars(texpr(bg_strain)) >1",genomesubset=TRUE)

> results_transcripts_strain = stattest(bg_strain_filt, feature="transcript", covariate="Strains", getFC=TRUE, meas="FPKM")

> results_transcripts_strain =data.frame (geneNames=ballgown::geneNames (bg_strain_filt), geneIDs=ballgown::geneIDs (bg_strain_filt), results_transcripts_strain)

> results_transcripts_strain = arrange (results_transcripts_strain, pval)

> results_genes_strain = arrange (results_transcripts_strain, pval)

> write.csv(results_transcripts_strain,"strain_transcript_results.csv",row.names=FALSE)

> write.csv(results_genes_strain,"strain_gene_results.csv",row.names=FALSE)

> subset(results_transcripts_strain,results_transcripts_strain$pval<0.05)

>write.csv(subset(results_transcripts_strain,results_transcripts_strain$pval<0.05),"strain_transcript_pval_0.05_results.csv",row.names=FALSE)

Obtaining FASTA sequences form a GTF file:

$ /location of gffread/gffread -w <name of output FASTA file>.fa -g /location of reference genomic FASTA file/GCF_000195955.2_ASM19595v2_genomic.fasta merged.gtf

Table S4: Mapping statistics of sRNA sequencing data.

Sample	Number of reads	Unmapped reads (%)	Uniquely mapped reads(%)	Multi-mapped reads(%)	Overall alignment (%)
H37Rv 1	229695	19.25	79.83	0.92	86.21
H37Rv 2	134040	18.02	81.06	0.92	87.48
H37Rv 3	174141	19.68	79.53	0.79	86.30
F15/LAM4/KZN 1	196760	16.73	82.48	0.79	88.69
F15/LAM4/KZN 2	180223	20	79.18	0.82	85.77
F15/LAM4/KZN 3	218883	35.99	63.32	0.69	68.54
Beijing 1	262453	15.98	83.08	0.93	89.32
Beijing 2	282735	17.31	81.81	0.87	88.43
Beijing 3	256899	16.43	82.69	0.88	89.17

Supplementary file 2: Source code used for mRNA sequencing reads trimming.

$ java -jar /<path to trimmomatic >/trimmomatic-0.39.jar PE /home/<path to sample sequence file Read 1>/S1R1.fastq.gz /home/<path to sample sequence file Read 2>/S1R2.fastq.gz /home/<path to output file Read 1>/S1R1_trimmed_paired_1.fastq.gz /home/<path to output file Read 1>/S1R1_trimmed_unpaired_1.fastq.gz /home/<path to output file Read 2>/S1R2_trimmed_paired_1.fastq.gz /home/<path to output file Read 2>/S1R2_trimmed_unpaired_1.fastq.gz ILLUMINACLIP:Nextera_XT_adapters.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

Table S5: Mapping statistics of mRNA sequencing data.

Sample	Number of reads	Unmapped reads (%)	Uniquely mapped reads(%)	Multi-mapped reads(%)	Overall alignment (%)
H37Rv 1	34253098	5.61	88.97	5.42	99.24
H37Rv 2	38572863	47.29	44.99	7.72	76.25
H37Rv 3	59219377	41.05	47.18	11.77	74.94
F15/LAM4/KZN 1	82148307	5.91	84.79	9.31	99.15
F15/LAM4/KZN 2	111908672	5.35	87.11	7.54	99.10
F15/LAM4/KZN 3	80868355	6.08	87.26	6.66	99.25
Beijing 1	51158506	5.08	92.67	2.26	99.19
Beijing 2	35760261	5.53	91.68	2.79	99.20
Beijing 3	35645802	5.40	91.23	3.37	99.21

Table S6: Primer sequences used for qualitative confirmation of the sRNAs and mRNAs from the small RNA and mRNA sequencing data

Name	Forward / Reverse (5'-3')	Sequence	Expected product size (bp)
MSTRG.34.1	Forward	CTCGCTAACCGCCAACTCA	276
	Reverse	CGAAATCCGTGACCTTTGCC	
MSTRG.40.1	Forward	CGCTCTTATGCGATTACGCC	
	Reverse	GGTCTGCGATGAGGCATTTTC	372
MSTRG.53.1	Forward	TGGACCGAGAACGATTCAGC	
	Reverse	ATCGTGGTAGGGCATGCAAA	242
MSTRG.26.1	Forward	TGGGCGCTAGTAACACGTAG	
	Reverse	ATGGGCTCAGTTCGAGGAAG	234
Rvnt01	Forward	GCTGATAACGAAGAGGTCGGA	
	Reverse	AATGCGAACCTCACCGAACG3	83
Rvnt02	Forward	GCAGTTGTGTCCTTTCGCAG	
	Reverse	TGAGTGAGCAGGTGGAAACC	557
MSTRG1.7	Forward	ACTCACCTTCACCTGCACAC	
	Reverse	GAGCATTCCCAGAGGCTACC	236
16S	Forward	TGTTTGGAGAGTTTGATCCTGGC	
	Reverse	AAAGGAGGTGATCCAGCCG	1532

Table S7: Primer efficiencies and R^2 values for the designed primer pairs.

Gene	Primer efficiency (%)	R^2
MSTRG.34.1	180,08	0,925
MSTRG.40.1	75,04	0,745
MSTRG.53.1	161.21	0,964
MSTRG.26.1	122,80	0,954
Rvnt01	127.21	0,993
Rvnt02	86,29	0,792
MSTRG1.7	67,69	0,500
16S	86,72	0,923

Table S8: Raw qPCR Ct values for selected sRNA and mRNA genes in clinical strains and the fold change difference when compared to the 16S housekeeping gene.

Gene	Sample	Ct1	Ct2	Average Ct	ΔCt	ΔΔ Ct	Fold change
16S	H37Rv	11,81		11,8091126			
	H37Rv	14,26	12,00	13,1310047			
	H37Rv	13,10	12,37	12,7319501			
	Beijing		11,83	11,828459			
	Beijing		10,96	10,9627258			
	Beijing	9,91	10,96	10,4347579			
	F15/LAM4/KZN	14,00	13,95	13,9765028			
	F15/LAM4/KZN		13,95	13,9509822			
	F15/LAM4/KZN	11,94	14,27	13,106686			
	Unique		11,94	11,9428457			
	Unique		11,30	11,2980746			
	Unique	10,65		10,6523161			
	F11	13,70	10,79	12,2457768			
	F11	11,64		11,6384388			
	F11	11,64		11,6384388			

Gene	Sample	Ct1	Ct2	Average Ct	ΔCt	ΔΔ Ct	Fold change
MSTRG.26.1	H37Rv	26,78	27,75	27,26	15,4540063	6,03566258	
	H37Rv	23,86	23,65	23,75	10,6223095	1,20396571	1.20
	H37Rv	2,77	27,05	14,91	2,17871545	-7,2396283	
	Beijing	17,92	17,47	17,69	5,86643164	-3,5519121	
	Beijing	25,85	24,14	24,99	14,0305087	4,612165	0.43
	Beijing	20,36	20,22	20,29	9,85296712	0,43462337	
	F15/LAM4/KZN	23,94	25,13	24,54	10,5611731	1,14282931	
	F15/LAM4/KZN	21,51	22,72	22,12	8,16502024	-1,2533235	4.37
	F15/LAM4/KZN	26,82	26,98	26,90	13,7932152	4,37487143	
	Unique	19,93	11,65	15,79	3,8449967	-5,5733471	
	Unique	27,78		27,78	16,4789651	7,06062136	4.76
	Unique	24,86	24,79	24,83	14,1742884	4,75594469	
	F11	2,49	21,42	11,96	-0,2903741	-9,7087178	
	F11	28,55	28,81	28,68	17,0417157	7,62337193	4.42
	F11	26,17	24,78	25,48	13,8378288	4,41948508	
MSTRG.34.1	H37Rv	21,15	20,28	20,71	8,90556175	5,41879782	
	H37Rv	2,14	17,74	9,94	-3,1889744	-6,6757383	1.23
	H37Rv	15,94	19,01	17,48	4,74370441	1,25694048	
	Beijing	14,58	14,07	14,32	2,49634946	-0,9904145	
	Beijing	20,71	20,82	20,76	9,80186003	6,3150961	4.70
	Beijing	18,87	18,37	18,62	8,18629882	4,69953489	
	F15/LAM4/KZN	20,33	20,22	20,27	6,29476304	2,80799911	
	F15/LAM4/KZN	18,19	18,86	18,53	4,57536761	1,08860368	2.81
	F15/LAM4/KZN	22,44	22,61	22,53	9,41909271	5,93232878	
	Unique	16,89	16,28	16,59	4,64417498	1,15741105	
	Unique	20,74	11,35	16,04	4,74563673	1,2588728	1.26
	Unique	20,72	10,37	15,54	4,89107163	1,4043077	
	F11	18,15	18,18	18,16	5,91733913	2,4305752	
	F11	21,27	21,09	21,18	9,5420349	6,05527097	5.63
	F11	20,80	20,70	20,75	9,11415398	5,62739005	
MSTRG.40.1	H37Rv	13,92	9,98	11,95	0,14235913	0,3506503	
	H37Rv		12,55	12,55	-0,5831436	-0,3748525	0.35
	H37Rv	12,55		12,55	-0,184089	0,02420216	

Gene	Sample	Ct1	Ct2	Average Ct	ΔCt	ΔΔ Ct	Fold change
	Beijing	34,88	10,42	22,65	10,8227054	11,0309966	25.06
	Beijing	35,82		35,82	24,8564544	25,0647456	
	Beijing	35,63		35,63	25,197202	25,4054932	
	F15/LAM4/KZN	35,40		35,40	21,4187521	21,6270433	
	F15/LAM4/KZN	35,40	35,12	35,26	21,3089863	21,5172775	21.63
	F15/LAM4/KZN		35,40	35,40	22,2885688	22,49686	
	Unique		14,70	14,70	2,75278159	2,96107276	
	Unique		13,02	13,02	1,71843378	1,92672495	1.93
	Unique		9,94	9,94	-0,7127596	-0,5044684	
	F11	15,78		15,78	3,53746733	3,7457585	
	F11	24,61	34,96	29,79	18,1487375	18,3570287	18.36
	F11		35,70	35,70	24,0594463	24,2677374	
MSTRG.53.1	H37Rv	19,09	19,03	19,06	7,25257256	2,07693343	
	H37Rv	16,48	15,91	16,19	3,06183801	-2,1138011	0.04
	H37Rv	17,75	18,14	17,94	5,21250681	0,03686768	
	Beijing	14,27	14,05	14,16	2,32829462	-2,8473445	
	Beijing	19,15	19,54	19,34	8,37930526	3,20366613	3.20
	Beijing	17,06	16,98	17,02	6,58465787	1,40901874	
	F15/LAM4/KZN	20,56	19,44	20,00	6,02344607	0,84780695	
	F15/LAM4/KZN	15,85	18,02	16,93	2,98231445	-2,1933247	1.61
	F15/LAM4/KZN	19,79	20,01	19,90	6,79044409	1,61480496	
	Unique	15,31	16,55	15,93	3,98866643	-1,1869727	
	Unique	19,67	8,51	14,09	2,79024127	-2,3853979	-1.38
	Unique	18,94	9,95	14,45	3,7973494	-1,3782897	
	F11	16,86	17,85	17,35	5,10759072	-0,0680484	
	F11	20,43	20,26	20,35	8,70943461	3,53379548	1.96
	F11	19,18	18,37	18,77	7,1350283	1,95938917	
Rvnt01	H37Rv	22,48	19,14	20,81	9,00288683	4,49867582	
	H37Rv	12,51	17,79	15,15	2,02016662	-2,4840444	-2.10
	H37Rv	11,68	18,76	15,22	2,48957959	-2,0146314	
	Beijing	16,05	16,45	16,25	4,42349977	-0,0807112	
	Beijing	19,89	19,58	19,74	8,77464478	4,27043377	4.27
	Beijing	18,91	19,06	18,98	8,55005808	4,04584707	

Gene	Sample	Ct1	Ct2	Average Ct	ΔCt	ΔΔ Ct	Fold change
	F15/LAM4/KZN	19,28	19,22	19,25	5,27376984	0,76955882	
	F15/LAM4/KZN	21,42	17,94	19,68	5,73113837	1,22692736	2.26
	F15/LAM4/KZN	19,82	19,92	19,87	6,7617404	2,25752938	
	Unique	16,98	17,98	17,48	5,53695245	1,03274143	
	Unique	19,38	12,23	15,80	4,50664618	0,00243517	1.03
	Unique	20,29	22,22	21,26	10,6043419	6,10013087	
	F11	11,42	19,76	15,59	3,34297019	-1,1612408	
	F11	16,71	19,29	18,00	6,36494947	1,86073846	1.86
	F11	11,71	18,58	15,15	3,50849243	-0,9957186	
Rvnt02	H37Rv	12,25	35,17	23,71	11,8981208	1,44352909	
	H37Rv	10,85	34,15	22,50	9,36950308	-1,0850886	1.44
	H37Rv	11,22	34,44	22,83	10,0961513	-0,3584405	
	Beijing	32,81	32,37	32,59	20,7636127	10,3090209	
	Beijing	32,48	30,55	31,51	20,5502144	10,0956227	10.31
	Beijing	32,81	34,17	33,49	23,0553713	12,6007795	
	F15/LAM4/KZN	33,40	34,01	33,71	19,7314598	9,27686807	
	F15/LAM4/KZN	30,53	30,54	30,54	16,5881436	6,13355185	6.13
	F15/LAM4/KZN	33,60	33,33	33,46	20,3556922	9,90110049	
	Unique	32,10	15,22	23,66	11,7166035	1,26201181	
	Unique	30,25	18,07	24,16	12,8617681	2,40717634	2.41
	Unique	33,18	13,91	23,55	12,8940734	2,43948171	
	F11	10,95		10,95	-1,2936467	-11,748238	
	F11	15,65	31,61	23,63	11,9935004	1,53890866	1.54
	F11	22,23	32,02	27,13	15,4869471	5,03235539	
MSTRG.1.7	H37Rv	19,11	20,31	19,71	7,90042405	4,59494747	
	H37Rv	7,63	18,69	13,16	0,02536328	-3,2801133	4.59
	H37Rv	8,28	21,17	14,72	1,9906424	-1,3148342	
	Beijing	31,94	10,38	21,16	9,33405798	6,02858141	
	Beijing	31,35	31,46	31,40	20,4403189	17,1348423	17.13
	Beijing	30,05	32,82	31,44	21,0030588	17,6975822	
	F15/LAM4/KZN	20,55	20,33	20,44	6,46424858	3,15877201	
	F15/LAM4/KZN	17,85	19,00	18,42	4,47256235	1,16708577	6.42
	F15/LAM4/KZN	23,65	22,02	22,84	9,72913262	6,42365605	

Gene	Sample	Ct1	Ct2	Average Ct	ΔCt	ΔΔ Ct	Fold change
	Unique	17,01	17,14	17,08	5,13299912	1,82752254	
	Unique	21,25	9,41	15,33	4,03008709	0,72461052	0.72
	Unique	20,19	8,42	14,31	3,65305786	0,34758128	
	F11	11,14	18,15	14,65	2,39967222	-0,9058044	
	F11	8,39	20,38	14,38	2,7449069	-0,5605697	-0.905
	F11	9,00	18,39	13,69	2,05640012	-1,2490765	

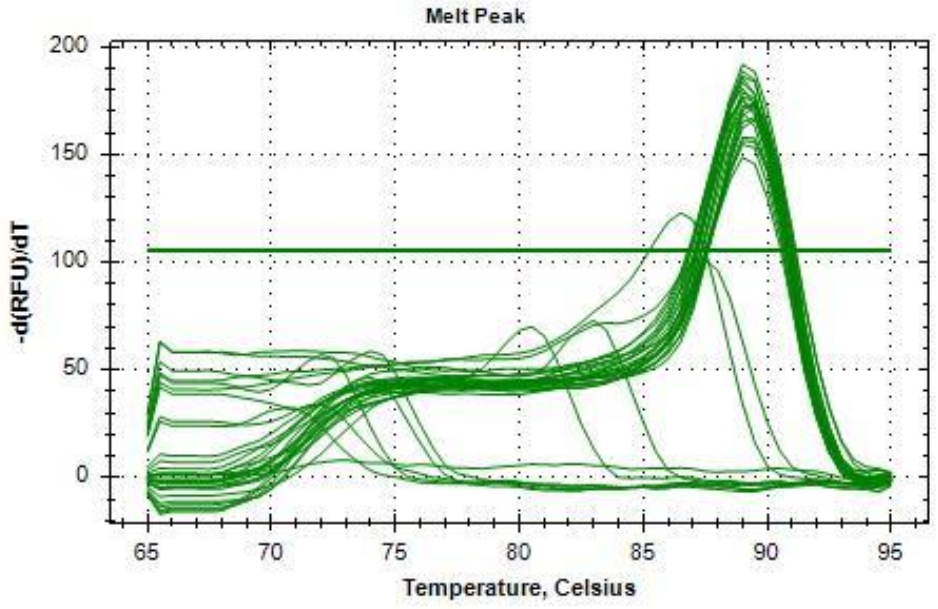


Figure S5: Melt peak of 16S qPCR data for clinical strains.

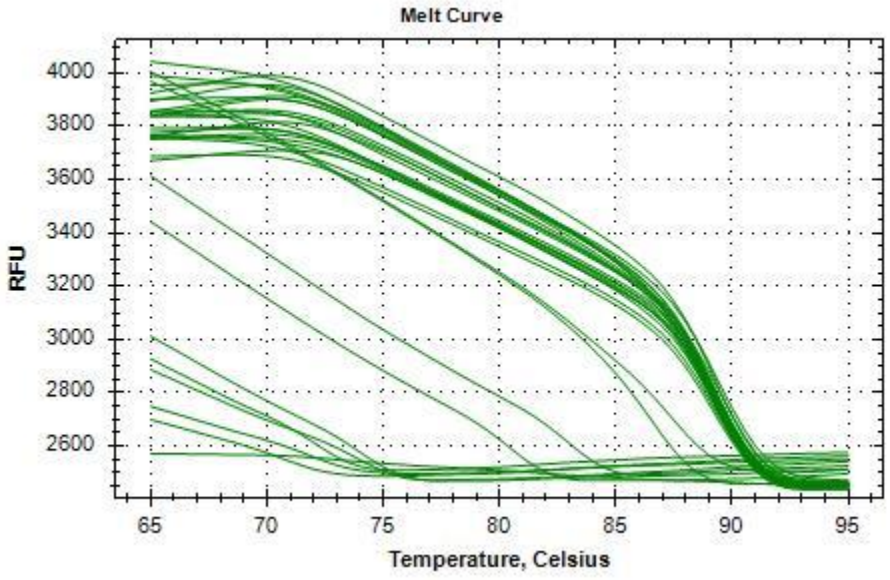


Figure S6: Melt curve of 16S qPCR data for clinical strains.

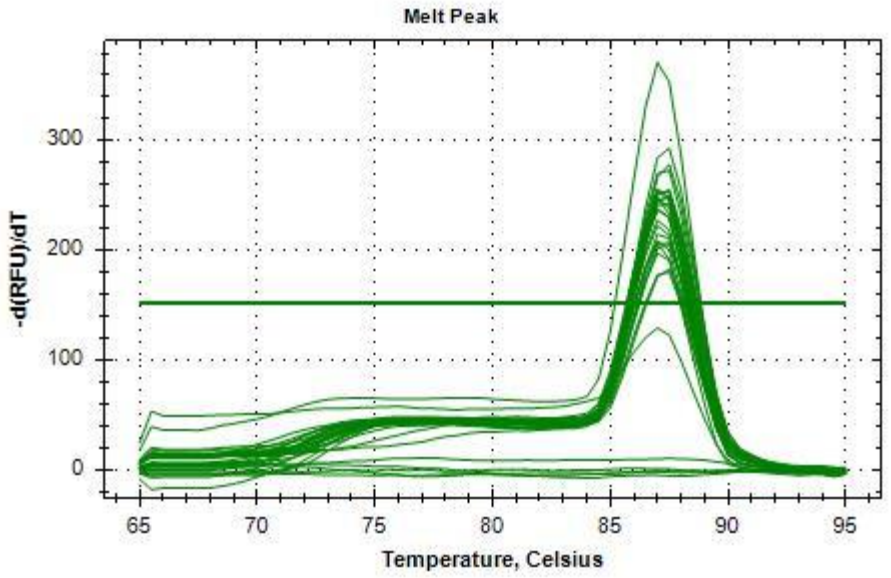


Figure S7: Melt peak of MSTRG.26.1 qPCR data for clinical strains.

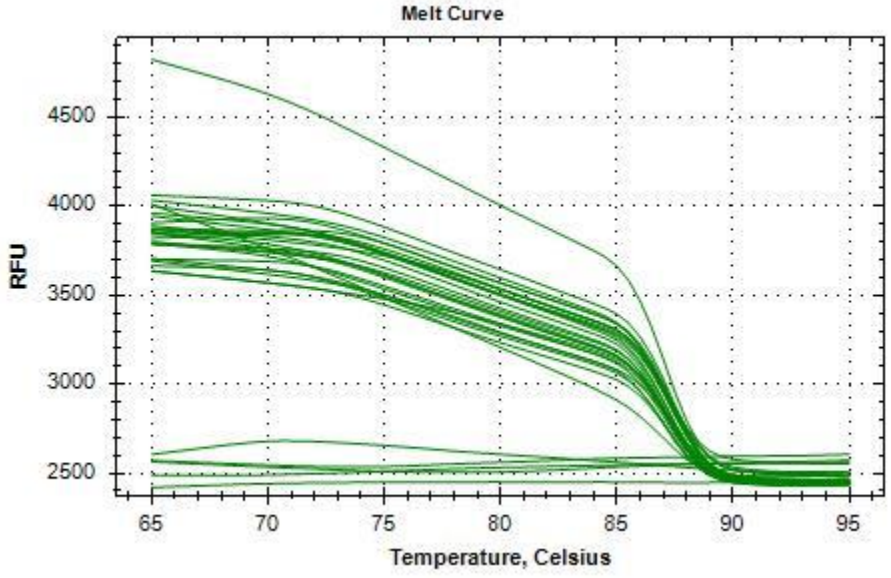


Figure S8: Melt curve of MSTRG.26.1 qPCR data for clinical strains.

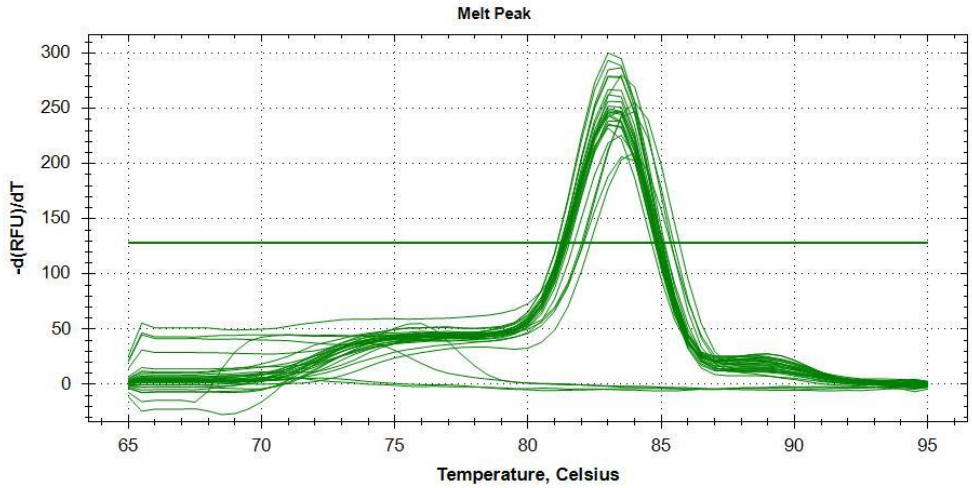


Figure S9: Melt peak of MSTRG.34.1 qPCR data for clinical strains.

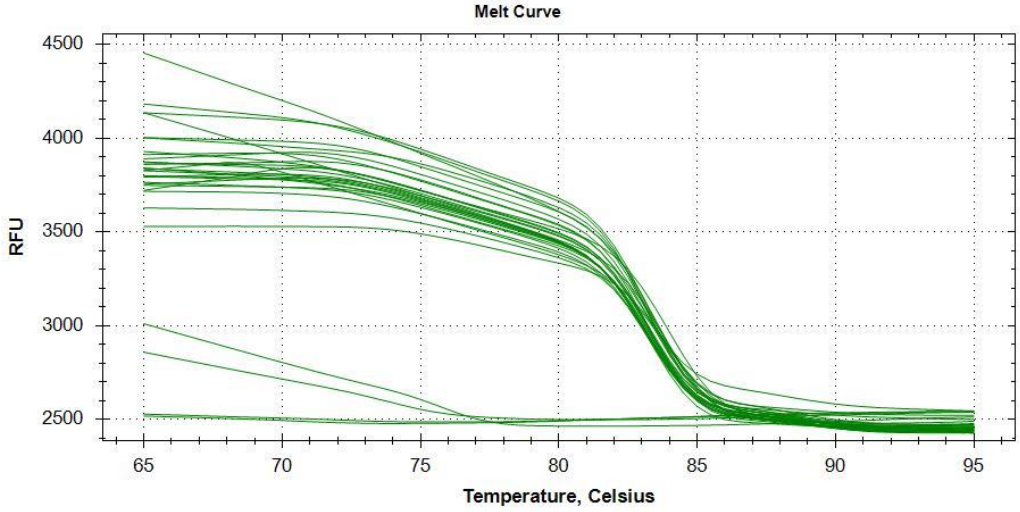


Figure S10: Melt curve of MSTRG.34.1 qPCR data for clinical strains.

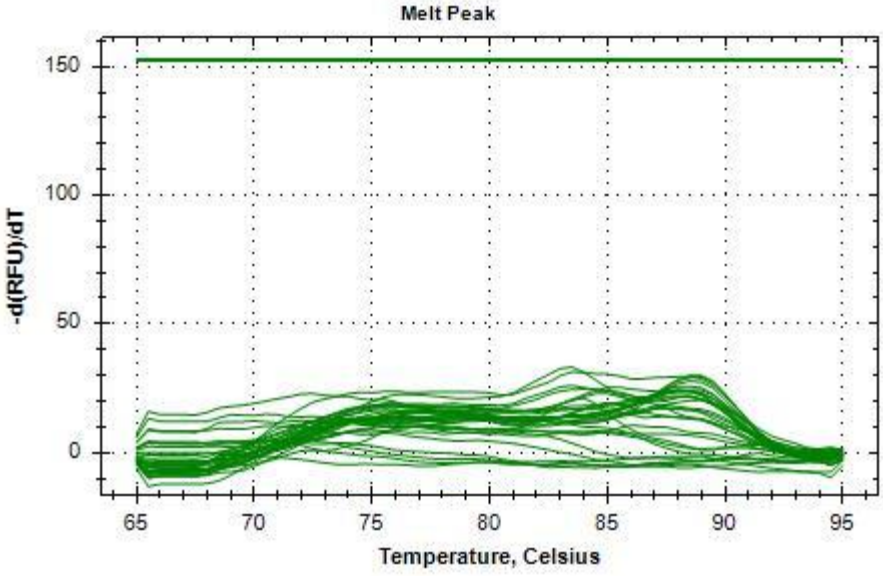


Figure S11: Melt peak of MSTRG.40.1 qPCR data for clinical strain.

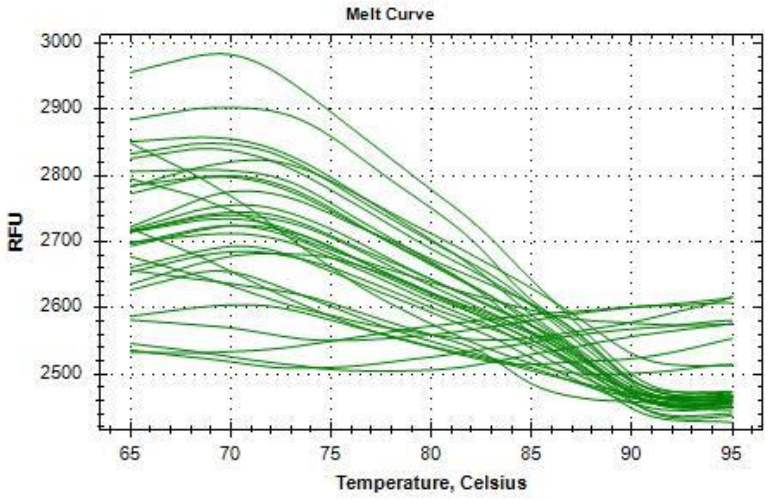


Figure S12: Melt curve of MSTRG.40.1 qPCR data for clinical strain.

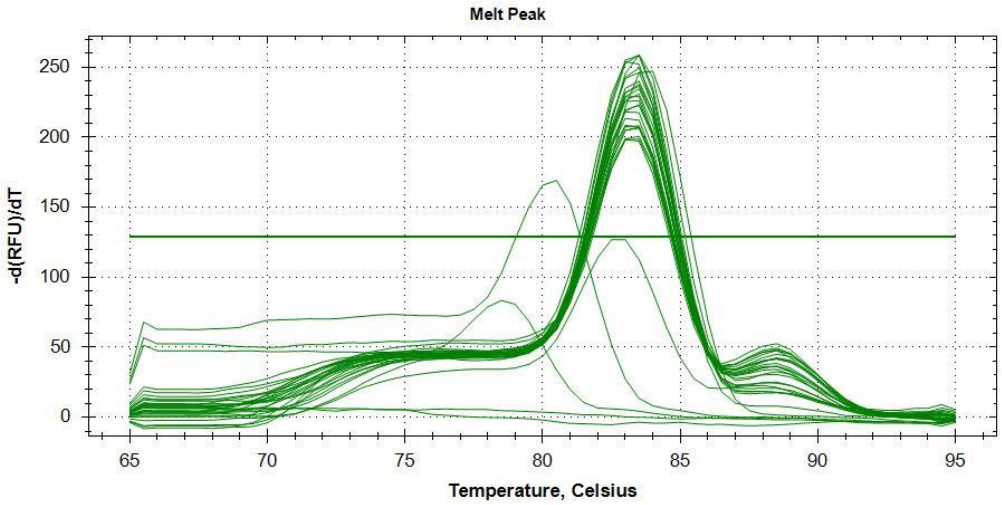


Figure S13: Melt peak of MSTRG.53.1 qPCR data for clinical strain.

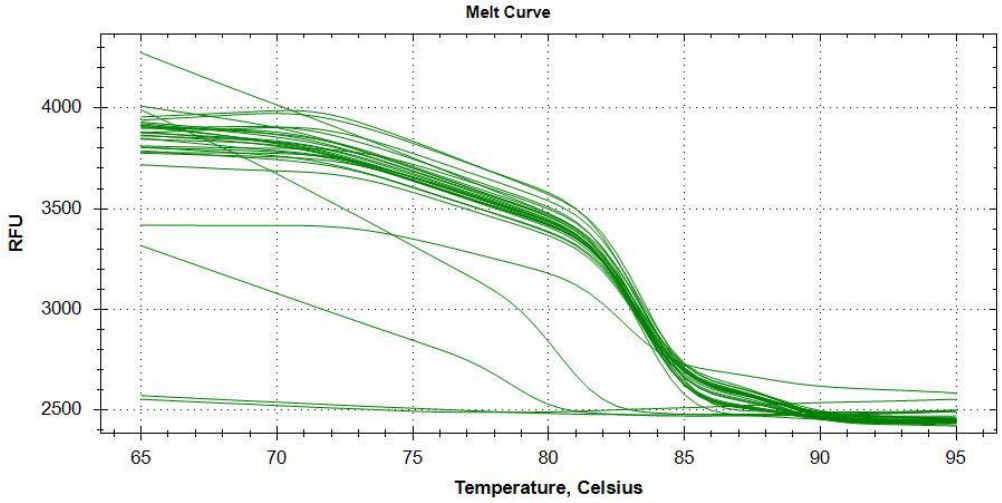


Figure S14: Melt curve of MSTRG.53.1 qPCR data for clinical strain.

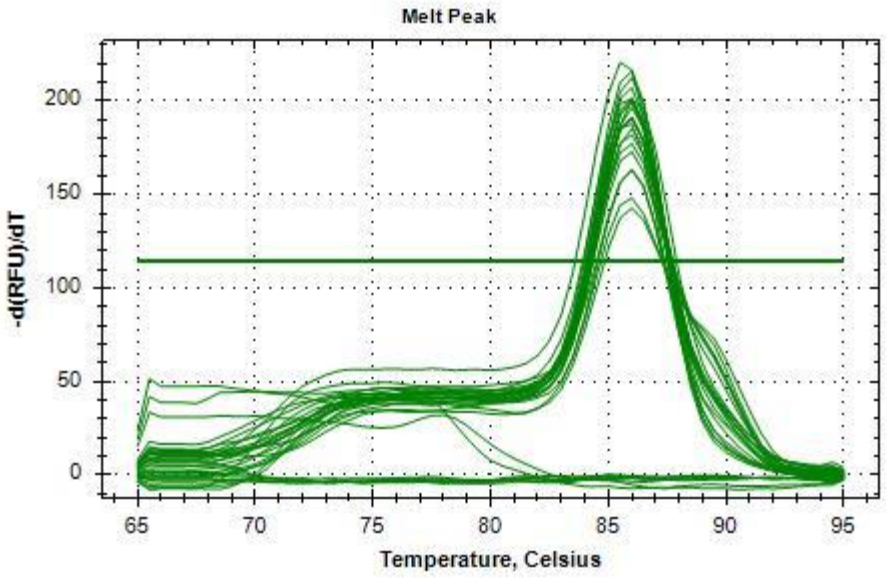


Figure S15: Melt peak of Rvnt01 qPCR data for clinical strain.

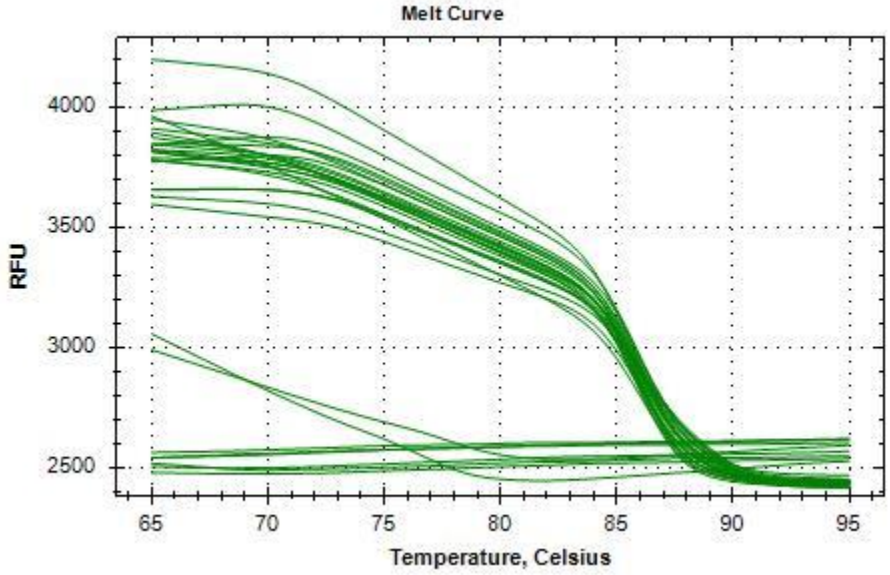


Figure S16: Melt curve of Rvnt01 qPCR data for clinical strain.

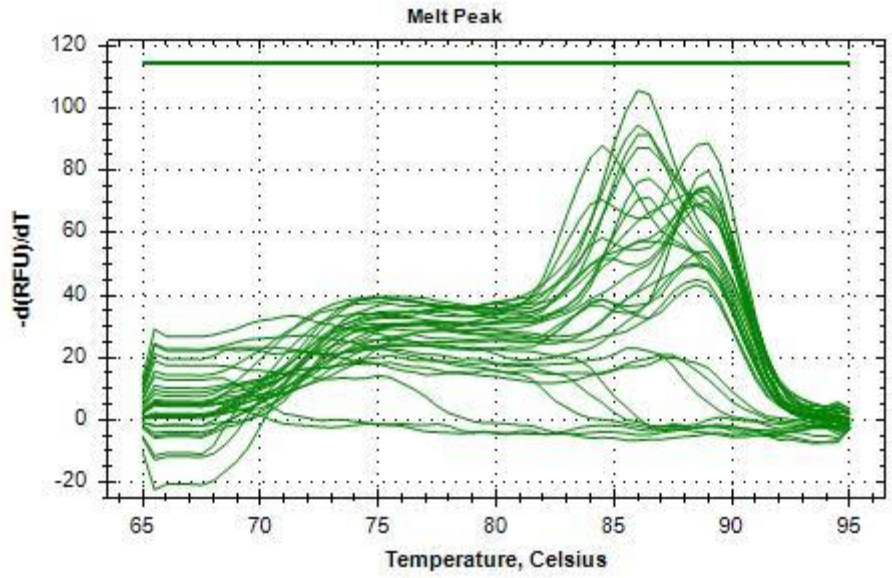


Figure S17: Melt peak of Rvnt02 qPCR data for clinical strain.

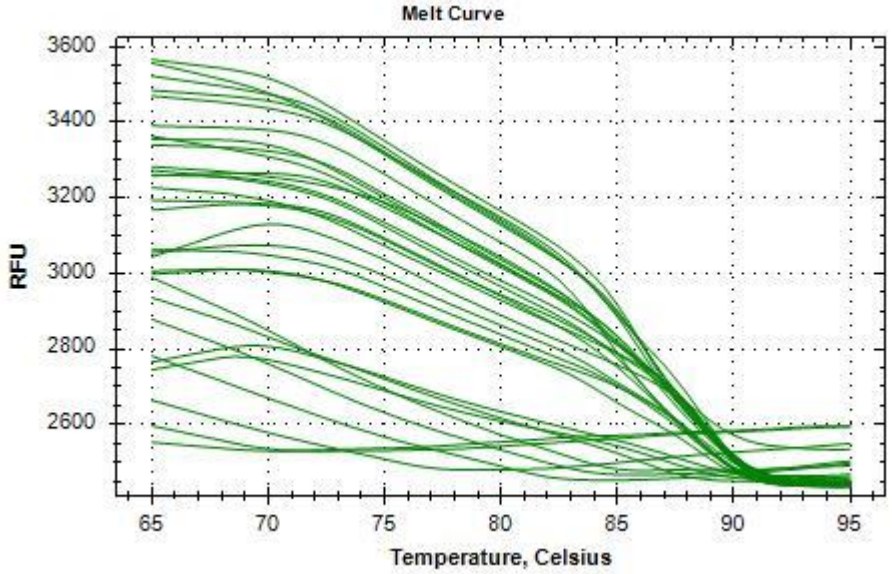


Figure S18: Melt curve of Rvnt02 qPCR data for clinical strain.

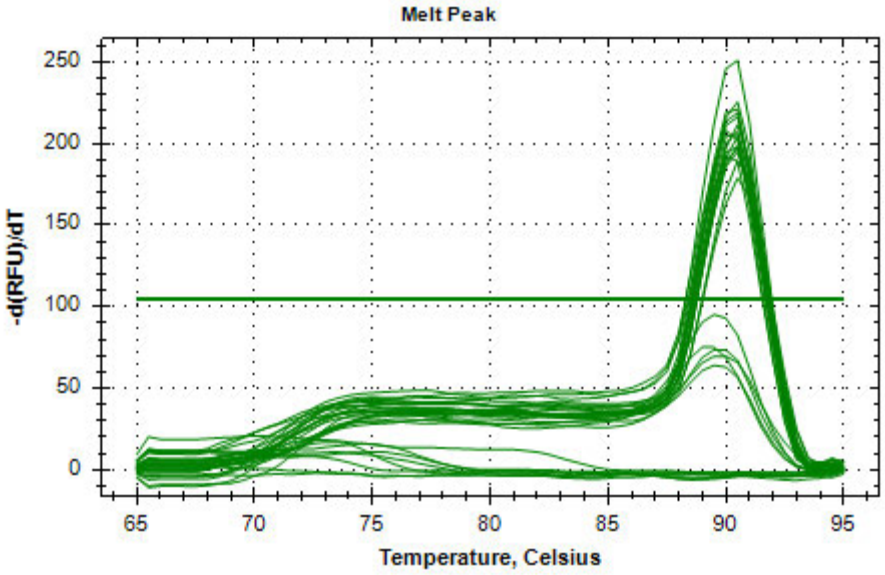


Figure S19: Melt peak of MSTRG.1.7 qPCR data for clinical strain.

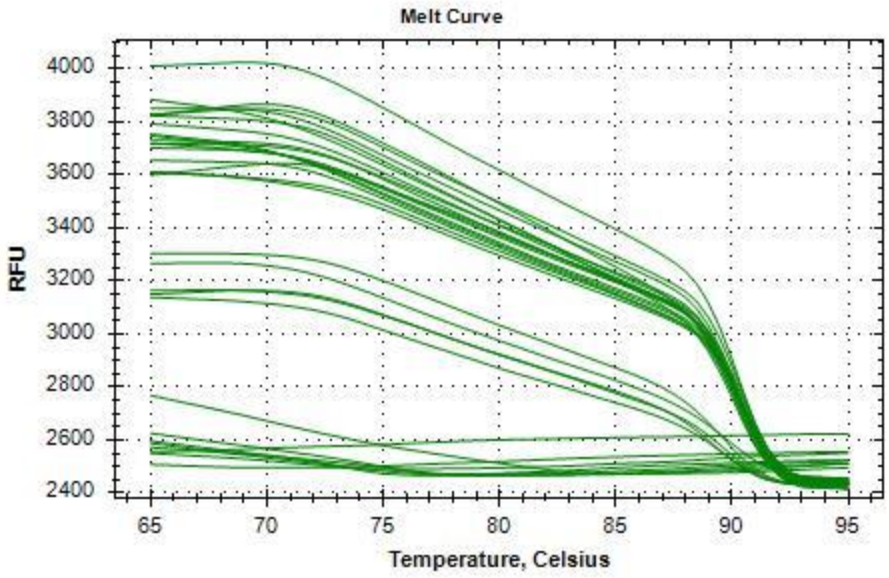


Figure S20: Melt curve of MSTRG.1.7 qPCR data for clinical strain.