

BIROn - Birkbeck Institutional Research Online

Enabling Open Access to Birkbeck's Research Degree output

Using global genomic and transcriptomic approaches to explore host specificity and the role of non-coding RNA in the Mycobacterium tuberculosis complex

<https://eprints.bbk.ac.uk/id/eprint/55008/>

Version: Full Version

Citation: Stiens, Jennifer Jane (2025) Using global genomic and transcriptomic approaches to explore host specificity and the role of non-coding RNA in the Mycobacterium tuberculosis complex. [Thesis] (Unpublished)

© 2020 The Author(s)

All material available through BIROn is protected by intellectual property law, including copyright law.

Any use made of the contents should comply with the relevant law.

[Deposit Guide](#)
Contact: [email](#)

Using global genomic and transcriptomic approaches to explore host specificity and the role of non-coding RNA in the *Mycobacterium tuberculosis* complex



A thesis submitted to:
Birkbeck University of London

For the degree of:
Doctor of Philosophy
27 September 2024
Jennifer Stiens

The copyright in this thesis is owned by the author. Any quotation from the thesis or use of any of the information contained in it must acknowledge this thesis as the source of the quotation or information

DECLARATION

I hereby declare that this thesis and the work presented herein is the result of my own efforts. Any ideas, data, images or text resulting from the work of others (whether published or unpublished) are fully identified as such within the work and attributed to their originator in the text and references. Where I have relied upon data produced in collaboration with others, this has been indicated in the text.

ABSTRACT

Human-adapted *Mycobacterium tuberculosis* and animal-adapted *Mycobacterium bovis* are members of the *Mycobacterium tuberculosis* complex (MTBC) with nearly identical genomes but different host preferences. In this thesis, the basis for these phenotypic differences is explored using global assays. Transposon insertion sequencing (tn-seq) is used to identify the essential genes for both species in *in vitro* growth in identical conditions. Differences in the importance of genes involved in nitrogen and sulfur assimilation, and lipid and amino acid metabolism, highlight the evolution of metabolic adjustments made to exploit a microecological niche in a particular mammalian host. *M. tuberculosis* and *M. bovis* are reported to be differently sensitive to oxidative stress and tn-seq was also used to determine conditional gene requirements for *M. bovis* under oxidative stress with menadione treatment. A fatty-acid ligase, *fadD30* and an iron transport regulator, *irtA*, were only conditionally essential with oxidative stress in *M. bovis* but essential for *in vitro* growth in *M. tuberculosis*. Regulation of these differently-required genes may be influenced by the expression of regions outside of protein-coding genes. Transcriptomic studies in both genomes have revealed pervasive non-coding transcription, increasing in stress conditions. Beginning with the best studied member of the MTBC, this thesis begins to address this question by presenting a whole genome co-expression network analysis of *M. tuberculosis* RNA-seq data to infer the function of these transcripts. This valuable resource can be used by mycobacterial researchers to find potential regulators among predicted non-coding transcripts expressed in a multitude of experimental conditions. Finally, one of these candidate transcripts is investigated using CRISPR inhibition to silence its expression. Antisense-phoR is located within an important regulatory operon, *phoPR*, which functions at the host-pathogen interface of the MTBC. Antisense silencing impacted sense transcript abundance; perhaps an indication that the transcript plays a role in stabilising *phoR* mRNA.

ACKNOWLEDGEMENTS

I have thoroughly enjoyed the last four years spent on this PhD project; undoubtedly because I have been incredibly lucky in my PhD supervisors. I am forever indebted to my supervisor, the brilliant and creative Dr. Irilenia Nobeli, who, over and over again, has encouraged me to push myself and made me believe I was capable of doing more than I ever thought I could (starting with the MSc!). And my 'second' supervisor in name only, Dr. Sharon Kendall, who has been incredibly generous with her time and resources, answered about a million questions, and most importantly, convinced me to get over my fears and get back to the bench. Thank you both for all your support and, especially, for your friendship. I am also very grateful to my thesis committee chair, Dr. Andrew Osborne, for his guidance, positivity and valuable advice.

Birkbeck is a fantastic place for pursuing your dreams--and for research. I am grateful there exists such a place with flexible learning for working people looking to restart a career, whatever their age. I've never felt anything but validation and support from anyone in the Birkbeck community, despite being a bit older than the typical student. Thanks to Professor Carolyn Moores, Professor Katherine Thompson, and Dr. Mark Williams for ongoing support and encouragement. Much gratitude goes to Dr. Dave Houldershaw for the friendly computing support--you always went above and beyond. A special shout-out to my Birkbeck/LIDo PhD bestie and fellow 'mature student', Dr. Trupti Gore. Thank you to everyone who was involved in the collaborative research without which I would have no project, including Dr. Amanda Gibson, Yen Yi Tan and, especially, Dr. Ian Passmore for his incredible patience and for being an absolute pleasure to work with. I am grateful to the Bloomsbury Colleges and LIDo fellowship programs for the generous financial support and for the opportunities to attend conferences and present my research.

I want to thank my mom for always believing I'd become a doctor, in one form or another. I sure made you wait a long time! Lastly, and most profoundly, I want to thank my husband and partner, Rick, and my children, Chloe and Ollie. I wouldn't even be able to get out of bed without you. Thank you for making my life so much richer.

TABLE OF CONTENTS

<i>Using global genomic and transcriptomic approaches to explore host specificity and the role of non-coding RNA in the Mycobacterium tuberculosis complex.....</i>	<i>1</i>
DECLARATION.....	2
ABSTRACT	3
ACKNOWLEDGEMENTS	4
TABLE OF CONTENTS.....	5
LIST OF FIGURES	10
LIST OF TABLES.....	14
LIST OF ABBREVIATIONS.....	15
Chapter 1: Introduction.....	17
1.1 The <i>Mycobacterium tuberculosis</i> Complex and tubercular disease.....	17
1.2 Host adaptation in the MTBC.....	19
1.3 Studying host-adapted gene-regulation	21
1.4 Evaluating host-specific gene requirements.....	22
1.5 Focus on a host-specific gene sensor	24
1.6 Study Aims	24
Chapter 2: Using a Whole Genome Co-expression Network to Inform the Functional Characterisation of Predicted Genomic Elements from Mycobacterium tuberculosis Transcriptomic Data.....	25
2.1 ABSTRACT	25
2.2 AIMS	26
2.3 INTRODUCTION.....	26
2.3.1 Non-coding RNA in the MTBC.....	27
2.3.2 How many functional non-coding RNAs are there in the MTBC?.....	30
2.3.3 Computational prediction of non-coding RNA from genomic and transcriptomic data ...	34
2.3.4 Using WGCNA to implicate functional associations of non-coding RNA	40
2.4 MATERIALS AND METHODS.....	42
2.4.1 Data Acquisition and Mapping.....	43

2.4.2 Non-coding RNA prediction and quantification.....	43
2.4.3 Creation of the WGCNA network.....	45
2.4.4 Module Enrichment.....	46
2.4.5 Data exploration	47
2.5 RESULTS AND DISCUSSION	49
2.5.1 <i>M. tuberculosis</i> expresses an extensive range of ncRNA transcripts over a wide variety of experimental conditions.....	49
2.5.2 Module networks represent groups of co-expressed genes and predicted non-coding RNA	51
2.5.3 Focus on selected module networks.....	59
2.5.4 Comparison with other global <i>M. tuberculosis</i> networks.....	69
2.6 CONCLUSION.....	73
<i>Chapter 3: Using transposon-insertion sequencing to identify the different essential gene requirements in vitro between human-adapted and animal-adapted members of the MTBC.....</i>	75
3.1 ABSTRACT	75
3.2 AIMS	76
3.3 INTRODUCTION.....	76
3.3.1 Host-adapted species may have different gene requirements	76
3.3.2 Determining gene essentiality with transposon insertion sequencing	77
3.3.3 Statistical analysis of transposon insertion sequencing results	79
3.4 MATERIALS AND METHODS	82
3.4.1 Creation of Libraries	82
3.4.2 Processing sequencing reads	82
3.4.3 Data Analysis	83
3.4.4 Compiling set of orthologous genes	84
3.4.5 Analysis of non-coding RNA.....	84
3.4.6 Functional enrichment.....	85
3.5 RESULTS.....	85
3.5.1 Libraries show good saturation of 'TA' sites	85
3.5.2 Essential genes overlap with published datasets.....	90
3.5.3 Comparing essentiality between <i>M. bovis</i> and <i>M. tuberculosis</i> orthologous genes with a qualitative approach	92
3.5.4 Determination of different non-coding RNA requirements.....	94
3.5.5 Quantitative comparison of gene requirements between orthologous genes in <i>M. tuberculosis</i> and <i>M. bovis</i> libraries	95
3.6 DISCUSSION	98

3.6.1 Insertions in identical metabolism and respiration genes show different fitness effects in <i>M. bovis</i> versus <i>M. tuberculosis</i>	98
3.6.2 Differences in requirements for cell wall-associated genes may reflect species-specific cell-wall lipid repertoires.....	100
3.6.3 Other membrane-associated genes important for in vitro survival in <i>M. bovis</i>	101
3.6.4 Comparing orthologous genes with different numbers of 'TA' sites	103
3.6.5 Comparing results from HMM and Resampling Methods.....	103
3.6.6 Limitations of the study	105
3.7 CONCLUSIONS	106
 Chapter 4: Using transposon insertion sequencing to identify the gene requirements for adjustment to redox stress in <i>Mycobacterium bovis</i>.....	
4.1 ABSTRACT	108
4.2 AIMS	108
4.3 INTRODUCTION.....	109
4.4 MATERIALS AND METHODS.....	113
4.4.1 Transposon Library Construction.....	113
4.4.2 Sequencing adapter and primer design.....	113
4.4.3 Sequencing Library Preparation.....	116
4.4.4 Sequencing and read processing	119
4.4.5 Data analysis.....	120
4.5 RESULTS.....	122
4.5.1 Sequencing of independent libraries and technical replicates.....	122
4.5.2 Calculation of the insertion density of menadione-treated and untreated samples.....	126
4.5.3 Determination of conditional essentiality with menadione treatment	127
4.6 DISCUSSION	130
4.6.1 Menadione treated libraries have increased requirements for genes involved in cell wall integrity	130
4.6.2 The requirement for oxidoreductases with menadione treatment varies	131
4.6.3 Oxidative stress response genes required for survival in menadione	133
4.6.4 Limitations of the study and further work.....	136
4.7 CONCLUSIONS	137
 Chapter 5: Exploring the regulation of <i>phoR</i> expression by antisense RNA	
5.1 ABSTRACT	138
5.2 AIMS	138
5.3 INTRODUCTION.....	139

5.3.1 Antisense RNA and bacterial gene expression	139
5.3.2 The PhoPR two-component system	140
5.3.3 An antisense RNA transcribed opposite <i>phoR</i>	143
5.3.4 CRISPR inhibition as a strategy to silence antisense RNA.....	144
5.4 MATERIALS AND METHODS	145
5.4.1 RT-qPCR.....	145
5.4.2 Design of sgRNAs to target antisense- <i>phoR</i>	145
5.4.3 Cloning of sgRNAs into pRH2521 plasmid	147
5.4.4 Transformation into <i>M. tuberculosis</i> and Induction of CRISPRi system	150
5.4.5 RNA extraction and sequencing.....	151
5.4.6 Quantification and Data Analysis	152
5.4.7 RNA structure and binding prediction.....	153
5.4.8 Analysis of publicly available RNA-seq datasets.....	153
5.5 RESULTS.....	154
5.5.1 An antisense transcript opposite <i>phoR</i> gene is predicted from <i>M. tuberculosis</i> RNA-seq data	154
5.5.2 Antisense- <i>phoR</i> is expressed in multiple RNA-seq datasets, including from <i>M. bovis</i>	156
5.5.3 Transcripts in region of antisense- <i>phoR</i> are detected at similar levels to housekeeping gene, <i>sigA</i> , in exponential growth conditions.....	157
5.5.4 Antisense- <i>phoR</i> is silenced using CRISPRi	158
5.5.5 Antisense silencing impacts <i>phoR</i> expression	161
5.5.6 Protein-coding genes associated with the cell membrane were differentially expressed with antisense-silencing.....	164
5.5.7 Two antisense transcripts are differentially expressed with as- <i>phoR</i> silencing.....	167
5.5.8 ATc treatment results in differentially expressed genes in both control and sgRNA expressing strains.....	169
5.6 DISCUSSION	170
5.6.1 As- <i>phoR</i> silencing reduces expression of <i>phoR</i>	170
5.6.2 The intergenic region between <i>phoP</i> and <i>phoR</i> may be involved in translational regulation of <i>phoR</i>	172
5.6.3 Decrease in <i>phoR</i> expression does not impact genes of the PhoP regulon in exponential growth	174
5.6.4 Further Work	176
5.7 CONCLUSIONS	177
Chapter 6: Conclusion	179
6.1 Exploring beyond the protein-coding genome.....	179
6.2 Back to the basics: discovering what is essential in the MTBC	180
6.3 Making 'antisense' of the PhoPR two-component system	182

6.4 Non-coding RNA in host-adapted gene systems.....	182
<i>APPENDIX.....</i>	<i>184</i>
Appendix 1. Published works.....	184
Appendix 2. Chapter 2.....	184
Appendix 3. Chapter 3.....	191
Appendix 4. Chapter 4.....	193
Appendix 5. Chapter 5.....	196
<i>REFERENCES.....</i>	<i>197</i>

LIST OF FIGURES

Figure 1.1. Phylogenomic tree based on maximum likelihood topology of human-adapted and animal-adapted members of MTBC.....	22
Figure 2.1. Analysis workflow.....	42
Figure 2.2. Hierarchical dendrogram of rlog transformed and limma batch corrected expression data by sample.	45
Figure 2.3. Heat map of correlation of module eigengene (ME) of each module with selected experimental conditions.....	53
Figure 2.4. Relative proportion of annotated CDS, predicted UTRs and predicted sRNAs in each module	55
Figure 2.5. Plot of number of UTRs against number of CDS in each module.	56
Figure 2.6. Expression of antisense transcripts.....	61
Figure 2.7. Antisense sRNA, ncRv2489/putative_srna:p2801108_2801678, (magenta bar) overlaps two transcripts and may encode a short peptide.	62
Figure 2.8. Antisense sRNAs (magenta bars) overlap Rv3230c and Rv3231c.....	63
Figure 2.9. Expression of antisense transcript.....	65
Figure 2.10. Overlapping 3' UTRs for Rv0292 (EccE3) and Rv0293c, (light green bars) may regulated transcription termination or transcript stability.....	67
Figure 2.11. 5' and 3' UTRs for Rv2081c (green bars) are overlapped by predicted sORFs.....	69
Figure 2.12. Protein coding genes involved in responses to hypoxia and adaptation to cholesterol cluster together in overlapping modules in different network approaches.....	71
Figure 3.1. Overview of transposon insertion sequencing (tn-seq) experiments..	79
Figure 3.2. 'TA' site saturation for tn-seq libraries.	86
Figure 3.3. The number of mapped reads is loosely correlated to the number of unique 'TA' sites with insertions.....	87
Figure 3.4. Insertion sites were well-distributed across the genomes.....	89

Figure 3.5. Essentiality calls for <i>M. bovis</i> and <i>M. tuberculosis</i> genomes with TRANSIT HMM. 'ES'=essential, 'GD'=growth defect, 'NE'=non-essential, 'GA'=growth advantage.....	91
Figure 3.6. Essential genes in <i>M. bovis</i> and <i>M. tuberculosis</i> show a large overlap with published datasets.....	92
Figure 3.7. There is significant overlap of essential and growth defect calls among orthologous protein-coding genes in the two genomes.....	94
Figure 3.8. Volcano plot showing genes with statistically significant ($p_{\text{adj}} < 0.05$) \log_2 fold-difference.....	98
Figure 3.9. The normalised insertion counts for <i>wag31</i>	105
Figure 4.1. Transposon insertion sequencing analysis (tn-seq) can be used with selective pressures.....	110
Figure 4.2. Chemical structures of menaquinone and menaquinone analogue, menadione.	111
Figure 4.3. Strategy for tn-seq primer design.....	115
Figure 4.4. Adapter-ligation screen.....	117
Figure 4.5. Flowchart of read processing pipeline.	120
Figure 4.6. Representative paired-end sequencing reads after adapter trimming.	120
Figure 4.7. Distribution histograms and quantile-quantile plots from representative samples	122
Figure 4.8. There is a higher degree of correlation between technical replicates.....	124
Figure 4.9. Read correlation plots.	125
Figure 4.10. Barplots showing insertion density	126
Figure 4.11. Volcano plot of resampling results comparing menadione treated and untreated tn-seq datasets.	128
Figure 4.12. Redox cycling of quinones.....	131
Figure 4.13. Relative position and quantity of normalised transposon insertions in the known domains of <i>mtr</i> (Mb2880).....	133

Figure 4.14. Relative position and quantity of normalised transposon insertions in the known domains of Mb1383, <i>irtA</i>	135
Figure 5.1. Models of regulation by antisense RNAs in bacteria.....	140
Figure 5.2. Alphafold prediction of monomer of PhoR _{Mtb} with predicted domains	142
Figure 5.3. Several transcriptional regulators involved in acid and stress responses are regulated by PhoP.....	143
Figure 5.4. Schematic diagram of CRISPR-interference system.....	144
Figure 5.5. CRISPRi two-plasmid system.....	149
Figure 5.6. Design of CRISPR-inhibition experiment.....	151
Figure 5.7. Diagram showing the orientation and length of as-phoR transcript relative to <i>phoPR</i> polycistronic operon.....	155
Figure 5.8. Sequence of <i>phoR</i> (blue) and as-phoR (green).....	156
Figure 5.9. Boxplots showing relative expression levels of as-phoR across multiple conditions using normalised counts from publicly available RNA-seq data.....	157
Figure 5.10. RT-qPCR with primers specific to as-phoR and <i>sigA</i>	158
Figure 5.11. Colony PCR.....	159
Figure 5.12. PCA plots before and after batch correction with Limma	160
Figure 5.13. Volcano plot showing differentially expressed transcripts with knock-down of as-phoR expression.....	161
Figure 5.14. Read coverage from forward strand mapped to <i>phoR</i> annotations...	162
Figure 5.15. The log ₁₀ ratio of mean base pair read coverage.....	163
Figure 5.16. The orientation of the sgRNA: dcas9 complex relative to direction of transcription.	163
Figure 5.17. Plot of normalised counts versus strain.....	166
Figure 5.18. Secondary structure of as-phoR as predicted by RNAFold.....	166
Figure 5.19. Plots of normalised counts versus strain.....	168
Figure 5.20. Schematic of the relative position and lengths of differentially expressed antisense transcripts.....	169

Figure 5.21. Number of differentially expressed genes with ATc treatment.....	170
Figure 5.22. There is a sORF predicted between <i>phoP</i> and <i>phoR</i> that is actively translated.	173
Figure 5.23. Close up of intergenic region between <i>phoP</i> and <i>phoR</i>	173
Figure 5.24. Translation of a sORF found in 5' leader of <i>phoP</i> may be regulated by antisense transcript.	177

LIST OF TABLES

Table 2.1. Origins and targets of non-coding RNA types in bacteria.....	29
Table 2.2. Functionally characterised sRNA and asRNA in mycobacteria	31
Table 2.3. Conserved non-coding RNA families and sequence listings from the RFAM database	33
Table 2.4. Datasets used in analysis.....	43
Table 2.5. Conditions tested in the various samples used in this analysis	48
Table 2.6. Tally of predicted expressed elements in the baerhunter-generated combined annotation file.....	50
Table 2.7. UTRs and module assignment of adjacent ORFs.....	58
Table 3.1. Sequencing statistics from <i>M. bovis</i> and <i>M. tuberculosis</i> tn-seq libraries.	87
Table 3.2. Non-coding RNAs with different essentiality calls in <i>M. bovis</i> versus <i>M.</i> <i>tuberculosis</i>	95
Table 3.3. Orthologous genes showing statistically significant log ₂ fold-difference between <i>M. tuberculosis</i> and <i>M. bovis</i>	97
Table 4.1. Oligonucleotides used in this chapter.....	114
Table 4.2. Sequencing and processing statistics.....	123
Table 4.3. Resampling results for genes with statistically significant log ₂ fold- changes.....	129
Table 5.1. List of oligonucleotides used in this study.....	146
Table 5.2. List of bacterial strains and plasmids used in this study.....	148
Table 5.3. Differentially-expressed genes	165

LIST OF ABBREVIATIONS

as-phoR	antisense phoR
asRNA	antisense RNA
ATc	anhydrotetracycline
bp	base-pair(s)
CDS	coding sequence
CFU	colony-forming unit
CRISPR	Clustered regularly interspaced short palindromic repeats
CRISPRi	CRISPR interference/inhibition
<i>E. coli</i>	<i>Escherichia coli</i>
h	hours
IFN- γ	interferon-gamma
m	minutes
<i>M. bovis</i>	<i>Mycobacterium bovis</i>
<i>M. smegmatis</i>	<i>Mycolicibacterium smegmatis</i>
<i>M. tb</i>	<i>Mycobacterium tuberculosis</i>
<i>M. tuberculosis</i>	<i>Mycobacterium tuberculosis</i>
ME	module eigengene
MM	module membership
MTBC	<i>Mycobacterium tuberculosis</i> complex
ncRNA	non-coding RNA
nt	nucleotide(s)
ORF	open reading frame
PAM	proto-spacer adjacent motif
RBS	ribosome binding site
RNA-seq	RNA sequencing
RNAP	RNA polymerase
ROS	reactive oxygen species
RT-qPCR	Real-time quantitative polymerase chain reaction
s	seconds
SNP	single nucleotide polymorphism
sORF	short open reading frame
<i>SPy</i>	<i>Streptococcus pyogenes</i>
sRNA	short RNA
TCS	two-component system
TIS	translation initiation site
Tn-seq	transposon insertion sequencing

TSS	transcription start site
TTS	transcription termination site
uORF	upstream short ORF
UTR	untranslated region
WGCNA	weighted gene co-expression analysis

Chapter 1: Introduction

Some of the text in this chapter is adapted from work previously published in:

Stiens, J., Arnvig, K. B., Kendall, S. L., & Nobeli, I. (2022). Challenges in defining the functional, non-coding, expressed genome of members of the *Mycobacterium tuberculosis* complex. *Molecular Microbiology*, 117(1), 20–31.
<https://doi.org/10.1111/mmi.14862>

1.1 The *Mycobacterium tuberculosis* Complex and tubercular disease

Tuberculosis continues to be a leading cause of death worldwide, killing over 1.3 million and infecting over 10.6 million people in 2022 (Geneva: World Health Organisation, 2023). Tubercular disease affects the health of humans, livestock, and wild animals as well as incurring a large economic burden. It is caused by members of the *Mycobacterium tuberculosis* Complex (MTBC), a group of closely-related pathological bacteria descended from a common ancestor species (itself evolved from a soil-dwelling bacterium) through a series of gene deletions and acquisitions (Brites et al., 2018; Gagneux, 2018; Loiseau et al., 2020; Sapriel et al., 2019). The pathogens primarily infect the lungs of the host where they are engulfed by macrophages. Within the macrophage, the pathogen is enclosed in a phagosome where the invaders are presented with a series of challenges including restriction of nutrients, available oxygen and cofactors such as iron, and by lowering phagosomal pH. Members of the MTBC have evolved multiple defences against host challenges including a waxy outer cell envelope to maintain cell homeostasis in hostile environments and mechanisms to scavenge nutrients from the host.

M. tuberculosis is primarily restricted to infection and transmission among humans with very seldomly reported cases of animal infection and transmission¹ (Gagneux, 2018). In contrast, the animal-adapted members of the MTBC, including *Mycobacterium bovis* (*M. bovis*), *Mycobacterium orygis* and *Mycobacterium caprae*,

¹ Reverse-zoonosis has been reported in areas with high TB burden and needs careful monitoring to avoid creating animal reservoirs for multi-drug resistant strains of *M. tuberculosis* (Kock et al., 2021)

among others, are able to infect and transmit between a large variety of mammalian hosts, including humans. In fact, human tuberculosis can be caused by either *M. tuberculosis* or *M. bovis* infection and the clinical and pathological manifestations of *M. tuberculosis* and *M. bovis* infections in humans are considered indistinguishable² (Grange, 2001; Sawyer et al., 2023). Zoonotic tuberculosis is an under-appreciated threat, especially in countries with a high tuberculosis burden (Kock et al., 2021; Olea-Popelka et al., 2017; Sawyer et al., 2023).

M. tuberculosis infection is initiated by the respiration of aerosolised bacteria. The primary acute stage of infection includes the activation of the cell-mediated immune system and dissemination to the lymph nodes and other organs. This involves T-lymphocytes which release cytokines such as interferon-gamma (IFN- γ), which is used in assays to determine infection. At this point, it appears that most infections are cleared (Smith, 2003). However, if this is not successful, the disease progresses, and caseous granulomas form in the lungs and lymph nodes which progress into cavitary lung lesions opening into the airways and releasing the bacteria, leading to transmission (Smith, 2003). In a limited number of *M. tuberculosis* infections, the bacteria are successfully contained but lie dormant in a 'latent' infection which may be re-activated in response to an unknown trigger, perhaps when the host immune system is more compromised and conditions more favourable for the pathogen (Getahun et al., 2015). In cattle, *M. tuberculosis* infection is attenuated. A study of parallel infections of cattle with *M. bovis* or *M. tuberculosis* found that the granulomatous lesions in lungs and lymph nodes typical of *M. bovis* infection were not found with *M. tuberculosis* infection, despite the presence of the bacteria in the lymph nodes and positive IFN- γ responses (Waters et al., 2010). In cattle, *M. tuberculosis* appears to be more rapidly cleared, with less damage caused by inflammation and a lack of cavitary lesions (Basaraba & Hunter, 2017; Waters et al., 2010).

The initial response of cattle or humans to *M. bovis* (the causative agent of bovine tuberculosis) is similar to *M. tuberculosis*, first initiating the cell-mediated immune system followed by humoral response (Holder et al., 2024). Infection by *M. bovis* is

² It is thought that extra-pulmonary tuberculosis in humans is more common with *M. bovis* infection (Grange, 2001).

primarily by the respiratory route, but also through contaminated food, water and milk (Grange, 2001; Sawyer et al., 2023). Though there is some evidence that *M. bovis* is less virulent in humans (Gonzalo-Asensio et al., 2014), it is hyper-virulent in certain hosts such as rabbits, guinea pigs and mice (Rehren et al., 2007). It is unclear whether *M. bovis* utilises a dormancy strategy similar to *M. tuberculosis* (Sabio y García et al., 2020).

Though the immune systems of mammals are very similar, there is evidence of co-evolution between mammalian hosts and host-specific pathogens. For example, different mammals have evolved different repertoires of antigen receptors (TLR's, toll-like receptors) on their immune cells (Bailey et al., 2013) and a recent study showed that human and bovine macrophages have diverse metabolic responses to *M. bovis* antigens (Bartens et al., 2024). Furthermore, the levels of specific T-cell types differ between humans and other mammals, with rodents, cattle and other ruminants utilising a much higher level of TCR- $\gamma\delta$ T-cells (Bailey et al., 2013). Perhaps the different qualities of the immune systems of diverse mammalian hosts require the bacteria to either employ effective and specifically-adapted strategies to survive and spread infection within a single host group, or more generally-applicable systems to survive in multiple different hosts.

1.2 Host adaptation in the MTBC

The different lineages in the MTBC have evolved to maximise survival in preferred hosts with seven lineages known to be adapted to human hosts and other members adapted to animal hosts. The animal-adapted lineages of the MTBC have reduced genomes (four deletion regions, 'RD') compared to the human-adapted lineages and have evolved into a more generalist species to exploit new ecological opportunities. This could either have occurred independently from the more host-limited *M. tuberculosis*, through a common generalist ancestor, or directly from a human-adapted ancestor³ (Brites et al., 2018; Gagneux, 2018) (Figure 1.1). The metabolic needs of the members are known to differ--with *M. bovis* unable to

³ This may be true in respect to the 'classical' animal-adapted species including: *M. bovis*, *M. caprae*, *M. orygis*, *M. pinnipedii* and *M. microtii*; the case is less clear for *M. mungi* and *M. surricattae* which may have evolved to/from the human-adapted lineage, *M. africanus* through host jumps (Gagneux, 2018)

utilise carbohydrates such as glycerol as a carbon source⁴ as *M. tuberculosis* does (Keating et al., 2005). Proteomic comparisons of the two strains at log-phase in the same culture media have identified 450 identical proteins, in addition to 248 'variable' proteins with different amino acid sequences, that are differentially expressed (K. M. Malone et al., 2018), indicating the basic gene requirements for growth differ between the strains.

M. bovis shares greater than 99.5% nucleotide similarity with *M. tuberculosis*, and the reference genome, AF2122/97 (Garnier et al., 2003; Malone et al., 2017), has approximately 2000 SNPs compared to the reference genome of *M. tuberculosis*, H37Rv (Cole et al., 1998; Lew et al., 2011). Mutations and SNPs in orthologous protein coding genes among species members of the MTBC and between species-specific strains, explain some of the reported differences in gene expression (Golby et al., 2007; Malone et al., 2018; Rehren et al., 2007). For example, a *M. bovis* specific mutation in the gene for the anti-sigma factor, *rskA*, means the sigma factor, SigK, is constitutively active, leading to higher expression of much of the SigK regulon compared to *M. tuberculosis* (Golby et al., 2007). However, mutations that alter promoter sequences and create or destroy transcriptional start sites (TSS) can also alter the expression of genes and antisense transcripts (Chiner-Oms et al., 2019; Dinan et al., 2014; Golby et al., 2013). Flexible and transient regulation of gene expression is crucial for pathogens to respond quickly to the onslaught of host defences, with post-transcriptional regulation of the effect and stability of transcripts used as a parsimonious strategy that limits the waste of cell resources (Chakravarty & Massé, 2019). It is thus not unreasonable to hypothesise more generally, that variations in the genomic sequence of non-coding elements could contribute to differences in gene and protein expression through both transcriptional and post-transcriptional levels of regulation (Schwenk & Arnvig, 2018).

⁴ This is definitely true of the lab-adapted *M. bovis* strain, AF2122/97, but it is unclear whether this is a widely-shared characteristic of the field strains of *M. bovis*.

1.3 Studying host-adapted gene-regulation

In this thesis, I will consider the topic of host-adapted gene regulation in the MTBC using several different approaches. Firstly, in the introduction to Chapter 2, I describe how non-coding transcription in the MTBC remains underexplored with only a handful of non-coding regulators characterised in *M. tuberculosis*, and virtually none in the animal-adapted species, despite pervasive transcription in non-coding regions (Arnvig et al., 2011; Dinan et al., 2014; Golby et al., 2013; Ju et al., 2024). Gaining understanding of both the mechanisms of RNA-mediated regulation in the MTBC and identifying functionally relevant transcripts is a starting point for comparing these regulators among the MTBC members. To contribute to the landscape of gene pathways and functions that may involve ncRNA, the investigation begins by predicting unannotated ncRNA from transcriptomic data and creating a co-expression network that clusters ncRNA with protein-coding genes, some involved in known gene pathways or functions. Prediction of non-coding RNA depends on the availability of high-quality transcriptomic data. With the increasing availability and decreasing cost of RNA-seq, there are many datasets available for *M. tuberculosis* in a wide range of experimental conditions. Unfortunately, there are very few transcriptomic studies for the other members of the MTBC, with available wild-type *M. bovis* data currently limited to exponential growth and stationary conditions. Therefore, Chapter 2 is limited to exploring the non-coding genome of *M. tuberculosis* with the hope that these results can eventually be compared to a similar analysis applied to other members of the complex.

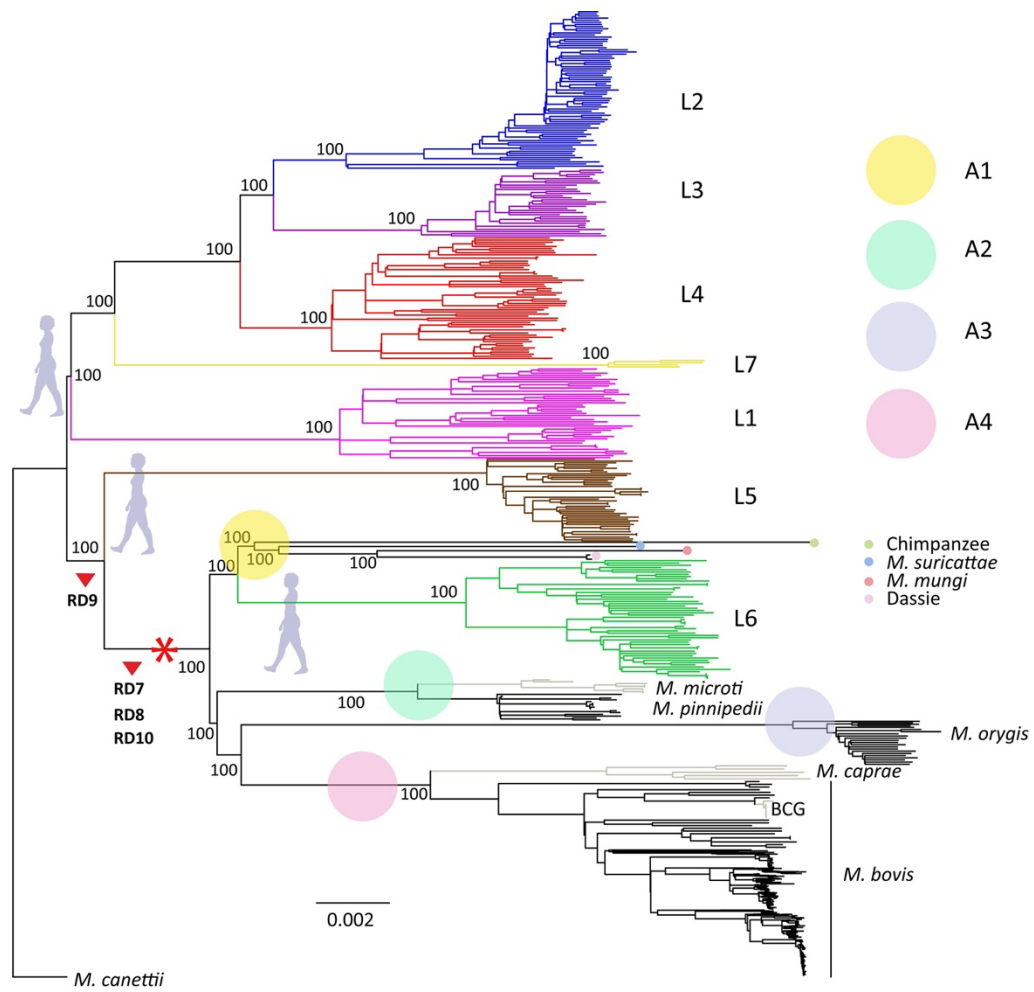


Figure 1.1. Phylogenomic tree based on maximum likelihood topology of human-adapted and animal-adapted members of MTBC shows that animal-adapted lineages (clades A1-A4) evolved through significant deletions in genomes relative to human-adapted lineages (L1-L6). Branch lengths are proportional to SNPs versus *M. canettii* ancestor genome. Red arrows mark large deletions (RD7-10) in animal-adapted lineages. Red asterisk marks host range expansion within the MTBC. From (Brites et al., 2018), Figure 1.

1.4 Evaluating host-specific gene requirements

As the lineages must evolve to adapt to different host immune challenges, the genes required for these adaptations may be different among the members of the MTBC. Despite the nearly identical genomes, *M. tuberculosis* and *M. bovis*, differ in metabolic requirements (Keating et al., 2005), ability to scavenge iron (Tullius et al., 2019) and their response to oxidative stress (Golby et al., 2007; Sohaskey & Modesti, 2009). Some of these differences are directly related to deleted genes and SNPs but it is not straightforward to determine which mutation is responsible for a particular phenotype. Transposon insertion sequencing (tn-seq) is a global method that directly implicates mutations in specific genes to survival in a particular environmental context. It has been used to implicate genes that are required for *M.*

tuberculosis survival in rich versus minimal media (Minato et al., 2019) and with specific carbon sources (Griffin et al., 2011). Tn-seq has also been used to highlight differences in susceptibility to antibiotics among clinical strains of *M. tuberculosis* which were not predicted by whole genome sequencing (Carey et al., 2018).

In Chapter 3, I compare parallel tn-seq libraries created in *M. bovis* and *M. tuberculosis* to indicate genes that are differentially required in animal versus human-adapted members of the MTBC with highly similar genomes. Using identical media and culture conditions, this study highlights the genes that are more or less tolerant of mutations in one species than the other. Some of the orthologous genes that are differently required contain SNPs or mutations, but some of them are identical in amino acid sequence, indicating different regulation in the host-adapted species. I also consider whether the genomic regions defined by the predicted intergenic ncRNAs from *M. tuberculosis* are essential in either species.

In Chapter 4, I use tn-seq to evaluate *M. bovis* genes required for survival in oxidative stress. Phagosomes in mammalian hosts engulf mycobacteria and attempt to restrict growth by creating oxidative stress with reactive nitrogen and oxygen species and by reducing pH of the intracellular environment (Rohde et al., 2007). The host-adapted species have evolved regulation pathways to sense and combat this challenge in different ways (Cumming et al., 2014; Magee et al., 2014; Pacl et al., 2018; Queval et al., 2021; Widdison et al., 2008). It has been proposed that *M. bovis* has different sensitivity to oxidative stress than *M. tuberculosis* (Golby et al., 2007; Ma et al., 2022; Sohaskey & Modesti, 2009) and SNPs in regulators of redox homeostasis in *M. bovis*, such as *whiB3* and *phoR* may be responsible for differences in lipid metabolism, secretion and cell envelope composition compared to *M. tuberculosis* (García et al., 2021; Goar et al., 2022; Gonzalo-Asensio et al., 2014; Malone et al., 2018; Singh et al., 2009; Urtasun-Elizari et al., 2024). In Chapter 4, I describe the creation of sequencing libraries from transposon libraries selected on media with and without menadione, a drug that causes oxidative stress in mycobacteria. I then identify the genes in *M. bovis* that show an increased requirement for survival under this selective pressure. This will add to the understanding of essential genes in the MTBC host-pathogen interface.

1.5 Focus on a host-specific gene sensor

The PhoPR two-component system is an essential part of MTBC detection and response to host-generated redox stress. Despite a SNP in the *phoR* ortholog, an intact *phoPR* operon is necessary for normal growth and virulence of both *M. tuberculosis* and *M. bovis* (Gibson et al., 2022; Urtasun-Elizari et al., 2024).

Exploring the co-expression network created in Chapter 2, I observed an antisense transcript, co-expressed with other genes involved in host-pathogen adaptation, that overlaps the *phoR* gene. Antisense transcription is pervasive in the MTBC and yet its purpose is obscure. To explore the possibility that this antisense transcript is involved in regulating *phoPR* expression, in Chapter 5, I present the results of an experiment using CRISPR-inhibition (CRISPRi) to silence the transcript.

Preliminary results indicate that the antisense transcript may be involved in post-transcriptional regulation of *phoR* mRNA. This would be a novel example of antisense regulation of mRNA stability in the MTBC and could help to illuminate this relatively unstudied phenomenon.

1.6 Study Aims

This thesis aims to explore host-specific adaptation of members of the MTBC by utilising genome-wide transcriptomic and phenotypic assays in *M. tuberculosis* and *M. bovis* including:

- inferring the functional associations between non-coding RNA and annotated protein-coding genes in *M. tuberculosis* using a gene co-expression network
- applying global phenotyping (tn-seq) to identify and compare essential gene requirements of *M. tuberculosis* and *M. bovis* for *in vitro* growth, and for *M. bovis* under oxidative stress
- describing antisense RNA regulation in the MTBC by targeting a well-studied gene system of the host-pathogen interface using CRISPRi and RNA-seq

Chapter 2: Using a Whole Genome Co-expression Network to Inform the Functional Characterisation of Predicted Genomic Elements from *Mycobacterium tuberculosis* Transcriptomic Data

This chapter includes both edited and verbatim text from the following previously published manuscripts of which I was the original and primary author. The introduction is based on:

Stiens, J., Arnvig, K. B., Kendall, S. L., & Nobeli, I. (2022). Challenges in defining the functional, non-coding, expressed genome of members of the *Mycobacterium tuberculosis* complex. *Molecular Microbiology*, 117(1), 20–31. <https://doi.org/10.1111/mmi.14862>

The results and discussion come from:

Stiens, J., Tan, Y. Y., Joyce, R., Arnvig, K. B., Kendall, S. L., & Nobeli, I. (2023). Using a whole genome co-expression network to inform the functional characterisation of predicted genomic elements from *Mycobacterium tuberculosis* transcriptomic data. *Molecular Microbiology*, 119(4), 381-400. <https://doi.org/https://doi.org/10.1111/mmi.15055>

2.1 ABSTRACT

A whole genome co-expression network was created using *Mycobacterium tuberculosis* transcriptomic data from publicly available RNA-sequencing experiments covering a wide variety of experimental conditions. The network includes expressed regions with no formal annotation, including putative short RNAs and untranslated regions of expressed transcripts, along with the protein-coding genes. These unannotated expressed transcripts were among the best-connected members of the module sub-networks, making up more than half of the ‘hub’ elements in modules that include protein-coding genes known to be part of regulatory systems involved in stress response and host adaptation. This dataset provides a valuable resource for investigating the role of non-coding RNA, and conserved hypothetical proteins, in transcriptomic remodelling. Based on their connections to genes with known functional groupings and correlations with

replicated host conditions, predicted expressed transcripts can be screened as suitable candidates for further experimental validation.

2.2 AIMS

- Predict non-coding RNA from *M. tuberculosis* RNA-seq data over wide variety of physiologically-relevant culture conditions
- Create WGCNA network to cluster non-coding RNA along with protein-coding transcripts into modules by co-expression over range of culture conditions
- Evaluate the network to observe trends in non-coding RNA expression
- Use the principle of 'guilt by association' to highlight non-coding RNA or uncharacterised protein candidates that may associate with proteins of known function for further study
- Create a resource for *M. tuberculosis* researchers to explore the network

2.3 INTRODUCTION

The members of the *Mycobacterium tuberculosis* complex (*MTBC*) have complex lifestyles that require rapid adaptation to host defences and immune pressure, including nutritional immunity, hypoxia and lipid-rich environments. To adapt to these environmental challenges, bacterial cells must make complex transcriptomic adjustments, and these are thought to be complemented and fine-tuned by post-transcriptional regulation and use of non-coding RNA (ncRNA). NcRNA can alter the abundance of RNA and proteins by controlling mRNA stability, processing and access to ribosome binding sites. Discovering the contribution of the non-coding genome to specific adaptation-response pathways may improve our ability to prevent the evolution of persistent phenotypes, design therapeutics to address zoonotic infection, and prevent reverse-zoonosis of drug-resistant human-adapted *M. tuberculosis* in animals. In this introduction, I will first summarise what is known about ncRNA in the MTBC, discuss the challenges of computational prediction of ncRNA, and outline a strategy to use a co-expression network to discover functional associations of predicted ncRNA using publicly-available *M. tuberculosis* RNA-seq data.

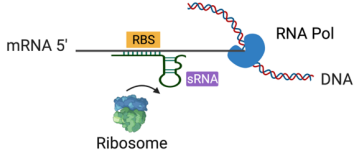

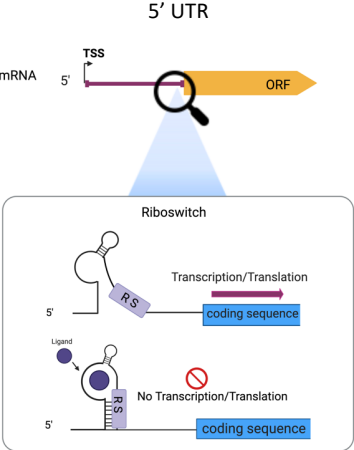

2.3.1 Non-coding RNA in the MTBC

The mycobacterial genome produces a range of conditionally expressed transcripts, including non-coding RNA, short, unannotated ORFs and untranslated regions at the 5' and 3' end of protein-coding sequences, many of which are poorly annotated and understood. In this chapter, I focus on 'non-coding' RNA (ncRNA), here referring to non-ribosomal RNA transcripts not known to be translated into peptides, such as short RNAs (sRNAs) acting on either distant or antisense mRNA targets and the expressed untranslated regions (UTRs) flanking coding regions (which may also contain short open reading frames (sORFs) upstream from coding regions) (Table 2.1). The proportion of non-ribosomal, ncRNA in the *M. tuberculosis* transcriptome has been shown to increase in stationary and hypoxic conditions, indicating a potential role in adjusting to environmental cues (Aguilar-Ayala et al., 2017; Arnvig et al., 2011; Gerrick et al., 2018; Ignatov et al., 2015). Several mycobacterial ncRNA transcripts (particularly, sRNA) have been extensively studied and found to be associated with regulatory systems controlling adaptation to stress conditions or growth phase, linked to virulence pathways and access to lipid media (Arnvig et al., 2011; Gerrick et al., 2018; Girardin & McDonough, 2020; Mai et al., 2019; Moores et al., 2017; Solans et al., 2014).

A definitive atlas of expressed non-coding elements in pathogenic mycobacteria does not exist. The lists available from databases and publications overlap only partially and are only available for the reference genomes of key representatives of the *Mycobacterium tuberculosis* complex (MTBC), such as *Mycobacterium tuberculosis* (*M. tuberculosis*) H37Rv. This gap in our knowledge impacts the successful analysis of the copious amounts of genomic and transcriptomic data that have become available in the last decade. For example, in the absence of a formal annotation of the non-coding transcriptome, the easiest and most common approach to differential expression analysis is to largely, or entirely, ignore information that does not relate to regions currently annotated as coding (CDS); this issue is more acute in studies focusing on non-reference *M. tuberculosis* strains or the animal-adapted strains, such as *Mycobacterium bovis*, where non-coding annotation is scarce or non-existent.

Bacterial non-coding RNA is a constantly evolving topic, but most of the research is focussed on the model organisms. Mycobacteria are different, in genome, physiology and lifestyle; and it appears that non-coding regulation in MTBC does not use the same accessory proteins or have the same sequence signatures as the model systems. Indeed, efforts to find an Hfq or ProQ analogue acting as an RNA chaperone in mycobacteria have so far been unsuccessful (Gerrick, 2018). These differences impact not only on our ability to transfer knowledge from model organisms to the MTBC species, but also on how applicable current experimental and computational methods are to discovering new regulators in mycobacteria.

Table 2.1. Origins and targets of non-coding RNA types in bacteria.

ncRNA Type	Description	Origin	Target/ Mechanism
<p>sRNA</p> 	Short structured RNA transcripts, 30-300 nt with short binding ('seed') region	Intergenic regions, UTRs, or antisense strands of coding genes; transcribed from own promoter, or by cleavage of longer transcripts	Involved in binding interactions with distant gene targets ('trans-acting') to regulate translation, including: mRNAs of other genes (e.g. UTRs of transcription factors), other sRNAs (known as 'sponge sRNAs' or 'ceRNAs'), and RNA-binding proteins
<p>asRNA</p> 	RNA transcript, 75-10,000 nt long	Complementary strand of UTR or coding sequence of regulated gene; transcribed from own promoter	Cognate RNA strand ('cis'-regulatory). Regulates by binding to mRNA transcript with perfect complementarity, forming duplex RNA: altering sensitivity to RNases, action of terminators, or access to RBS (ribosome binding site), can also act in 'trans'-regulatory manner with complementary sequences transcribed elsewhere in the genome
<p>5' UTR</p> 	<p>5-100s nt long, including transcriptional start sites (TSS), ribosome binding site (RBS), alternative transcriptional terminators (TTS) and cleavage sites</p> <p>Riboswitches are structured 5' UTRs that change secondary structure in response to ligand binding, controlling either transcription or translation of downstream gene by changing access to a 'regulatory sequence' (R S) which could be an anti-terminator sequence or RBS</p>	Upstream sequence of coding sequence, between TSS and start codon (alternative TSS may exist in gene locus)	<p>Binding interactions with sRNAs, proteins, metabolites and second messengers ('cis'-regulation of downstream ORF)</p> <p>Antisense binding with other UTRs. Potential source of sRNAs that can act on distant RNA targets ('trans'-regulatory)</p>
<p>3' UTR</p> 	5-100s nt long, following the coding ORF of the upstream gene. Can include RNase cleavage sites, alternative transcriptional start sites (TSS) and sRNA binding sites	Downstream sequence between stop codon and TTS. Alternative TSS and TTS may exist in gene locus.	<p>Binding interactions with sRNAs, proteins, metabolites and second messengers to regulate upstream ORF ('cis'-regulatory)</p> <p>Antisense binding with other UTRs, source of sRNAs ('trans'-regulatory)</p>

2.3.2 How many functional non-coding RNAs are there in the MTBC?

Only a handful of sRNAs and asRNAs have been functionally characterised in the mycobacteria literature (Table 2.2). In most cases, top-down approaches, such as differential expression studies and ChIP-seq (chromatin immunoprecipitation with sequencing), have been employed to discover *M. tuberculosis* sRNAs, such as the RNAP-associated, Ms1 (Arnvig et al., 2011; Šiková et al., 2019) and the PhoP-regulated, Mcr7 (Solans et al., 2014). Verification of the transcript size and abundance by Northern blot analysis has also established the stability of many ncRNAs in *M. tuberculosis* but identifying targets and functional associations requires extensive research. It is curious, that even among the eight well-characterised examples in Table 2.2, there are two (MrsI and asRelE2) not listed in the current official annotation of the reference H37Rv genome, available from the corresponding NCBI annotation (GFF) file (GCF_000195955.2_ASM19595v2_genomic.gff), most likely because they were relatively recent discoveries. This annotation file currently includes 20 features labelled as non-coding RNAs, 15 of which are listed in Arnvig et al., 2011 (Arnvig et al., 2011) and 9 in DiChiara et al. 2010 (DiChiara et al., 2010), with 4 listed in both. It also includes 10 “sequence features” which are annotated as fragments of putative small regulatory RNAs (sourced from DiChiara et al., 2010; Pelly et al., 2012), and two “misc RNA” including a tmRNA and the ribonuclease, P RNA. Although twenty or even thirty non-coding elements is almost certainly an underestimate of the total number of ncRNAs in *M. tuberculosis*, it is of note that the corresponding *E. coli* reference genome annotation (GCF_000005845.2_ASM584v2_genomic.gff) contains currently 72 elements labelled ncRNAs, suggesting that either functional non-coding elements are not very common in bacteria, or that, even for a well-studied organism, our understanding of non-coding regulation is incomplete.

Table 2.2. Functionally characterised sRNA and asRNA in mycobacteria. *Annotation according to Lamichlane et al., 2013.

Name (H37Rv annotation*, other names)	Mycobacterial organism	Genomic coordinates (H37Rv)	Citation	Pathway / targets
DrrS (ncRv11733, MTS1338)	<i>M. tuberculosis</i>	1960667- 1960783 (+)	Moore et al., (2017)	DosR regulon / unknown
Mcr7 (ncRv002, MTB000067)	<i>M. tuberculosis</i>	2692172- 2692521 (+)	Solans et al., (2014)	PhoP regulon / tatC
MrsI (ncRv11846)	<i>M. tuberculosis</i> , <i>M. smegmatis</i>	2096758- 2096863 (+)	Gerrick et al., (2018)	Iron-sparing response / brfA
Ms1 (ncRv0036a, MTS2823)	<i>M. smegmatis</i> , <i>M. tuberculosis</i>	4100669- 4100968 (+)	Šiková et al., (2019)	Transcription regulation/ RNAP
6C sRNA (ncRv13660c, B11)	<i>M. tuberculosis</i> , <i>Msmeg</i> , (homologues in all GC-rich gram+ bacteria)	4099386- 4099478 (-)	Mai et al., (2019)	Growth, virulence (ESX-1) / panD, dnaB
Mcr11 (ncRv11264Ac)	<i>M. tuberculosis</i>	1413227 - 1413107/8 (-)	Girardin and McDonough, (2020)	Growth, metabolism / unknown
F6 (ncRv10243)	<i>M. tuberculosis</i> , <i>M. smegmatis</i>	293604 - 293705 (+)	Houghton et al., (2021)	SigF regulon / unknown
asRelE2 (ncRv2866c)	<i>M. tuberculosis</i>	3178333 - 3177822 (-)	Dawson et al., (2022)	Toxin-antitoxin / relE2

Whereas functional characterisation is ultimately needed to create a reliable list of non-coding RNAs, homology to known families of RNAs from other organisms remains the most popular approach for predicting non-coding RNAs in the absence of experimental evidence. The RNA families described in the RFAM database (Kalvari et al., 2021) derive from the application of covariance models (and where structure information is not available, Hidden Markov Models) representing meticulously curated multiple sequence and secondary structure alignments of homologous RNAs. RFAM thus represents some of the most reliable predictions for non-coding elements in genomes and its predictions for *M. tuberculosis* H37Rv are summarised in Table 2.3. As conservation of structure is at the heart of RFAM families, non-coding RNAs with few or no known relatives in other species, and those that do not fold into strongly conserved structures, are unlikely to be found in RFAM. Hence, this database too is likely to miss elements that are specific to a small

number of pathogenic mycobacteria or that are too short to fold into a stable structure. In general, homology-based approaches to discovering *novel* non-coding elements will be limited in pathogenic mycobacteria as there are few closely-related genomes outside the phyla. One notable exception, 6C sRNA, is well-conserved among Gram-positive bacteria with over-expression leading to altered growth phenotypes in *M. tuberculosis*, *M. smegmatis* and another GC-rich bacterium, *Corynebacterium glutamicum*. Perhaps as a result, it is one of the few sRNAs for which target molecules have been identified and experimentally validated (Mai et al., 2019).

Table 2.3. Conserved non-coding RNA families and sequence listings from the RFAM database (<https://rfam.xfam.org>). Ribozymes (Group II catalytic introns and Bacterial RNase P class A), tRNAs and rRNAs have not been included in this table.

RNA Type	Family Name	Rfam ID	Number Sequences in RFAM	Length (nt)
Riboswitch	Cobalamin (B ₁₂)	RF00174	2	173-218
Riboswitch	ykok leader/Mbox (Mg ⁺)	RF00380	2	169-174
Riboswitch	TPP/Thi-box (thiamine)	RF00059	2	110
Riboswitch	ydaO/yuA leader (Cyclic di-AMP)	RF00379	1	222
Riboswitch	Glycine	RF00504	2	90-97
Riboswitch	S-adenosyl methionine (SAM-IV)	RF00634	1	119
sRNA	Mcr7	RF02671	1	348
sRNA	npcTB_6715	RF02886	2	211
sRNA	Ms1	RF02566	1	301
sRNA	ncRv12659	RF02659	1	171
sRNA	ncrMT1302	RF02341	1	108
sRNA	b55	RF01783	1	60
sRNA/asRNA	ASdes	RF0781	2	67
sRNA	F6	RF01791	1	57
sRNA	Ms_AS-5	RF02465	1	44
sRNA/5'UTR	5_ureB_sRNA	RF02514	1	294
asRNA	ASpks	RF01782	5	68-77

Expanding our exploration to resources beyond the official NCBI annotation, further complicates the question of what is known about functional, non-coding RNAs in mycobacteria. Mycobrowser (Kapopoulou et al., 2011), arguably the most popular internet resource for the exploration of representative mycobacterial genomes, currently lists 92 non-coding RNAs, labelled as 'ncRNA' (including sRNAs and asRNAs under this moniker) for H37Rv: 40 overlap the official NCBI GFF annotation and originate from the four key publications listing experimentally-verified non-coding RNAs (Arnvig et al., 2011; Arnvig & Young, 2009; DiChiara et al., 2010; Pelly et al., 2012) and the remaining 52 overlap the list compiled by DeJesus et al. using

their in-house computational tool, *BS_finder*, applied to RNA-seq data derived with a small-RNA sequencing protocol (Dejesus et al., 2017). Despite including annotations from nine other species and strains, including *M. bovis*, *M. smegmatis* and *M. tuberculosis 18b*, non-coding RNAs annotated with the tag “ncRNA” appear in the GFF files of only three additional species/strains in Mycobrowser, and only *M. tuberculosis 18b* has more than two ncRNAs listed. Strikingly, *M. bovis*, sharing more than >99.95% sequence identity to *M. tuberculosis*, has no other entries for RNAs apart from rRNAs, tRNAs and the same two RNAs tagged “misc_RNA” in the *M. tuberculosis* annotation; it is highly unlikely that many of the ncRNAs present in *M. tuberculosis* do not have a counterpart in *M. bovis*; and thus, the list must be assumed to be incomplete. In fact, at least 41 of experimentally-verified sRNAs found in various mycobacterial species, including in the above studies, can be mapped to the *M. bovis* genome (Dinan et al., 2014) and a sequence comparison (A2.1 Supplemental Tables: Ch2_Supp_Table_1) finds that only three of the listed *M. tuberculosis* ncRNAs have less than 99.0 % sequence identity in *M. bovis* (and all have greater than 92% similarity). The lack of standardised tags and incomplete listings of non-coding elements (even within the same resource), together with the absence of a clear justification for which elements are included and why, likely adds to the confusion about non-coding regulation in mycobacteria. A more systematic approach to the annotation tags of these elements, similar to approaches suggested for consistent naming of non-coding RNA (Lamichhane et al., 2013), could go some way towards eliminating this confusion.

2.3.3 Computational prediction of non-coding RNA from genomic and transcriptomic data

The most extensive lists of putative non-coding RNAs in mycobacteria are the result of computational predictions based on genomic or transcriptomic data (or sometimes both). Computational prediction algorithms have been used with moderate success in other bacteria, including *Salmonella enterica* (Sridhar et al., 2010) and *Staphylococcus aureus* (Liu et al., 2018) and new tools continue to be developed with increasing sophistication. However, the utility of these tools is even further limited when applied to mycobacteria. Genomics-based methods rely on the conservation of non-coding elements across several species and, like Rfam, are likely to miss elements specific to a small subset of the genus or unique to a species.

Such comparative genomics methods are typically enhanced by the search for characteristic sequence features and other signals of regulatory RNAs such as promoters, terminator structures and transcription-factor binding sites. For example, SIPHT begins with conserved intergenic sequences (defined as the sequence between two annotated genes or open reading frames (ORFs), on one strand) and looks for characteristic features of sRNAs in these regions, such as conserved promoters and rho-independent terminator motifs (Livny et al., 2008). Other genomics-based programs rely entirely on sequence features and genomic context (ignoring conservation). sRNAScanner determines intergenic sequences using genome annotation files and differentiates coding from non-coding sequences using position scoring matrices for sequence signals such as RBS and start codons (Sridhar et al., 2010). A recently published tool, the Pred-GsRNA feature of the PresRAT server, extracts intergenic sequences, also based on genome annotation, and excludes candidates that have an 8 nt sequence found to be depleted in known sRNAs. It scores each predicted sequence with weighted Minimum Free Energy scores for predicted paired and loop regions and scores for the predicted U-rich consensus sequences typical of intrinsic terminators (Kumar et al., 2020). The server offers 405 possible ‘non-genic sRNA’ predictions for the *M. tuberculosis* H37Rv genome (<http://www.hpppi.iicb.res.in/presrat/>). When the predicted sRNA coordinates are compared with the coordinates of the 92 ‘stable’ RNAs in the H37Rv genome on Mycobrowser (<https://mycobrowser.epfl.ch>), there are no PresRAT predicted sRNAs overlapping the boundaries of the Mycobrowser listed RNAs, except for low-ranking predictions that were over 4000 bp long, indicating that this method has limited power to recognise intergenic sRNA elements in mycobacterial genomes.

Relying on the current annotation to define the intergenic search space is problematic given that ribosome occupancy studies suggest that there are a significant number of unannotated proteins encoded at the 5’ ends of annotated genes (Shell et al., 2015; Smith et al., 2022). Furthermore, a considerable proportion of transcripts in the mycobacterial genome are either ‘leaderless’, meaning the transcription start site and the start codon are overlapping, and the transcripts therefore lack the canonical Shine-Dalgarno sequence used to identify ORF boundaries (Cortes et al., 2013; Ju et al., 2024; Martini et al., 2019; Sawyer et al.,

2021; Shell et al., 2015). Programs that search the intergenic regions for conserved sequence features based on sRNAs discovered in the model organisms are also less effective in mycobacteria as mycobacteria make use of a large number of alternative sigma factors which recognise diverse promoter sequences, and many lack a conserved -35 sequence (Newton-Foot & Gey van Pittius, 2013). Mycobacterial transcripts, including sRNAs, also generally lack the recognisable intrinsic terminator motifs at their 3' ends typical of Hfq-binding sRNAs (Arnvig et al., 2014; D'Halluin et al., 2023; DiChiara et al., 2010; Moores et al., 2017). Furthermore, identifying regions of high GC-content in order to detect RNA secondary structure in the intergenic space is even more challenging in the context of the GC-rich genome of mycobacteria. In any compact bacterial genome, tools that narrow the search space to strictly intergenic regions that lack annotated genes on either strand, effectively ignore sRNAs and asRNAs generated from coding regions, antisense regions or 5'/3' UTRs; this may bias our understanding of non-coding regulation in mycobacteria.

Transcriptomics-based detection methods are essentially versions of sliding window approaches looking for abrupt increases and drops in the expression signal and using such changes to delineate the limits of putative non-coding elements. High-throughput RNA sequencing (RNA-seq) has exposed a multitude of short transcripts from intergenic sequences, 5' and 3' UTRs and antisense to coding regions. Identifying functional transcripts in the conditions examined is the main challenge when using these data in non-coding RNA discovery. For example, sensitive methods are able to pick up expressed elements in regions of low read coverage; this signal may represent true low-abundance transcripts but it can also be the result of either technical noise or stochastic gene expression. The more sensitive computational methods will therefore inevitably over-predict putative non-coding elements. Ironically, high-depth sequencing has magnified this problem (Mao et al., 2015; Tarazona et al., 2011). Non-fragmented, size-selected libraries, where small transcripts remain intact, are superior for discerning between signal and noise for small RNA transcripts (Leonard et al., 2019; Wang et al., 2016). For all the reasons discussed above, detecting the existence of sRNAs expressed in low levels against very strongly expressed coding genes remains a computational challenge. Sequencing strand-specific cDNA libraries, where the information about

which strand the transcript originates from is preserved, is necessary for the discovery of new ncRNAs. Preservation of the strand information avoids mis-mapping asRNAs or other overlapping sRNAs that might otherwise be mapped to a coding gene on the opposite strand.

Many labs have developed their own computational pipelines and scripts to map RNA-seq data, normalise signals and identify ncRNA transcripts across the genome (Ami et al., 2020; Dejesus et al., 2017; Gómez-Lozano et al., 2014; Miotto et al., 2012; Wang et al., 2016), whereas others have carried out this process semi-manually (Arnvig et al., 2011). Progress in the field, and an easy comparison between approaches, has been hindered by the fact that not all of the labs publishing computational predictions have made their code readily available. In response to this challenge, several groups have created publicly-available prediction programs or workflows such as *Rockhopper* (McClure et al., 2013), *DETR'PROK* (Toffano-Nioche et al., 2013), *ANNOgesic* (Yu et al., 2018), *APERIO* (Leonard et al., 2019) and *baerhunter* (Ozuna et al., 2019). All of these transcriptomics-based methods require users to set thresholds for separating background noise (whatever its origin) from signal in the data. Indeed, most programs need adjustment to their default parameters in order to respond to sequencing depth and signal abundance, but tuning these parameters can be a matter of art rather than science.

The more sophisticated among the transcriptomics-based approaches use a combination of sources, such as transcriptional start sites (TSSs) or conservation across species, to reduce false positives. *DETR'PROK* is a Galaxy-based workflow, coordinating over 40 publicly-available Galaxy sequence comparison tools into a pipeline which streamlines the number of user-defined parameters. However, there are still 14 different user inputs, most of which concern filtering to account for read depth and transcriptional noise (Toffano-Nioche et al., 2013). The *ANNOgesic* suite of tools utilises multiple third-party software packages, as well as its own scripts to analyse RNA-seq data and filter predictions. Although, the suite includes an *sRNA-finder* module, using this module in isolation on user-generated alignment files requires specific file formats for the alignment (wig) and several reference annotation files. Multiple levels of filtering are possible to identify *bona fide* ncRNAs, but such filtering requires downloading of tools and databases such as RNAfold

(Denman, 1993), BSRD (Li et al., 2013) and the NCBI nr protein database (NCBI Resource Coordinators, 2014). In the context of validating mycobacteria ncRNA predictions, such databases may possibly be less relevant, given the lack of homology or shared sequence features between mycobacterial and other bacterial ncRNAs. Additionally, fine-tuning cut-off parameters to distinguish signal from noise is ultimately still up to the user. Somewhat surprisingly, the added complexity of such methods does not always translate into more accurate results: in limited comparisons between methods that use additional information and the simpler, signal-only-based method of *baerhunter*, used in this chapter, it was found that a naïve approach performs comparatively well, most likely because more sophisticated methods often require more tuning of their parameters to take advantage of their added complexity (Ozuna et al., 2019). Rockhopper is an independent, Java-based tool designed for bacterial RNAseq data (McClure et al., 2013). To eliminate guesswork by the user to adjust for noise vs. signal, the program normalises for read counts using the upper quartile of non-zero gene expression values and generates a transcriptional map of the predicted non-coding elements. *Baerhunter* (Ozuna et al., 2019) and APERO (Leonard et al., 2019) are lighter tools to install, both written in R and requiring only the most commonly used BAM format alignment files and relevant reference annotations. Like Rockhopper, the output of *baerhunter* is a transcriptional map (in gff format) and can consolidate annotations from multiple samples. APERO exploits improvements in sequencing technology by requiring paired-end reads (where each fragment is sequenced from both ends, creating two barcoded reads for each fragment) and optimising parameters for non-fragmented libraries. The output consists of a set of flat files of the predicted transcript 5' and 3' ends for each sample that can then be filtered for read counts and assembled into a genomic context.

Steps can be taken to lend support to computational predictions of sRNAs and 5' UTRs in mycobacteria. In a recent study to identify differentially expressed, verifiable sRNAs in *M. tuberculosis*, software predictions based on RNA-seq produced over 200 candidate sRNAs (Dejesus et al., 2017), 82 of which were differentially expressed by 6-fold in at least one experimental condition (Gerrick et al., 2018). Applying additional filters to the 92 'stable ncRNAs' listed in Mycobrowser, we compared their 5' boundaries with a compendium of published

predicted TSSs (Cortes et al., 2013; Shell et al., 2015), and found 40 with predicted TSS within 10 nucleotides of the annotated 5' boundary. 62 of the Mycobrowser ncRNAs are putative sRNAs originating from the DeJesus et al. study (DeJesus et al., 2017), 25 of which have TSSs within 10 nucleotides of the 5' boundary. These putative sRNAs were also compared with the transcripts found to be differentially expressed in Gerrick et al (Gerrick et al., 2018), and found 17 putative ncRNAs with both TSSs and differential expression (A2.1 Supplemental Tables: Ch2_Supp_Table_2). However, the available lists of *M. tuberculosis* TSS sites (Cortes et al., 2013; D'Halluin et al., 2023; Shell et al., 2015) have so far been mapped only in starvation and exponential growth and may not include TSSs that are utilised under different experimental conditions. Furthermore, 3' UTRs that are functionally independent from their cognate coding sequence (CDS) have been identified in other bacteria (Desgranges et al., 2021; Menendez-Gil et al., 2020; Ponath et al., 2022) and those generated from the 3' UTRs of coding genes through RNase processing would presumably lack a TSS. RNase cleavage sites could also lend support to the existence of other sRNA candidates cleaved from longer transcripts or otherwise processed (Martini et al., 2019; Zhou et al., 2023). Finally, polycistronic transcripts often include non-coding sequence between the genes of an operon, and this may contain functional elements and/or processing sites (DeLoughery et al., 2018; Durand et al., 2015; Martini et al., 2019; Zhou et al., 2023).

Sequence conservation of non-coding elements in mycobacterial genomes outside the MTBC can help to identify *bona fide* predictions made by RNA-seq methods. A comprehensive analysis of the genomic context, structural conservation and expression profiles of non-coding RNA homologues both within the MTBC, and in the wider phyla, would be a valuable resource for the mycobacterial research community. In the absence of such a resource, I have performed a sequence similarity search with each of the non-coding RNAs annotated in *M. tuberculosis* in three related genomes: one member of the MTBC (*M. bovis*), the non-pathogenic strain widely used surrogate for *M. tuberculosis*, *M. smegmatis*, and a pathogenic species outside the MTBC, *Mycobacterium abscessus*, using the web-based application, *fastA* (Madeira et al., 2019). 43 of the 92 Mycobrowser ncRNAs have significant (E-value < 0.01) sequence matches in both *M. smegmatis* and *M. abscessus* with sequence identities ranging from 52-87% (A2.1 Supplemental Tables:

Ch2_Supp_Table_1). 18 of these have been experimentally verified by Northern blot, but 25 of them were predicted by RNA-seq methods alone.

A further complication in defining the non-coding transcriptome is that putative non-coding elements predicted by computational algorithms may actually be (or contain) as yet unannotated ORFs; there is no way of asserting from the RNA-seq signal alone whether a transcript is coding or non-coding. Early ribosome profiling studies pointed to the presence of hundreds of small peptides encoded in the 5' UTR of mycobacterial transcripts (Shell et al., 2015), and more recent efforts have shown pervasive translation in *M. tuberculosis*, uncovering over 1000 novel ORFs (Smith et al., 2022). The majority of these were short ORFs (sORFs) with non-canonical features that would thus be missed by regular gene prediction algorithms. Comparing this list with the annotated ncRNAs listed in Mycobrowser, two of the ncRNAs overlap with predicted sORFs (A2.1 Supplemental Tables: Ch2_Supp_Table_2). Although translation of these transcripts does not necessarily render them functional, they may constitute a pool of peptides that are available to use under the right conditions. The observation that leaderless transcripts are translated more efficiently under stress conditions (Sawyer et al., 2021) also points to the fact that mycobacterial non-canonical ORFs may play increasingly important roles in conditions of nutrient starvation or other stresses. Ribosome profiling will continue to be instrumental in resolving ambiguities in annotation of ORFs versus non-coding elements in untranslated regions. Although such information can already be integrated in a subset of computational pipelines (Yu et al., 2018), the corresponding data is only available for a limited number of reference mycobacterial strains.

2.3.4 Using WGCNA to implicate functional associations of non-coding RNA

To include a complete picture of the interaction of the non-coding genome with coding genes involved in adaptation pathways, we have generated a novel set of ncRNA sequence-based predictions (sRNAs and UTRs) from publicly available *M. tuberculosis* datasets using our in-house software package, *baerhunter* (Ozuna et al., 2019). Some of these predicted non-coding transcripts overlap with those of previous studies, but many represent novel predictions. The expression of these transcripts is quantified along with the protein-coding genes and used in network

analysis to provide a more complete picture of the functional groupings involved in adaptation to environmental changes. Including a variety of culture conditions that replicate aspects of the host environment improves the chances that the expression of any ncRNA that is restricted to one or more conditions is included in the network (Ami et al., 2020).

Weighted gene co-expression network analysis (WGCNA) (Zhang & Horvath, 2005) has been widely used to identify functional groups of genes, called ‘modules’, through the application of hierarchical clustering to differential expression levels of RNA transcripts in microarray or RNA-seq experiments. Recent studies have focussed entirely on the protein-coding portion of the transcriptome, using WGCNA with RNA-seq to cluster the differentially expressed genes of *Mycobacterium marinum* in response to resuscitation after hypoxia (Jiang et al., 2020) and *Mycobacterium aurum* infected macrophages (Lu et al., 2021). *M. tuberculosis* microarray data have been used to cluster protein-coding genes that show differential expression among clinical isolates (Puniya et al., 2013) and in response to two different hypoxic models to identify potential transcription factors (Jiang et al., 2016). Another recent network analysis, using a matrix deconvolution method followed by module clustering, uses a large number of *M. tuberculosis* RNA-seq samples including deletion mutants, infection models and antibiotic-treated samples as well as restricted media and culture conditions (Yoo, et al., 2022). Here the authors identify 80 modules of protein-coding genes that each approximate an isolated source of variance, together estimated to account for 61% of the total variance seen in the dataset. This proportion is reportedly lower than results from similar analyses in other organisms, potentially due to the bias in the types of conditions available in the database and/or the complex nature of regulation in *M. tuberculosis* (Yoo, et al., 2022). However, the contribution of regulatory ncRNA elements may be a considerable unexplored source of variance in this complex system. Here we use an alternative, complementary approach by including ncRNA, as well as annotated protein-coding genes, in the modules.

In this study, WGCNA was applied to multiple *M. tuberculosis* H37Rv datasets covering 15 different culture conditions replicating various growth conditions, nutrient sources and stressors encountered in the host environment. A global view

of the non-coding genome across an extensive WGCNA network is presented and selected modules are interrogated to identify functional groupings between protein-coding and non-coding transcripts, as well as between well-characterised genes and those with little functional annotation. The correlation of the modules with the various conditions can identify participants in large-scale transcriptomic remodelling programs in response to changes in environmental conditions.

2.4 MATERIALS AND METHODS

The overall workflow for this analysis is presented in Figure 2.1. All scripts for *baerhunter*, WGCNA and subsequent analysis are available at: <https://doi.org/10.5281/zenodo.7319853>.

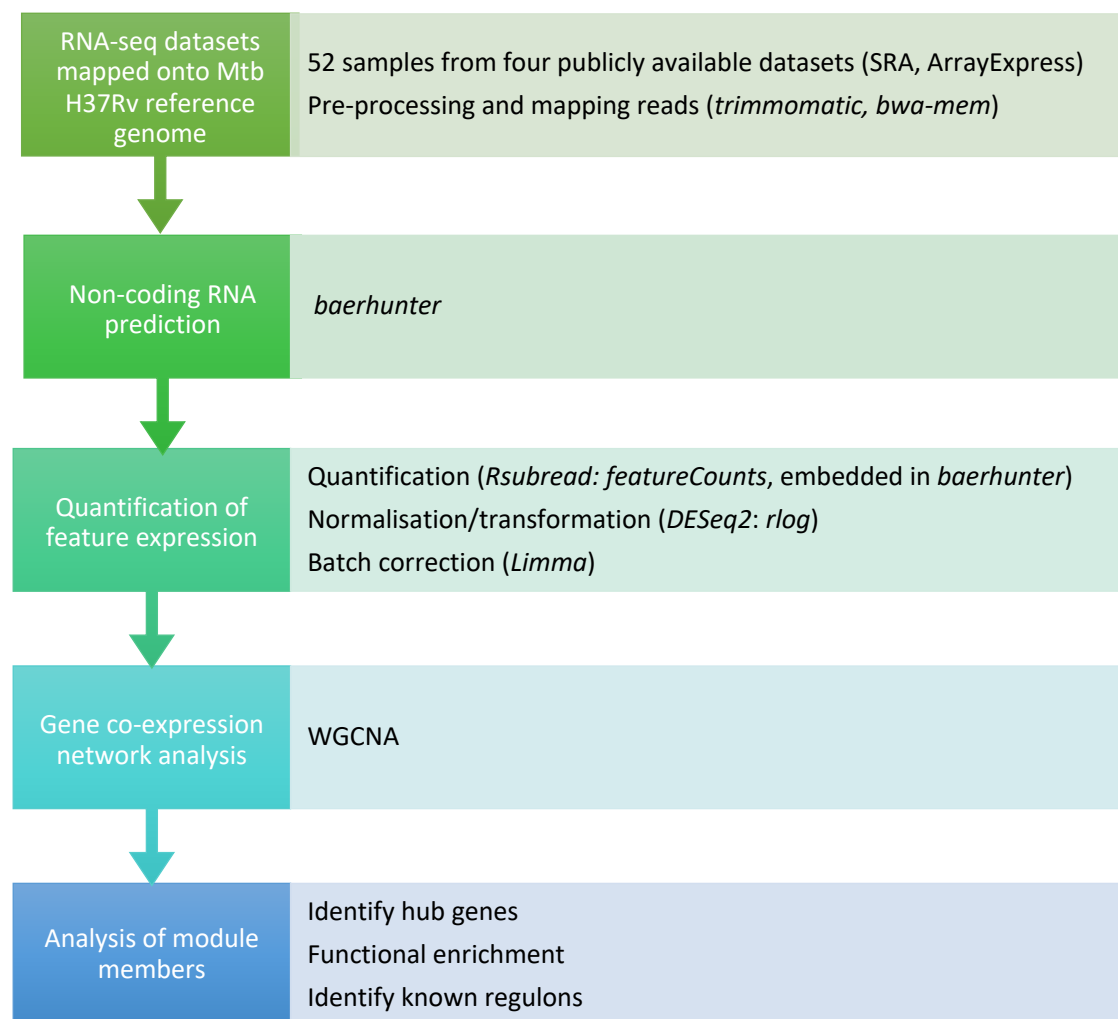


Figure 2.1. Analysis workflow.

2.4.1 Data Acquisition and Mapping

Datasets were downloaded from SRA (<https://www.ncbi.nlm.nih.gov/sra/docs/>) or Array Express (<https://www.ebi.ac.uk/arrayexpress/>) using the accession numbers listed in Table 2.4. To minimise batch effects and ensure compatibility with RNA prediction software, we limited analysis to datasets with similar library strategies. Samples were included based on inspection to confirm that 1) samples were from monocultures of wild-type *M. tuberculosis* H37Rv strain and 2) sequencing was using a paired-end, stranded protocol. Reads from samples that passed quality control thresholds were trimmed using *Trimmomatic* (Bolger et al, 2014) to remove adapters and low-quality bases from the 5' and 3' ends of the sequences. Trimmed reads were mapped to the H37Rv reference genome (GenBank AL123456.3) using *BWA-mem* in paired-end mode (Li, 2013). All samples had >70% percent reads mapped with an overall mean of ~ 27.75M mapped reads and a range of 3.97M to 60.68M mapped reads per sample (A2.1 Supplemental Tables: Ch2_Supp_Table_3).

Table 2.4. Datasets used in analysis. Project accession numbers from SRA and Array Express.

Dataset	Num of samples	Instrument	Library Layout	Library Strand	Library Strategy	Avg Spot Length	Ribo depleted
PRJEB65014_3 E-MTAB-6011	3	Illumina MiSeq	paired end	reversely stranded	cDNA	150	Y
PRJNA278760 GSE67035	22	Illumina HiSeq 2000	paired end	reversely stranded	cDNA	50	Y
PRJNA327080 GSE83814	15	Illumina HiSeq 2000	paired end	reversely stranded	cDNA	180	Y
PRJNA390669 GSE100097	12	Illumina NextSeq 500	paired end	reversely stranded	cDNA	287	N

2.4.2 Non-coding RNA prediction and quantification

Each dataset was run through the R-package, *baerhunter* (Ozuna et al., 2019), using the 'feature_file_editor' function optimised to the most appropriate parameters for the sequencing depth https://github.com/jenjane118/mtb_wgcna. 'Count_features'

and *'tpm_norm_flagging'* functions were used for transcript quantification and to identify low expression hits (less than or equal to 10 transcripts per million) in each dataset, which were subsequently eliminated. When viewed on a genome browser, coverage at the 3' ends of putative sRNA and UTRs often appears to decrease gradually, with the actual end of the transcript appearing indistinct, compared to the 5' end. Prokaryotic ncRNA transcripts may not demonstrate a clear fall-off of expression signal in RNA-seq due to incomplete RNAP processivity and pervasive transcription regulated by the changing levels of Rho protein observed in different conditions (Bidnenko & Bidnenko, 2018; Wade & Grainger, 2014). These very long predictions can mask predicted transcripts in the same region from other samples, obscuring potentially interesting shorter transcripts expressed in different conditions. For this reason, transcripts longer than 1000 nucleotides were eliminated before combining the predictions between datasets. The predicted annotations for each dataset were combined into a single annotation file, adding the union of the predicted boundaries to the reference genome for H37Rv (AL123456.3). Predictions that overlapped with annotated ncRNAs and UTR predictions that overlapped sRNA predictions from a different dataset were eliminated. Transcript quantification was repeated on each dataset using the resulting combined annotation file and the count data from each dataset was merged into a single counts matrix.

DESeq2 v1.30.1 (Love et al., 2014) was used on the complete counts matrix including the filtered *baerhunter* predictions to calculate size factors, estimate dispersion and normalise the data with the regularised log transformation function (Appendix A2.1, Figures S1 and S2). The normalised data was checked for potential batch effects using PCA plots and hierarchical dendrograms. *Limma* v3.46.0 (Ritchie et al., 2015) *'removeBatchEffect'* was applied with a single batch argument to remove batch effects associated with the first component (batching the data according to dataset due to technical differences) while preserving differences between samples. The final hierarchical dendrogram, post-batch correction, indicates successful application as samples cluster by similar experimental conditions, rather than by dataset alone (Figure 2.2 compared to A2.2 Supplemental Figures, S3). Samples from experiment PRJEB65014 continue to group together, but as they represent single replicates in unique conditions, it is difficult to estimate the influence of confounding

batch effects for these samples. The normalised, batch-corrected data is accessible as an R data object at https://github.com/jenjane118/mtb_wgcna/tree/master/R_data.

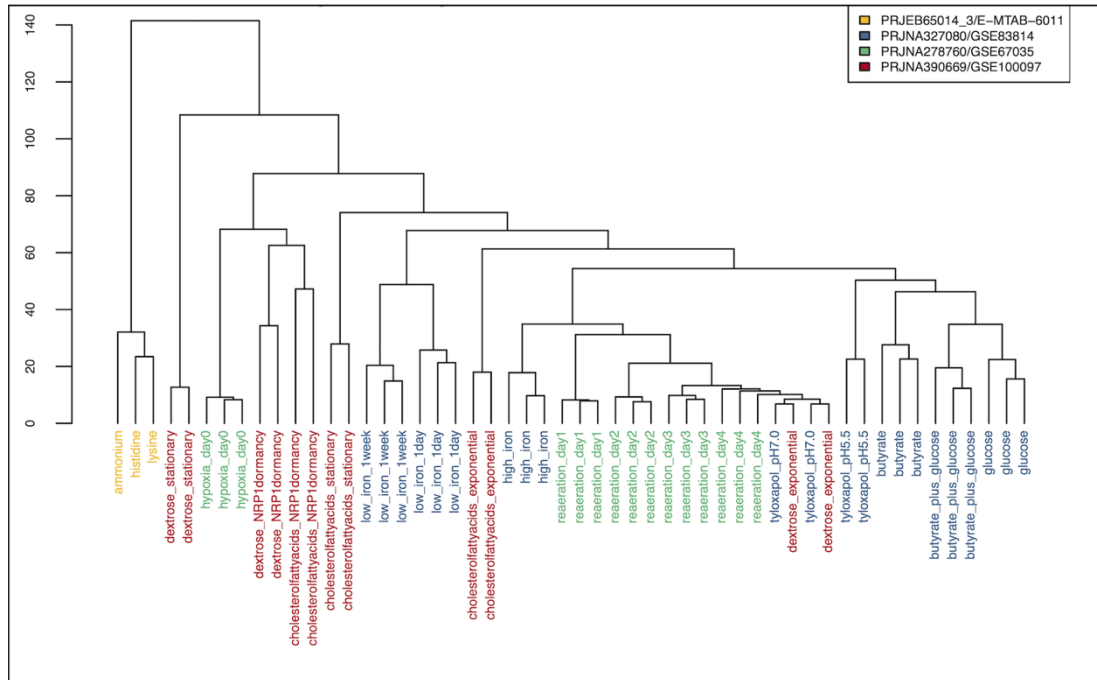


Figure 2.2. Hierarchical dendrogram of rlog transformed and limma batch corrected expression data by sample. The sample labels are coloured by dataset, demonstrating that they are clustering by condition, rather than experiment.

2.4.3 Creation of the WGCNA network

The normalised and batch-corrected expression matrix was used to create a signed co-expression network using the R package, *WGCNA* v1.69 (Langfelder & Horvath, 2008), with the following parameters: corType = "pearson", networkType = "signed", power = 12, TOMType = "signed", minModuleSize = 25, reassignThreshold = 0, mergeCutHeight = 0.15, deepSplit = 2. In this type of network, the 'nodes' are the genes, and the 'edges', or links, are created when gene expression patterns correlate. In contrast to unweighted binary networks where links are assigned 0 or 1 to indicate whether or not the genes are linked, in a weighted network the links are given a numeric weight based on how closely correlated the expression is. *WGCNA* first calculates the signed co-expression similarity for each gene pair. The absolute value of this correlation is raised to a power (determined by the user, based on a scale-free topology model that mimics biological systems (A2.2 Supplemental Figures, S4) in order to weight the strong connections more highly than the weaker connections. The resulting similarity matrix is used to cluster groups of genes with strong connections to each other in a non-supervised manner

(i.e., it doesn't use any previous information about gene groups or connected regulons). A cluster dendrogram is created (A2.2 Supplemental Figures, S8) and closely connected branches of the dendrogram are merged into modules based on a cut-off value (also a parameter controlled by the user). Pairwise correlations were calculated between all of the genes in each module, and between module 'hubs' and all of the other genes in the module, using the Pearson correlation coefficient. The mean of these values for each module are available in A2.1 Supplemental Tables: Ch2_Supp_Table_5. The modules are defined by a 'module eigengene' (ME), which explains most of the variance in the expression values in the module. The connectivity of the MEs define the shape of the overall network (A2.2 Supplemental Figures, S9). The modules can then be tested for potential correlations with experimental conditions while reducing the degree of penalties for multiple testing. In signed networks, correlation of the module with a condition can be in either the positive or negative direction, as modules include transcripts that are similar in both the degree and direction of correlation, allowing for a more fine-grained analysis than with unsigned networks (A2.2 Supplemental Figures, S10).

To test correlations of modules with experimental conditions, the individual RNA-seq samples were assigned to a condition based on the experimental description in the project metadata. Some of these conditions were shared among the different projects, so when appropriate, samples from different datasets were assigned the same condition, resulting in 15 tested conditions. For example, late-stage reaeration samples were tested along with exponential growth samples, and samples that tested hypoxia and cholesterol utilisation together were included in multiple conditions. Models of hypoxia differed between the RNA-seq projects, and these samples were assigned to different conditions: 'hypoxia' versus 'extended hypoxia' (Table 2.5). Network correlations were made using robust biweight midcorrelation tests and all p-values were corrected for multiple testing with the Benjamini-Hochberg (BH) method (Benjamini & Hochberg, 1995). Significance was evaluated as an adjusted p-value (p_{adj}) of < 0.05 .

2.4.4 Module Enrichment

Modules were interrogated for enrichment for Gene Ontology (GO) terms (Ashburner et al., 2000; The Gene Ontology Consortium, 2021), Clusters of

Orthologous Groups (COG) (Galperin et al., 2021), KEGG pathway genes (Kanehisa et al., 2022), functional categories and literature searches for known regulons. GO terms, COG terms and KEGG pathway enrichment were accessed programmatically using the DAVID web service (Huang et al., 2009b, 2009a; Jiao et al., 2012) to query the list of protein-coding genes from each module for enrichment. Enrichment was determined using a modified one-sided Fisher's Exact Test ('EASE' score) with BH correction for multiple testing, with $p_{adj} < 0.01$ considered significantly enriched for a particular term, pathway or COG term. Enrichment for the 11 functional categories from Mycobrowser annotation (Kapopoulou et al., 2011) was determined using a one-sided Fisher's Exact Test with BH correction for multiple testing. Modules were enriched for a particular functional category if $p_{adj} < 0.01$. Lists of genes associated with known regulons were mined from literature and enrichment was tested using the same one-sided Fisher's Exact Test as above with a $p_{adj} < 0.01$ cut-off for enrichment.

2.4.5 Data exploration

Non-coding RNA prediction, network analysis and subsequent data manipulation was performed with R (v4.0.5, 2021-03-31). All plots were made in R with the following packages: *WGCNA* (v1.69), *dendextend* (v1.15.2), *ggplot2* (v3.3.5). Scripts and expression data are available at https://github.com/jenjane118/mtb_wgcna. A downloadable app (https://github.com/jenjane118/mtb_wgcna) was made to query and explore the network using R *shiny* (v1.7.5), *JBrowseR* (v0.10.2) and *shinyjs* (v2.1.0).

Table 2.5. Conditions tested in the various samples used in this analysis. Sample run acquisition numbers from (NCBI Sequence Read Archive, <https://www.ncbi.nlm.nih.gov/sra>)

Condition	Number of samples	Samples
ammonium	1	ERR2103718
histidine	1	ERR2103722
lysine	1	ERR2103723
hypoxia	4	SRR5689228, SRR5689229, SRR5689234, SRR5689235
extended hypoxia	3	SRR3725585, SRR3725586, SRR3725587
reaerated culture	6	SRR3725588, SRR3725589, SRR3725590, SRR3725591, SRR3725592, SRR3725593
exponential	10	SRR3725594, SRR3725595, SRR3725596, SRR3725597, SRR3725598, SRR3725599, SRR1917712, SRR1917713, SRR5689224, SRR5689225
butyrate	3	SRR1917694, SRR1917695, SRR1917696
butyrate and glucose	3	SRR1917697, SRR1917698, SRR1917699
glucose	3	SRR1917700, SRR1917701, SRR1917702
high iron	3	SRR1917703, SRR1917704, SRR1917705
low iron	6	SRR1917706, SRR1917707, SRR1917708, SRR1917709, SRR1917710, SRR1917711
acid	2	SRR1917714, SRR1917715
cholesterol	6	SRR5689230, SRR5689231, SRR5689232, SRR5689233, SRR5689234, SRR5689235
stationary	4	SRR5689226, SRR5689227, SRR5689232, SRR5689233

2.5 RESULTS AND DISCUSSION

2.5.1 *M. tuberculosis* expresses an extensive range of ncRNA transcripts over a wide variety of experimental conditions

M. tuberculosis RNA-seq datasets were selected from publicly available data to find experiments using the wild-type H37Rv strain and representing a range of growth conditions the pathogen may encounter in a host environment. Four datasets passing our quality standards were subjected to our analysis pipeline (see Material and Methods) and included 52 samples under 15 different experimental conditions (Table 2.5). The R package, *baerhunter* (Ozuna et al., 2019), was used to predict ncRNA in intergenic regions, antisense RNA (opposite a protein-coding gene) and UTRs at both the 5' and 3' ends of genes by searching the mapped RNA-seq data for expression peaks outside of the annotated regions in the reference sequence for H37Rv. Non-coding RNA predictions from each dataset were filtered for low expression and combined to create a single set of non-overlapping annotations that encompassed all predictions made from any sample under any experimental condition. In total, 1283 putative sRNAs were predicted (including both truly intergenic transcripts as well as those antisense to a protein-coding gene, or annotated RNA) and 1715 UTRs which includes all transcribed regions outside of annotated protein-coding sequences at both 5' and 3' ends, as well as the non-coding regions between adjacent genes in operons. All putative ncRNA transcripts (sRNAs and UTRs) were searched for a TSS near the start of the predicted 5' boundary using previously published annotations (Cortes et al., 2013; Shell et al., 2015). Annotated TSSs were found within 20 nucleotides of the 5' end in 43% of the predicted sRNA transcripts. Predicted 5' UTRs had a TSS within 10 nucleotides of the start in 42% of cases, compared with 3% of the predicted 3' UTRs. Where the UTR covered the entire sequence between two protein-coding regions (labelled as 'between' UTRs), 9% had a TSS in the first 10 nucleotides of the sequence (Table 2.6 and A2.1 Supplemental Tables: Ch2_Supp_Table_4, 'putative_UTRs').

Table 2.6. Tally of predicted expressed elements in the baerhunter-generated combined annotation file. 4018 protein-coding genes were included in the annotation. ‘Between’ UTRs cover the entire sequence between two protein-coding regions. *TSS predictions under exponential and starvation conditions from (Cortes et al., 2013; Shell et al., 2015).

Predicted element	Number predicted	With predicted TSS*
Total sRNA	1283	553
sRNA ‘intergenic’	88	23
sRNA ‘antisense’	1195	530
Total UTRs	1715	273
5’ UTRs	475	200
3’ UTRs	602	16
‘Between’ UTRs	638	57

The predicted sRNAs were further annotated using the accepted nomenclature (Lamichhane et al., 2013) which identifies the putative ncRNA relative to annotated gene loci and differently signifies truly intergenic sRNAs and those that overlap any part of a protein-coding region on the opposite strand. Most of the putative sRNAs are antisense to the protein-coding region of one or more genes, but 88 putative sRNAs have predicted boundaries that do not overlap an annotated transcript on either strand (or overlap an annotated transcript on the opposite strand by fewer than 10 nucleotides). This number is most probably an underestimate of the truly ‘intergenic’ sRNAs in the genome, as many of the sRNA predictions appear over-estimated at the 3’ end, effectively classifying them as an antisense RNA even though the 5’ half of the transcript does not overlap any genes on the opposite strand. Isoforms of annotated sRNAs can be subject to post-transcriptional processing to create an active transcript (Moores et al., 2017) and post-transcriptional processing of 3’ ends *in vivo* is more likely the norm for most prokaryotic transcripts (Wang et al., 2019). However, for our purposes, any RNA-seq transcripts that extend to overlap a protein-coding gene on the other strand in any dataset will be labelled as antisense RNA.

The generated combined annotation file was used to quantify the expression of all 7046 expressed elements, including every annotated CDS, annotated ncRNA and predicted ncRNA, in each sample. Raw counts of expression varied greatly among the datasets due to different sequencing depth, as well as between some samples within datasets (as would be expected with different environmental conditions).

The raw expression counts were transformed using DESeq2's *rlog* function (Love et al., 2014), and plots of the dispersion of count data show that the median expression level between samples and between datasets has been normalised (A2.2 Supplemental Figures, S1 and S2). The distribution of the normalised expression levels of protein-coding regions alone shows consistent median expression levels across the entire dataset, however distribution of the normalised data restricted to putative sRNAs shows more variability, with certain conditions showing increased or decreased expression of these transcripts (A2.2 Supplemental Figures, S5-S7). This is not unexpected, given that several studies have identified pervasive transcription in hypoxic infection models, stationary phase and dormancy. This is accompanied by a concomitant increase in non-rRNA abundance (especially antisense RNA transcripts) and in the number of predicted TSSs in *M. tuberculosis* and *M. smegmatis* (a fast-growing, non-pathogenic strain) (Arnvig et al., 2011; Ignatov et al., 2015; Martini et al., 2019).

2.5.2 Module networks represent groups of co-expressed genes and predicted non-coding RNA

A weighted co-expression network was created from the normalised RNA-seq expression data using *WGCNA* (Langfelder & Horvath, 2008) (see Materials and Methods). This program segregates transcripts with similar patterns of expression over a range of samples into modules. The modules represent sub-networks of connected genes, and functional relationships can be explored among the members of the individual modules. The 'hub' genes represent the most highly connected genetic elements within a module and have highest module membership values. Module membership (MM) is measured by correlation of the expression of the individual genes with the module eigengene (ME), the vector that best represents the variation in the module. This value is highly correlated with the level of interconnectivity between the gene and the other genes of the module and can be used to find the best-connected genes in the module.

The signed co-expression network presented in this paper consists of 54 different modules, assigning 99.3% of the expressed elements (CDS, putative UTRs and putative sRNAs) into 53 modules, with 46 unassigned elements clustered in the 'grey' module (A2.1 Supplemental Tables: Ch2_Supp_Table_4, 'Module Overview').

Module size ranged from 766 to 25 expressed elements. The modules (using the ME) were tested for correlations with the various conditions used in the RNA-seq experiments (see Materials and Methods). The RNA-seq data was categorised into 15 different experimental conditions in total with varying numbers of replicates (Table 2.5), therefore, a statistically significant correlation of modules with every condition was not expected. However, some modules do show significant correlations with conditions such as iron restriction, cholesterol media, hypoxia and growth phase and this can be informative when considering the association of the gene groups with biological processes (Figure 2.3).

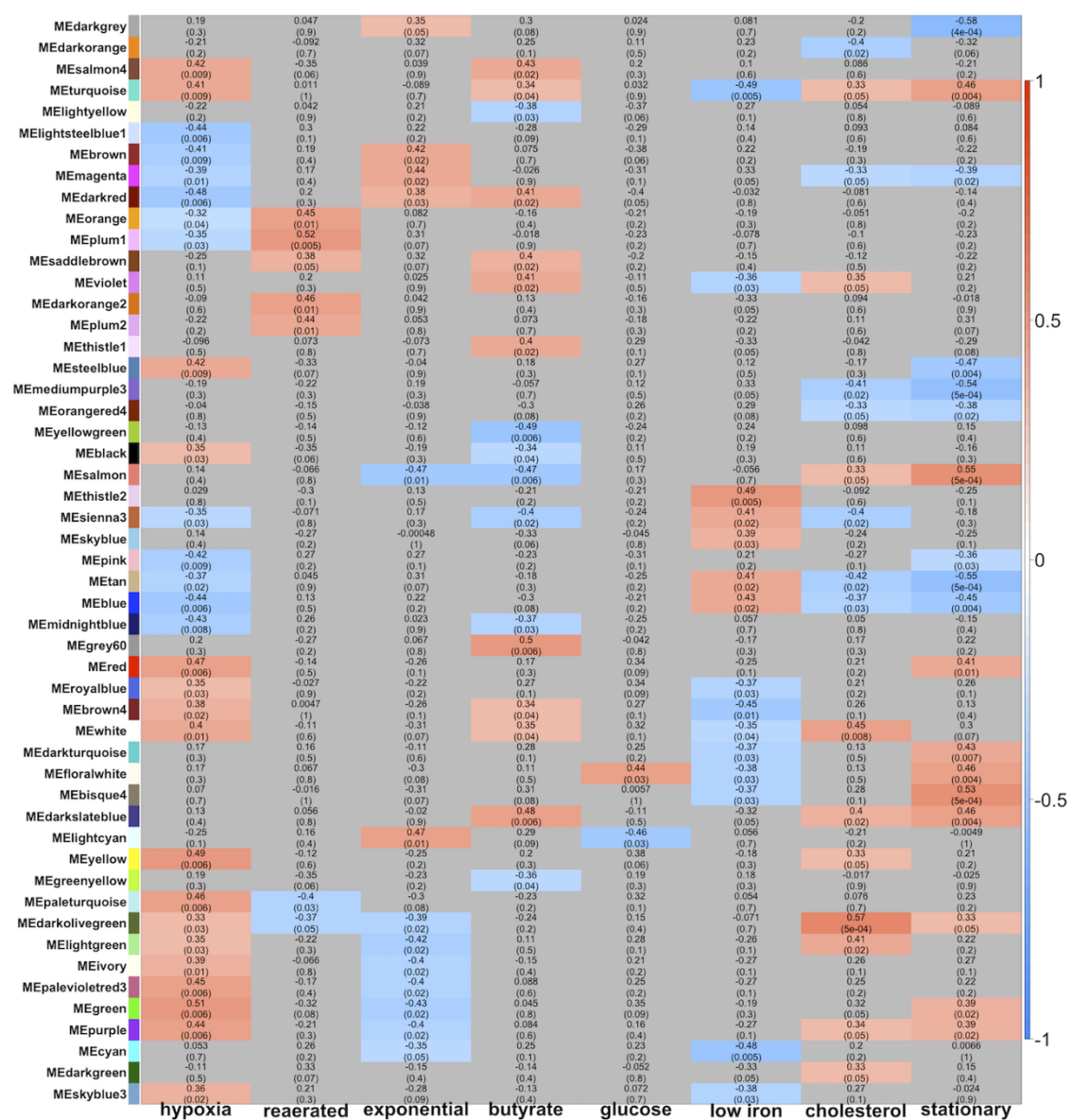


Figure 2.3. Heat map of correlation of module eigengene (ME) of each module with selected experimental conditions. Correlation was calculated using biweight midcorrelation (bicor) and p-values were adjusted for multiple testing (BH-fdr). Positive correlation is red, negative correlation is blue. Non-significant correlations in grey ($p_{adj} > 0.05$).

2.5.2.1 Well-established regulons cluster together in single modules

In many cases, the gene membership of the modules includes well-established regulons or groups of functionally related genes, establishing the biological relevance of the module sub-networks and proof of concept for the application of WGCNA on such a heterogenous dataset. For example, the DosR regulon is a well-studied regulon associated with hypoxia and stress responses (Du et al., 2016; Rustad et al., 2008; Voskuil et al., 2004). 47 of 48 previously identified DosR-regulated genes are found in a single module, 'cyan', representing statistically significant enrichment of DosR-regulated genes in the module (one-sided Fisher's

exact test, $p_{\text{adj}}=3.81\text{e-}53$). The '*cyan*' module also includes 5 genes from the PhoP regulon which is associated with hypoxic response and coordination with the DosR regulon (Gonzalo-Asensio et al., 2008; Singh et al., 2020) and the DosR-regulated ncRNA, DrrS/MTS1338, known to be upregulated in hypoxic conditions (Ignatov et al., 2015; Moores et al., 2017). Unsurprisingly, the '*cyan*' module is enriched for the GO term, 'response to hypoxia', however, a statistically significant correlation was not seen with the hypoxia condition (though it is negatively correlated with the exponential growth condition, $\text{bicor}=-0.35$, $p_{\text{adj}}=0.05$) (Figure 2.3). The KstR regulon includes 74 genes under control of the TetR-type transcriptional repressor, KstR, known to be involved in lipid catabolism and upregulated during infection (Kendall et al., 2007, 2010; Nesbitt et al., 2010). The '*royalblue*' module is significantly enriched for known KstR-regulated genes (one-sided Fisher's exact test, $p_{\text{adj}} = 5.06\text{e-}30$) with 30 of 72 KstR-regulated genes clustering together in the module. This module is enriched for genes of the KEGG pathway for steroid degradation ($p_{\text{adj}}=3.32\text{e-}10$) and the GO term 'steroid metabolic process' ($p_{\text{adj}} = 5.62\text{e-}16$). The module shows statistically significant positive correlation for hypoxia ($\text{bicor}=0.35$, $p_{\text{adj}}=0.03$) and negative correlation with the low iron condition ($\text{bicor}=-0.37$, $p_{\text{adj}}=0.03$) (Figure 2.3). Genes involved in mycobactin synthesis are nearly all found in the '*grey60*' module (one-sided Fisher's Exact test, $p_{\text{adj}}= 1.23\text{e-}17$), a module enriched for the KEGG pathways 'siderophore metabolic processes' and 'arginine biosynthesis'. As these examples show, known associated genes are co-located in modules which represent a functional group of genes that have co-regulated expression under various experimental conditions. The modules can be further explored to identify novel associations.

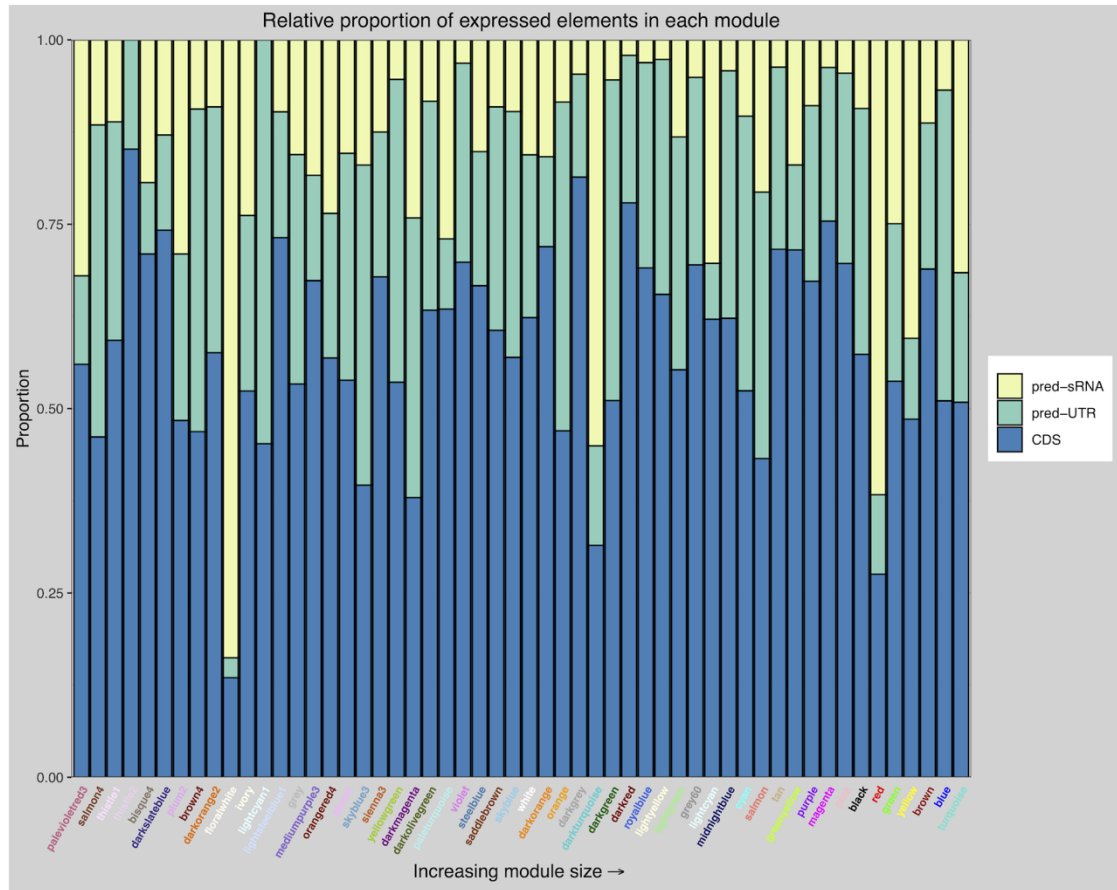


Figure 2.4. Relative proportion of annotated CDS, predicted UTRs and predicted sRNAs in each module, ordered by module size.

2.5.2.2 Predicted non-coding RNAs are enriched in certain modules

Putative sRNAs and/or predicted UTRs were distributed throughout all modules in the network (Figure 2.4). The number of predicted sRNAs were statistically enriched in seven modules and predicted UTRs enriched in another seven modules (one-sided Fisher's exact test, $p_{adj} < 0.01$, A2.1 Supplemental Tables: Ch2_Supp_Table_4, 'Module Overview'). A roughly linear relationship between the number of CDS and the number of UTRs, is to be expected, given that UTRs are defined by the *baerhunter* algorithm by their position at the start or end of protein-coding genes (Ozuna et al., 2019). However, if the UTRs are positioned in an operon, there will be a smaller increase in the relative number of UTRs with an increasing number of protein-coding genes, as UTRs between two protein-coding genes are predicted as a single UTR. As expected, the two modules that include the highest number of predicted operons (from OperonDB, Chetal & Janga, 2015), 'turquoise' and 'brown', have a lower relative proportion of UTRs; however, the 'blue' module, which includes 15 complete predicted operons, is significantly enriched for UTRs ($p_{adj} = 6.79e-21$) (Figure 2.5).

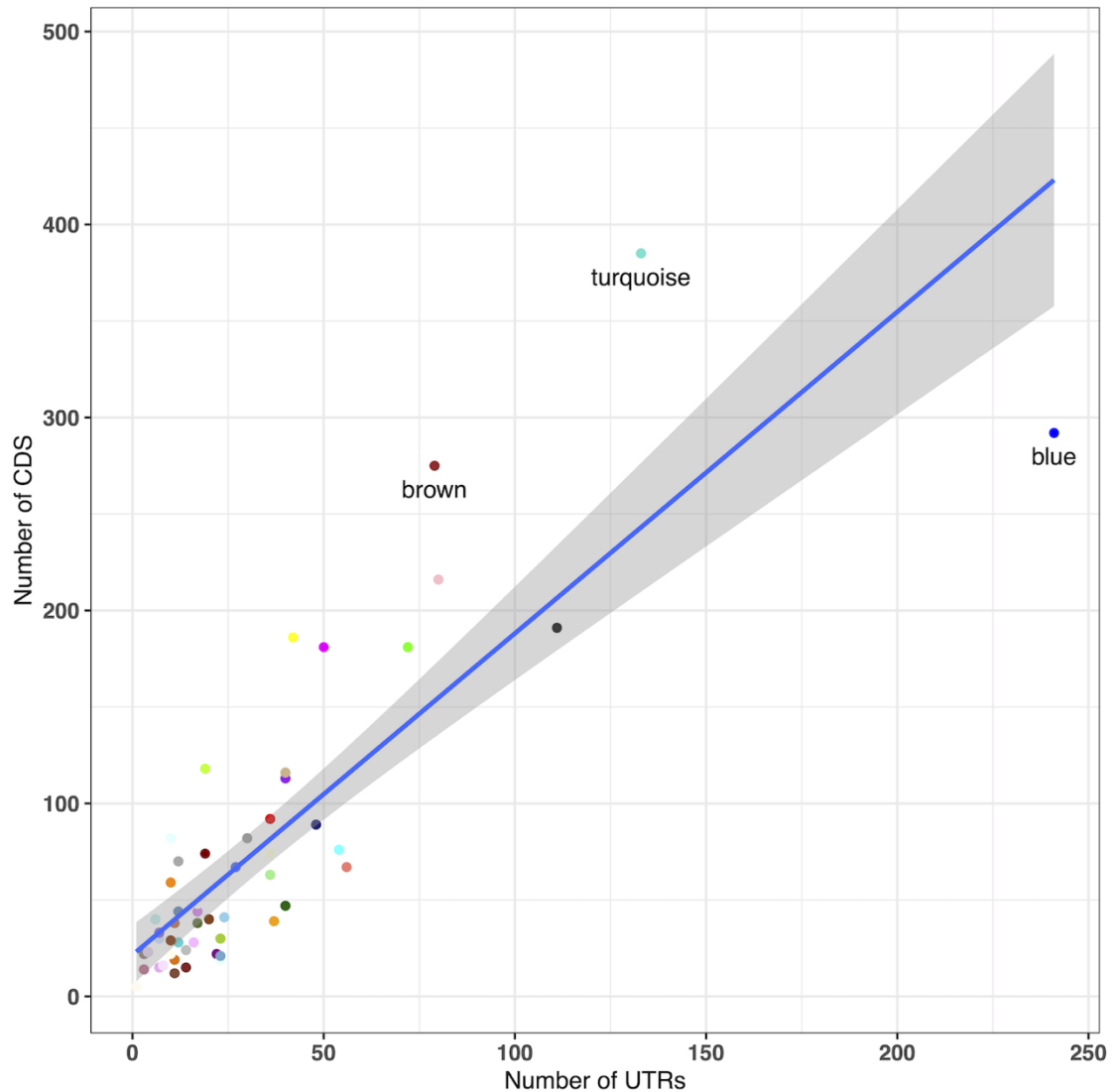


Figure 2.5. Plot of number of UTRs against number of CDS in each module. Grey shading indicates confidence interval of 0.95.

Within the module sub-networks, the tight co-expression of protein-coding genes and ncRNA is reflected by the number of ncRNA found among the most connected elements in the module. The ‘hub’ elements are those with the best correlation to the ME and therefore the most tightly connected elements in the individual module networks. In 14 modules, ncRNA (both predicted and annotated) make up more than half of the elements with module membership values (MM) ≥ 0.80 (our threshold for identifying hub elements) (A2.1 Supplemental Tables: Ch2_Supp_Table_4, 'Module Overview'). These associations may implicate ncRNA as co-conspirators in regulatory pathways implemented to adapt to conditions such as hypoxia, cholesterol media and low iron. The 30 annotated ncRNAs in the *M. tuberculosis* reference genome (AL123456.3) are spread over 20 modules, with 10

of them hubs of the module, and one unassigned ('grey' module) (A2.1 Supplemental Tables: Ch2_Supp_Table_4, 'Annotated ncRNA'). For example, Ms1/MTS2823, observed to be the most abundantly expressed ncRNA in expression studies over various stress conditions (Arnvig et al., 2011; Arnvig & Young, 2012; Ignatov et al., 2015; Šiková et al., 2019), is a hub element in a module that is positively correlated with cholesterol-containing media conditions ('darkgreen', bicor=0.35, p_{adj} =0.04) (Figure 2.3). This module is significantly enriched for KEGG pathways, including: Pyruvate metabolism (p_{adj} = 3.1e-3) and two-component systems (p_{adj} = 3.8e-3), and GO terms: plasma membrane respiratory chain complex II and plasma membrane fumarate reductase complex. Mcr7/ncRv2395A, found to be part of the PhoP regulon (Solans et al., 2014), is a hub in the 'violet' module enriched for lipid metabolism and PE/PPE functional categories, correlated positively with growth in cholesterol (bicor= 0.35, p_{adj} = 0.04) and butyrate (bicor= 0.41, p_{adj} = 0.02) and negatively correlated with low iron (bicor= -0.36, p_{adj} = 0.03) (Figure 2.3). F6/ncRv10243/SfdS, a sRNA upregulated in starvation and mouse infection models, is thought to be involved in regulating lipid metabolism and long-term persistence (Houghton et al., 2021). This ncRNA is a hub in a module found to be enriched in 'lipid metabolism' genes ('saddlebrown') and found to be correlated positively with reaerated culture (bicor= 0.38, p_{adj} = 0.04) and butyrate (bicor= 0.4, p_{adj} = 0.02) conditions (Figure 2.3).

2.5.2.3 UTR and adjacent ORF expression differ in over 50% of cases

We were interested to see how many of the predicted UTRs were assigned the same module as the adjacent ORF—indicating whether the ORF and its adjacent UTR were co-regulated. Intuitively, the UTR of a protein-coding gene would be expected to be expressed as a single transcript along with the ORF and show similar expression patterns. However, both 5' and 3' UTRs can act independently of the attached ORF and RNA abundance in RNA-seq experiments reflects both transcription activity and transcript stability. For example, some 5' UTRs are known to contain regulatory elements, such as riboswitches, that alter the transcription of the downstream ORF (Dar et al., 2016; Kipkorir et al., 2021; Schwenk & Arnvig, 2018; Warner et al., 2007), whereas sRNAs cleaved from 3' UTRs have been shown to regulate the stability of the remaining transcript—with different half-lives as a result (Chao et al., 2012; Dar & Sorek, 2018; Menendez-Gil & Toledo-Arana, 2021). Of the *baerhunter* -predicted

UTRs labelled 5' and 3', the UTRs co-segregated with the ORF they were closest to in fewer than half of cases (Table 2.7. UTRs and module assignment of adjacent ORFs). We would expect correctly-identified 5' UTRs to utilise a TSS (whether or not there is a known predicted TSS), whereas it appears functional 3' UTRs are more likely to be cleaved from the longer mRNA transcript (Dar & Sorek, 2018; Menendez-Gil & Toledo-Arana, 2021; Ponath et al., 2022). Our data confirms this: transcripts classified as 5' UTRs are much more likely to have a predicted TSS in the first 10 nucleotides than transcripts classified as 3' UTRs (42% vs 2.7%). Approximately 9% of the UTRs predicted to be between ORFs (labelled, 'Between' UTRs) have predicted TSS (Table 2.7. UTRs and module assignment of adjacent ORFs). The presence of a TSS in the first 10 nucleotides of the predicted UTR appeared to have little bearing on whether or not the UTR and its adjacent ORF are assigned to the same module, with 43% of 5' and 19% of 3' UTRs with a predicted TSS co-assigned with their adjacent ORF partner. 42% of the 'Between' UTRs do not segregate with either the ORF upstream or downstream, indicating their expression is, to some degree, independent of either adjacent ORF. 195 UTRs were found to be hubs in modules independent of their adjacent ORF(s), with 27 including a predicted TSS. All 'independent' UTRs are found in the A2.1 Supplemental Tables: Ch2_Supp_Table_4, 'independent_UTRs'.

Table 2.7. UTRs and module assignment of adjacent ORFs excluding those in 'grey' module (unassigned transcripts). DS=downstream, US=upstream. TSS indicates presence of annotated TSS in first 10 nucleotides of predicted UTR (Cortes et al., 2013; Shell et al., 2015).

	Total (excluding grey)	Number with TSS	Number in same module as adjacent ORF	Proportion of UTRs in same module as adjacent ORF
5' UTR	471	198	173 DS	37%
3' UTR	597	16	254 US	43%
BTWN UTR	633	56	112 DS	18%
			116 US	18%
			137 both	22%

2.5.2.4 Antisense RNAs are hubs in modules independent of cognate ORF

It has been observed that the overall abundance of antisense RNA and other non-ribosomal RNA increases upon exposure to stress such as hypoxia and nutrient restriction (Arnvig et al., 2011; Ignatov et al., 2015), and in our network, ncRNA are

well-connected in various modules that include known transcription factors and gene regulons associated with stress responses. Not unexpectedly, very few (5%) of the predicted antisense transcripts were assigned to the same module as the protein-coding region overlapping on the opposite strand (choosing the most downstream locus in the event of multiple overlapping ORFs), signifying distinct patterns of expression for transcripts on opposite strands, possibly due to independent or bi-directional promoters and/or overlapping transcription termination sites. Bi-directional promoters have been identified in multiple prokaryotic genomes, and competition for RNA polymerase (RNAP) binding among divergently transcribed sense/antisense pairs may function as a mechanism for regulation of gene expression (Ju et al., 2019; Warman et al., 2021). Long 3' UTRs that overlap with converging protein-coding genes on the opposite strand (or with the 3' UTR) can create an 'excludon' regulatory arrangement, where transcription of the two opposite mRNAs is simultaneously regulated by RNase targeting, or mutually exclusive due to RNAP collision (Sáenz-Lahoya et al., 2019; Toledo-Arana & Lasa, 2020). Examining the module groupings of the antisense RNAs and their base-pairing target on the other strand may provide insight on which genes are regulated by antisense transcription.

2.5.3 Focus on selected module networks

The large-scale transcription analysis presented here is useful for the more global analysis of the overall trends related to ncRNA and transcription, but there is a great deal of information to be gleaned by more fine-grained inspection of individual module groupings. To discover novel associations in such a large and complex dataset, we have selected a few modules for closer examination, focussing on those that contain gene groups or regulons related to the tested conditions. Many of the modules that contain interesting correlations or gene regulon enrichments also include an abundance of putative sRNAs and UTRs. Using the 'guilt by association' principle, we can hypothesise that the well-connected ncRNAs found among the module hub elements have a role in transcriptional 'remodelling' in response to changes in environmental conditions such as growth on cholesterol-containing media, restricted iron or hypoxia.

2.5.3.1 Detoxification-linked proteins cluster in the module best correlated with cholesterol media condition

The '*darkolivegreen*' module showed positive correlation with the cholesterol media condition (bicor=0.57, $p_{\text{adj}}=5.0\text{e-}04$) and negative correlation with low iron (bicor = -0.48, $p_{\text{adj}} = 0.001$) (Figure 2.3). Many protein-coding genes involved in detoxification pathways are hubs in the module, including several encoding transmembrane proteins such as the *mmpL5-mmpS5* efflux pump operon (Rv0676c-Rv0677c), as well as the next gene downstream, Rv0678, which was identified as part of a 'core lipid response' in differential expression analysis in lipid-rich media (Aguilar-Ayala et al., 2017). The 5' UTR for Rv0677c and 3' UTRs for Rv0676c and Rv0678 are also hubs. This operon is involved in siderophore transport and expressed in cholesterol and lipid-rich environments (Aguilar-Ayala, et al., 2017; Pawełczyk et al., 2021). The module contains several Type II toxin-antitoxin systems including VapBC12 (Rv1720c1721c), VapBC41 (Rv2601A-2602), RBE2 (relFG, Rv2865-2866) and vapB36 and vapB40 which may have roles in adaptation to cholesterol and the evolution of persisters (Ramage et al., 2009; Sala et al., 2014). VapBC12, specifically, has been shown to inhibit translation and promote persister phenotypes in response to cholesterol (Talwar et al., 2020). Other detoxification-linked genes in the module, such as the ABC-family transporter efflux system, Rv1216c-1219c, have also been implicated in transcriptomic remodelling in response to cholesterol (Aguilar-Ayala et al., 2017; Pawełczyk et al., 2021).

Two adjacent predictions, the 3' UTR for Rv1772 (putative_UTR:p2006948_2007063) followed by ncRv1773/putative_sRNA:p2007213_2007377, are hubs in the '*darkolivegreen*' module. Together, they extend to overlap the antisense strand of a large portion of Rv1773c, a probable transcriptional regulator in the IclR-family, found in a different module ('*turquoise*'). The 3' UTR for Rv1772 has been previously identified as an abundant antisense transcript during exponential growth (Arnvig et al., 2011). The start of the predicted sRNA transcript has no known TSS and could instead be an extension of the predicted 3' UTR (Figure 2.6). (When combining predicted annotations from different datasets, long predicted UTRs that overlapped shorter sRNA predictions were discarded, see Methods). In *E.coli*, the IclR-family transcriptional regulators demonstrate both activating and repressing activities on

targets such as multidrug efflux pumps and the *aceBAK* operon which regulates the glyoxylate shunt (Zhou et al., 2012). *Icl2a* (Rv1915) is one of the *M. tuberculosis* isoforms of the isocitrate/methylcitrate lyase gene, *aceA*, and may be regulated by Rv1773c, as seen in *E.coli*. *Icl2a*, Rv1772, its predicted UTR and the antisense RNA (ncRv1773) are all hubs in the ‘darkolivegreen’ module. *Icl2a* has been observed to be upregulated with cholesterol as the sole carbon source and likely has a second function as part of the methylcitrate cycle to convert the fatty acid metabolites propionate and propionyl CoA to less toxic compounds (Bhusal et al., 2017; Pawełczyk et al., 2021).

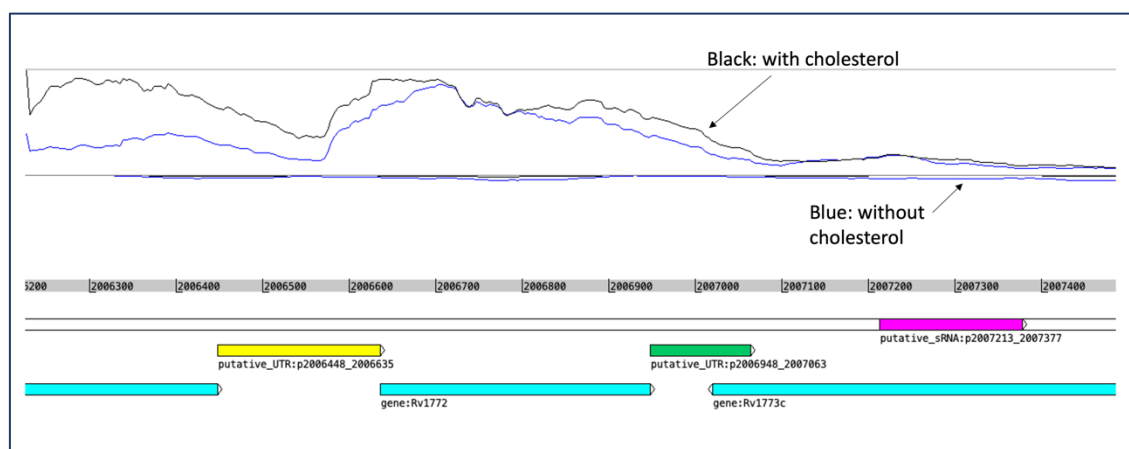


Figure 2.6. Expression of antisense transcripts, putative_UTR:p2006948_2007063 (highlighted in green) and ncRv1773/putative_sRNA:p2007213_2007377 (highlighted in magenta), appear to suppress expression of convergently transcribed gene, Rv1773c in cholesterol and fatty acid-containing media vs standard media without cholesterol. RNA-seq samples SRR5689230 and SRR5689224 from PRJNA390669. Strand coverage using the ‘second’ read of each pair mapping to the transcript strand, visualised using Artemis genome browser (Carver et al., 2012).

2.5.3.2 Module correlated with reaeration after non-replicating persistence includes genes for amino-acid synthesis and cell wall remodelling

The module, ‘saddlebrown’ is enriched for GO-terms for various amino-acid metabolic processes and COG ‘lipid metabolism’. It is positively correlated with reaeration after non-replicating persistence (bicor= 0.38, p_{adj} = 0.04) and butyrate-containing media (bicor= 0.4, p_{adj} = 0.02) (Figure 2.3). This pairing of upregulation of amino-acid synthesis and upregulation of the synthesis of cell wall lipids has been observed in the ‘lag phase’ after reaeration for increased protein synthesis (Du et al., 2016). The hubs of the ‘saddlebrown’ module include several predicted sRNAs, and the annotated sRNA, F6. F6/ncRv10243/SfdS is a sigF-dependent ncRNA which has been shown to be induced in nutrient starvation, oxidative stress, acid stress (Arnvig

& Young, 2009; Houghton et al., 2021) and the fatty acid hypoxia model (Del Portillo et al., 2019). In addition to being expressed from its own promoter, F6/SfdS has been proposed to be co-transcribed with the upstream gene *fadA2* (Rv0243), a probable acetyl-CoA acyltransferase; however, *fadA2* is clustered in a different module from SfdS ('darkred').

One of the predicted sRNAs among the 'saddlebrown' module hubs is antisense transcript ncRv2489/putative_srna:p2801108_2801678 with a TSS at 2801108. This overlaps the 3' end of PE-PGRS43 (Rv2490c) (Figure 2.7). There is a short reading frame (30 nucleotides, 10 amino acids) initiating from a Methionine at this TSS that suggests a possible dual-function sRNA or sORF with independent function. A shorter, possibly-leadered, sORF was predicted by Shell et al. (2015) that falls within this region (2801238..2801261). The TSS for the predicted sRNA overlaps the 5' end of Rv2489c, a short, hypothetical 'alanine-rich protein'. The TSSs for these convergently overlapping transcripts are 42 nts apart

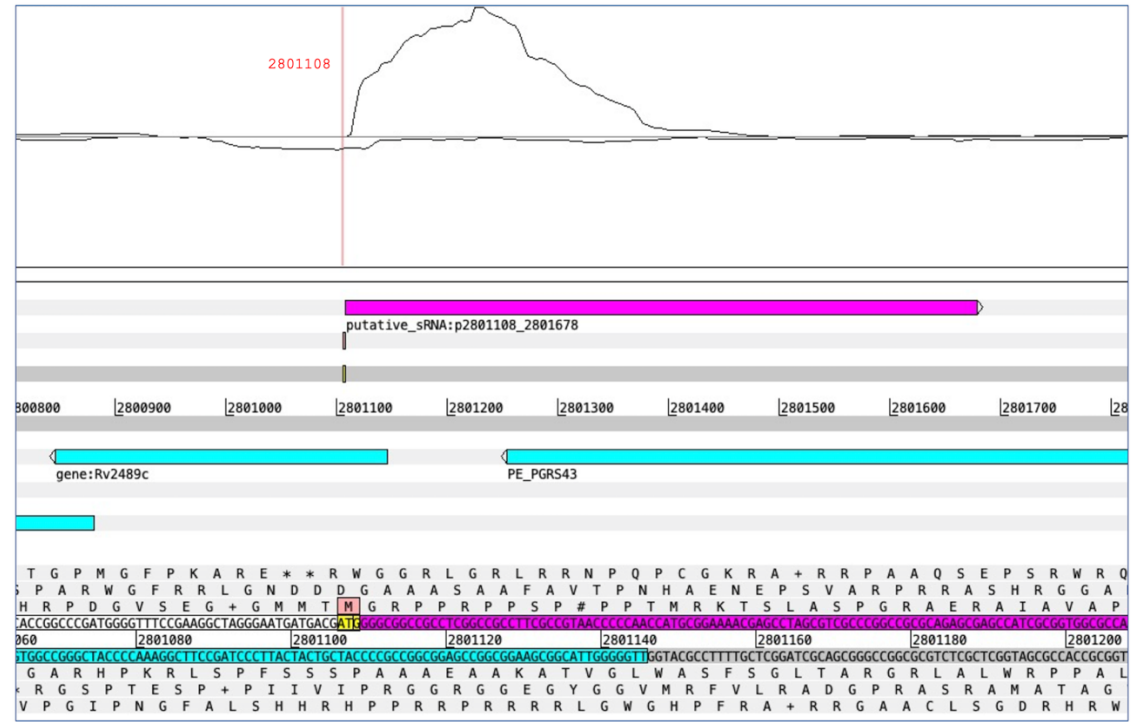


Figure 2.7. Antisense sRNA, ncRv2489/putative_srna:p2801108_2801678, (magenta bar) overlaps two transcripts and may encode a short peptide. TSS for sRNA indicated in red and corresponding amino acid highlighted in pink. Sample SRR5689230 from PRJNA390669, exponential growth on cholesterol and fatty acid media. Strand coverage using the 'second' read of each pair mapping to the transcript strand, visualised using Artemis genome browser (Carver et al., 2012).

and may involve RNAP collision if both are transcribed simultaneously. Therefore, transcription of the predicted sRNA could impact either Rv2489c and/or PE-PGRS43 expression through two different mechanisms. Another hub sRNA in ‘saddlebrown’ includes ncRv1450/putative_sRNA:p1630466_1631246, which has a TSS at 1630466 and is likely to be an intergenic transcript between two divergently transcribed genes on the opposite strand, *tkl* (Rv1449c) and PE-PGRS27 (Rv1450c), both of which are assigned to different modules. The 3’ end of the prediction includes possible run-on transcription antisense to the 3’ end of PE-PGRS27.

The fatty-acid desaturase gene, Rv3229c (*desA3*) is a hub in the module, but without its operon partner, Rv3230c. However, the module does contain an antisense sRNA in this region, ncRv3230/putative_sRNA:p3607084_3607499 which is antisense to the 3’ end of Rv3230c, but lacks a known TSS. Interestingly, Rv3230c has an internal transcription termination site predicted at 3607550 which coincides with the 3’ end of the antisense sRNA (D’Halluin et al., 2023) (Figure 2.8). Another hub antisense sRNA, putative_sRNA:p3608313_3608866/ncRv3231c, overlaps the 3’ end of the upstream gene, Rv3231c, and has a predicted TSS at 3608313.

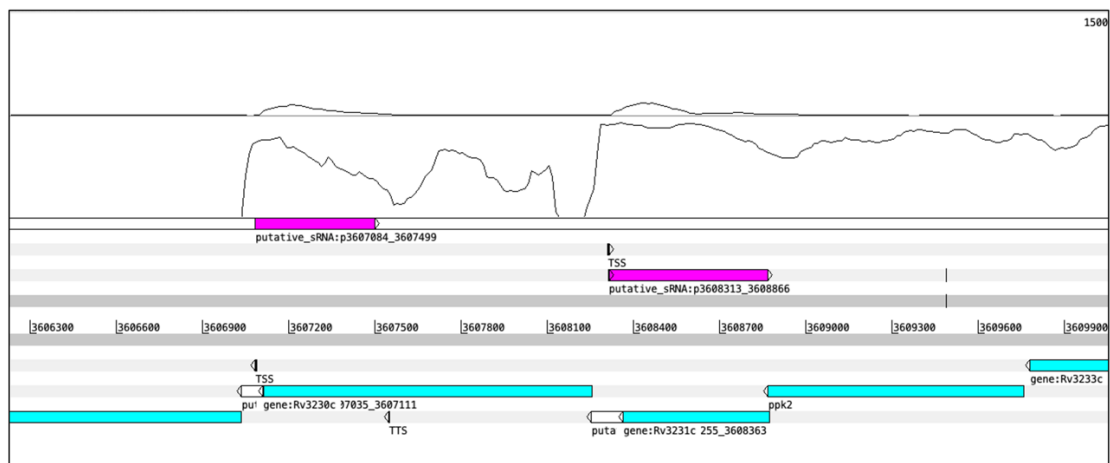


Figure 2.8. Antisense sRNAs (magenta bars) overlap Rv3230c and Rv3231c. TSSs and TTSs are indicated in black. Shown is sample SRR1917713 from PRJNA278760, exponential growth in dextrose media. Coverage is limited to 1500 reads to aid visualisation of coverage on the + strand. White bars are predicted UTRs. Strand coverage using the ‘second’ read of each pair mapping to the transcript strand, visualised using Artemis genome browser (Carver et al., 2012).

2.5.3.3 *Slow-growth correlated module is associated with transcriptional remodelling and metal ion homeostasis and enriched for sRNAs*

The '*green*' module contains genes that are associated with transcriptional remodelling in response to hypoxic or stationary growth conditions. It is positively correlated with hypoxic (bicor=0.49, p_{adj} =0.004) and stationary (bicor=0.4, p_{adj} =0.01) growth conditions, negatively correlated with exponential growth (bicor=-0.44, p_{adj} =0.01) (Figure 2.3) and is enriched for GO terms related to response to metal ions as well as regulation of gene expression. The '*green*' module contains at least 30 known transcription factors, with 14 of them hubs in the module, including FurA, Zur and sigma factor, SigH, as well as being enriched for SigH regulon genes. Three of the most well-connected transcription factors (furA, smtB and zur) are involved in iron uptake and utilisation, and the Zur-regulated ESAT-6 secretory proteins, esxR and esxS (Rv3019c, Rv3020c), are also present in the module, linking metal homeostasis with response to hypoxia (Maciag et al., 2007; Zhang et al., 2020). Two chaperonin protein targets of the non-coding RNA F6/Sfds, GroES (Rv3418c) and GroEL2 (Rv0440) are in the module, as well as the chaperonin protein, hsp (Rv0251c), all of which are part of the phoPR virulence-regulating system (Gonzalo-Asensio et al., 2008, 2014).

The '*green*' module is enriched for sRNAs (p_{adj} =0.011). Among the best-connected, are 27 predicted antisense RNAs. One of these hubs, putative_sRNA:p1404640_1404929/ ncRv1257 is antisense to the 3' end of Rv1257c, a probable oxidoreductase, and another (putative_sRNA:p1771044_1771498/ ncRv1546) is antisense to the 5' end of a trehalose synthetase, *treX*. Both of these sRNAs have TSSs and are expressed differentially among the tested conditions. Control of reactive oxygen species and synthesis of trehalose intermediates are important for cells in order to survive hypoxic conditions (Eoh et al., 2017; Harold et al., 2019) and antisense RNA may be involved in fine-tuning these responses. Another antisense RNA, ncRv1358c (putative_sRNA:m1530046_1530745) has a TSS near its start and is found antisense to Rv1359. Rv1359 and the upstream gene, Rv1358, on the opposite strand are very similar to each other (43.7% identity in 197 aa overlap) and to another gene elsewhere in the genome, Rv0891c (48.5% identity in 204 aa overlap) (Kapopoulou et al., 2011). All three genes are possible LuxR family transcriptional regulators

which are thought to be involved in quorum-sensing adaptations and contain a probable ATP/GTP binding site motif (Chen & Xie, 2011; Modlin et al., 2021) and are found in different modules. Expression of this antisense sRNA appears to suppress the expression of the transcript on the opposite strand to varying degrees in all conditions (Figure 2.9). In the cholesterol and fatty acid media samples, expression of a shorter transcript appears to begin inside the Rv1359 ORF, where the transcript is not overlapped by the antisense transcript, possibly utilising an internal TSS at 1530774.

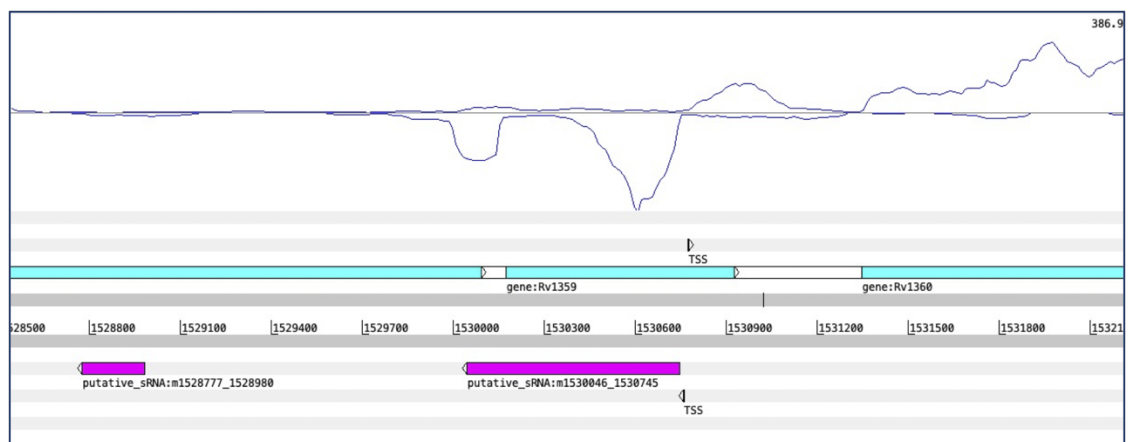


Figure 2.9. Expression of antisense transcript *putative_sRNA:m1530046_1530745* (magenta bar) seems to suppress the expression of most of Rv1359 and Rv1358 in cholesterol and fatty acid media. An internal TSS exists inside the Rv1359 CDS at 1530774 near where expression begins. Note, prediction of an individual sRNA is an aggregate of predictions under different conditions, so will not always match the expression of the sRNA in any particular sample. Sample SRR5689230 from PRJNA27860. Strand coverage using the 'second' read of each pair mapping to the transcript strand, visualised using Artemis genome browser (Carver et al., 2012).

2.5.3.4 Metal ion homeostasis genes cluster in module that is negatively correlated with the hypoxia condition.

The 'darkred' module is negatively correlated with the hypoxia condition (bicor=-0.46, $p_{adj}=0.005$, Figure 2.3). This module contains most of the ESX-3 genes (Rv0282-Rv0292) related to siderophore-mediated iron (and zinc) uptake in *M. tuberculosis* (Serafini et al., 2013; Zhang et al., 2020), with nine of these representing hubs in the module. The module is enriched for the PE/PPE functional category, and includes the two genes preceding the ESX-3 genes, Rv0280 (PPE3) and Rv0281 (a possible S-adenosylmethionine-dependent methyltransferase involved in lipid metabolism, though its position in the genome would suggest regulation could be linked to ESX-3 (Lunge et al., 2020)), and an ESX-5 gene, Rv1797 (*eccE5*). The module also contains

another Zur-regulated gene, Rv0106, which is a potential zinc-ion transporter (Zondervan et al., 2018). Among the hubs of the module are several genes related to lipid metabolism and fatty acid synthesis, including: probable triglyceride transporter, Rv1410; the operon consisting of Rv0241c (*htdX*), Rv0242c (*fabG4*), and Rv0243 (*fadA2*) (Dutta, 2018); and a gene involved in the pentose phosphate pathway, *zwf2* (Rv1447c).

There are some well-connected ncRNAs in the '*darkred*' module, including a predicted antisense RNA to Rv0281, 'ncRv0281c' (putative_sRNA:m341328_342075). This putative sRNA has a predicted TSS at the 5' end and is transcribed divergently from Rv0282 (*eccA3*). This is one of the rarer cases where the antisense transcript and cognate protein-coding gene (Rv0281) are clustered in the same module. The prevailing direction of transcription at this locus may be a result of competition for RNAP binding at a bi-directional promoter in the predicted 5' UTR of Rv0282 which also clusters in the module. There are several UTRs in the module hubs, including a 3' UTR for the gene Rv1133c, *metE* (also found in the module). This UTR was previously identified as abundantly expressed in exponential culture (Arnvig et al., 2011). There is a 3' UTR for Rv0292 (*eccE3*, also a hub in the '*darkred*' module) that is antisense to a large part of the 3' end of Rv0293c which has a converging orientation to Rv0292 (Figure 2.10). Rv0293c is a hub in a different module ('*lightgreen*') along with its 3' UTR. Overlapping 3' ends of genes could function to regulate transcription, possibly by bi-directional termination brought about by RNAP collision, or function post-transcriptionally by influencing transcript stability (Ju et al., 2019; Vargas-Blanco & Shell, 2020).

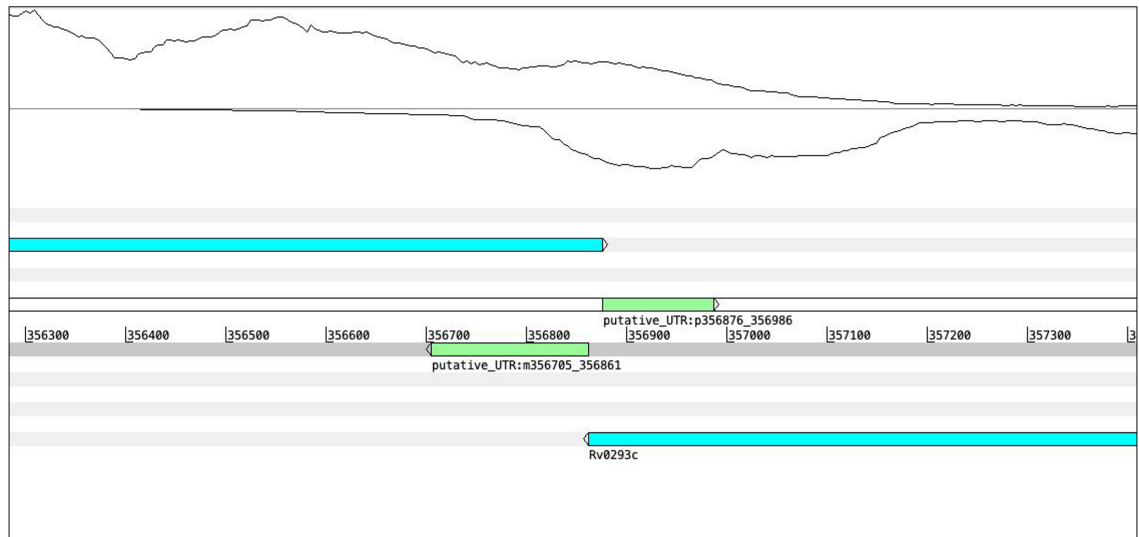


Figure 2.10. Overlapping 3' UTRs for Rv0292 (EccE3) and Rv0293c, (light green bars) may regulated transcription termination or transcript stability. Sample SRR5689224 from PRJNA390669, exponential growth in dextrose-containing media. Strand coverage using the 'second' read of each pair mapping to the transcript strand, visualised using Artemis genome browser (Carver et al, 2012).

2.5.3.5 Module enriched for sRNAs and PE/PPE genes is correlated with stationary condition

The 'darkturquoise' module is enriched with sRNAs, with 33 hub sRNAs. It is negatively correlated with the low iron condition (bicor = -0.37, $p_{adj}=0.03$) and positively correlated with stationary growth (bicor= 0.43, $p_{adj}=0.007$). The genes of the module are enriched for the PE/PPE functional category and there are several PE/PPE genes among the hubs. The previously annotated ncRNA, B11 (also known as 6C or ncRv13660c), is one of the most well-connected elements in the module and overexpression of B11 in *M.smegmatis* has been shown to cause growth arrest and downregulation of a large set of genes including those involved in cell division and virulence, including all the ESX-1 secretion system genes (Mai et al., 2019). Mcr11 is also found in the module. This sRNA is known to respond to the second messenger 3',5'-cyclic adenosine monophosphate and has been found to be expressed in hypoxic *M. tuberculosis* cultures and in a mouse infection model (Girardin & McDonough, 2020). Mcr11 regulates the expression of several genes that adapt central carbon metabolism during slow growth conditions (Girardin & McDonough, 2020).

There are two well-connected intergenic sRNAs predicted in the 'darkturquoise' module. Putative_sRNA:p1164036_1164162 / ncRv11040 is located between PE8

and a possible transposase, Rv1041c, but on the antisense strand. There is a predicted TSS at 1163697, 39 nucleotides upstream of the predicted start sequence. This transcript is in a converging orientation to the transposase and may be instrumental in regulating horizontal gene transfer (Ellis & Haniford, 2016; Lejars et al., 2019). The other intergenic hub is also upstream from possible transposase, Rv3114, but in diverging orientation on the opposite strand. The TSS is at 3481459, and the sRNA is within a predicted 'MT-complex-specific' genomic island associated with virulence genes (Becq et al., 2007). Rv3112-14 are clustered in the '*salmon*' module.

There are several interesting 'independent' UTRs that are well-connected in the module, but their assumed transcriptional partner clusters in another module. There are several predicted TSS's and transcriptional termination sites (TTS) (D'Halluin et al., 2023) within the predicted boundaries of a 3' UTR for the gene Rv2081c (putative_UTR:m2337218_2338064) and a predicted sORF based on ribosome profiling (Smith et al., 2022) (Figure 2.11). The adjacent gene, Rv2081c, is in the '*cyan*' module along with most of the DosR regulated genes. The 5' UTR of Rv0281c is also a hub in the module and contains predicted TSSs, TTS and sORF. It would be interesting to discover whether these UTRs could have dual functions as regulatory RNA elements as well as being translated into short peptides. Rv2081c is a conserved membrane protein containing a simple sequence repeat of 8 C's and has been identified as a source of sequence variation in *M. tuberculosis* sputum and culture (Shockey et al., 2019; Sreenu et al., 2007).

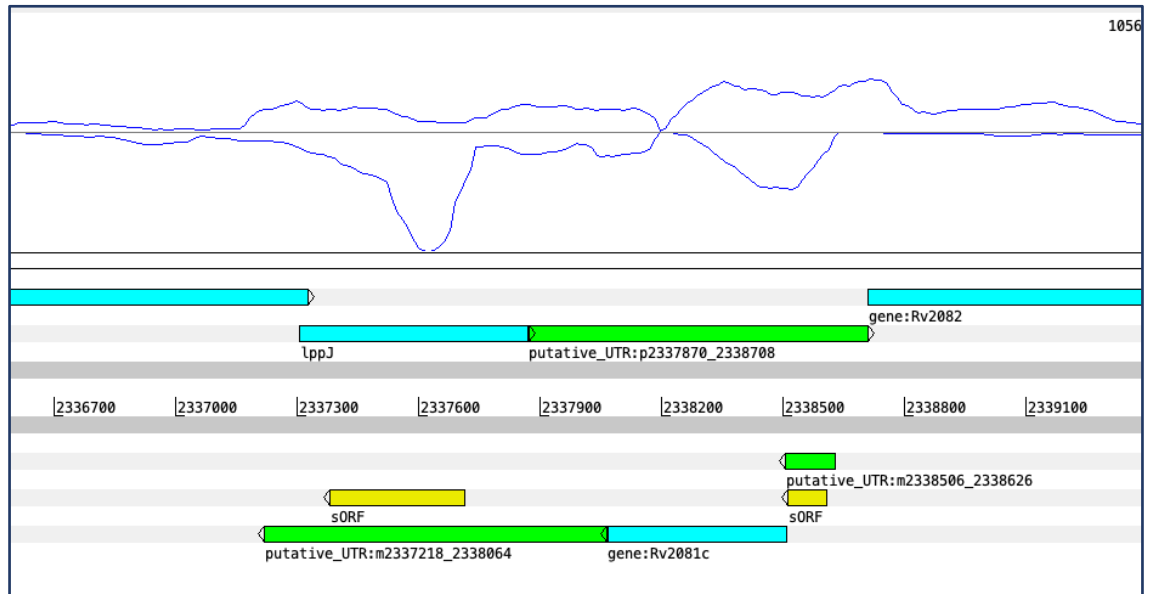


Figure 2.11. 5' and 3' UTRs for Rv2081c (green bars) are overlapped by predicted sORFs (yellow bars). (Cortes et al., 2013; Smith et al., 2022). Shown is sample SRR5689224, exponential growth, from PRJNA27860. Strand coverage using the 'second' read of each pair mapping to the transcript strand, visualised using Artemis genome browser (Carver et al., 2012).

The best connected elements in the module are antisense sRNAs, including putative_sRNA:p2081553_2082178/ ncRv1835, with a predicted TSS at its start. It is antisense to Rv1835c, the gene for a putative serine esterase clustered in the 'mediumpurple3' module, in particular to the 3' end of the peptidase domain (Xaa-Pro dipeptidyl-peptidase-like domain) (Blum et al., 2020). Putative_sRNA:m2497549_2498369 /ncRv2225c, with a TSS at 2498368, is antisense to Rv2225, coding for a 3-methyl-2-oxobutanoate hydroxymethyltransferase PanB. This gene clusters in the 'turquoise' module.

2.5.4 Comparison with other global *M. tuberculosis* networks

Other regulatory networks have been developed for *M. tuberculosis* that use transcriptomic data to cluster protein-coding genes according to their responses to environmental conditions (Peterson et al., 2014; Yoo et al., 2022). Peterson et al. (Peterson et al., 2014), utilises a 'biclustering' algorithm, *cMonkey*, that clusters genes and conditions based on co-expression in publicly available microarray data and the presence of common transcription factor binding motifs (Reiss et al., 2006). The network is pruned and shaped by adjusting the weights of particular lines of evidence *a priori* input such as binding motifs, protein homology, operon groupings and known protein-protein interactions (PPIs) (Peterson et al., 2014; Reiss et al.,

2006). This network's ability to assimilate both *a priori* and transcriptomic expression data was tested by its ability to recapitulate known associations and groupings found by overexpression of transcription factors and identification of transcription factor binding motifs. Thus, a 'parsimonious' network was created that uncovers novel transcriptomic responses to particular environmental conditions that are validated by several lines of evidence (Peterson et al., 2014; Reiss et al., 2006). This approach differs significantly from ours in several important ways. Firstly, the WGCNA network we present relies entirely on transcriptomic data alone—RNA-seq, in particular. RNA-seq is more sensitive than microarray data and is able to detect the expression of novel transcripts that may represent non-coding or unknown protein-coding RNA transcripts. Our network is more comprehensive in an attempt to include every detectable RNA transcript found in the included RNA-seq datasets. These novel transcripts naturally lack any *a priori* data to shape or reinforce associations, and we have not applied any filtering methods other than evaluating the strength of module membership.

A more recent approach uses a large number of RNA-seq datasets with deconvolution methods to reduce the noise in the network and find clusters of protein-coding genes ('iModulons') that together account for significant chunks of variation in expression levels in response to environmental conditions (Yoo et al., 2022). In both the Yoo et al and Peterson et al studies, genes can be members of more than one module, unlike our WGCNA network where all transcripts are assigned only to a single module, making any direct comparison of the entire network of limited value. However, several of the modules highlighted in the previous studies do show considerable overlap with the protein-coding members of some of the modules presented here, especially in modules associated with response to hypoxia or cholesterol media. For example, a comparison of the hypoxia-linked, 'DevR', iModulon and the protein-coding genes of the '*cyan*' module with a MM cutoff of 0.7, reveals 34 overlapping genes between them. All 13 of the hypothetical proteins in the '*cyan*' hubs are also in the 'DevR' iModulon. The hypoxia-linked 'Bicluster 182' shares 7 genes with both the iModulon and the '*cyan*' module (Figure 2.12A). The *kstR* regulon-enriched module, '*royalblue*', discussed earlier, shares 15 hub genes with the Rv0681 iModulon and 18 genes with the group

of three biclusters identified in the Peterson et al study as enriched for steroid ring degradation (Biclusters 199, 200 and 337) (Figure 2.12B).

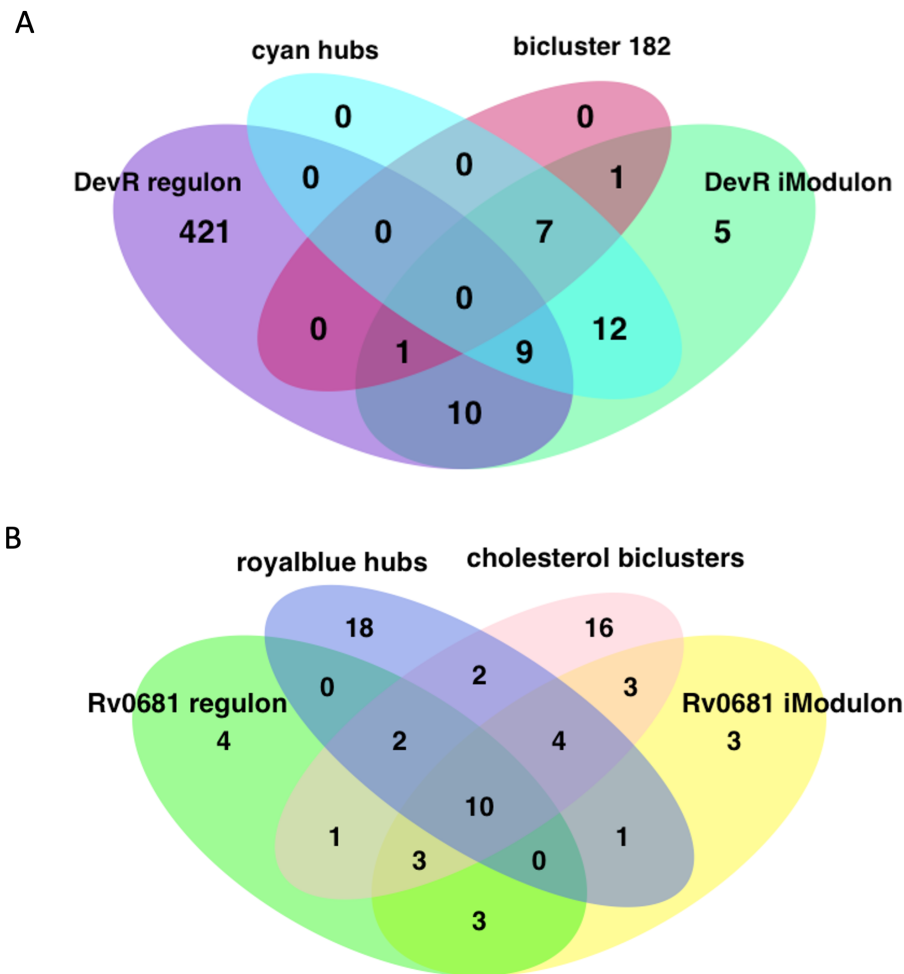


Figure 2.12. Protein coding genes involved in responses to hypoxia and adaptation to cholesterol cluster together in overlapping modules in different network approaches. A) Comparison of protein-coding genes with $MM > 0.7$ in 'cyan' module with Bicluster 182 (Peterson et al, 2014), DevR iModulon (Yoo et al, 2022) and DevR regulon. B) Comparison of cholesterol metabolism biclusters linked to steroid ring degradation (bc_0199, bc_0200, bc_337) (Peterson et al, 2014), Rv0681 iModulon (Yoo et al, 2022) and the protein-coding genes of 'royalblue' module with $MM > 0.8$. Regulons were defined as in Yoo et al, 2022 (downloaded from https://github.com/Reosu/modulome_mtb) and include genes with predicted binding.

As all the RNA-seq datasets included in this WGCNA analysis are also included in the iModulon analysis, overlaps between these two studies are perhaps not surprising. An important distinction between our study and these other approaches is that the network presented here seeks to identify not just groupings of protein-coding genes linked by transcriptional regulation, but associations involving non-coding RNA, as well. For example, the protein-coding hub genes of the 'violet' module overlap with the 'VirS' iModulon which was linked in Yoo et al (2022) to response to acid

environment and remodelling of cell membrane. In addition to the coding genes that overlap the 'VirS' iModulon, the hubs of the '*violet*' module include the non-coding RNA, Mcr7. Mcr7 is a ncRNA known to be activated by the PhoPR regulon which responds to acid pH (Solans et al., 2014). The hypothetical protein-coding transcript that overlaps this locus, Rv2395A, is found in the 'PhoP' iModulon. The '*violet*' module also includes several UTRs among the hub members that may represent important players in this adaptation response. Thus, our approach adds value to these previous methods by including unannotated elements that may have roles in the regulation of gene expression.

One advantage of the deconvolution method over WGCNA is that by filtering for only the strongest associations and allowing genes to be members of more than one iModulon, the modules are less 'noisy'. However, deconvolution methods require extremely large numbers of samples to perform well, may be subject to batch effect issues between experimental datasets and characterise a limited proportion of the protein-coding transcripts expressed by *M. tuberculosis* (Saelens et al., 2018; Yoo, et al., 2022). In order to include predicted ncRNA in the network, a significant degree of quality control, parameter adjustment and manual curation is required, limiting the number of datasets that could be included in our analysis. Including more data would most likely strengthen the correlations with certain conditions and improve the overall specificity of the WGCNA modules.

The gene modules presented here are somewhat 'blunt-force instruments' applied to transcripts that are part of overlapping, coordinated responses to various environmental cues, but restricted to a single module grouping. Recent work exploring differentially expressed genes in response to various environmental conditions have revealed highly integrated adaptation responses. In other words, a single environmental change, e.g. hypoxia or growth on fatty acids or cholesterol, stimulates transcriptomic remodelling across diverse cellular functions, perhaps acting as cues to stimulate anticipatory pathways and ready the pathogen for the next challenge (Eoh et al., 2017; Gerrick et al., 2018). Confounders such as dual-function, 'moonlighting', proteins may weaken the correlation of a module with a specific condition and may create noise in otherwise well-connected modules. Rather than using an arbitrary cutoff to decide which module associations are

relevant, we utilise a flexible measure of module membership that allows the user to filter the strength of associations. In our discussion, we used a relatively stringent threshold 'module membership' score of 0.8 to identify the transcripts in each module that have the tightest correlation to the module eigengene, but there has been no pruning or editing of the modules, in order to avoid any loss of information.

An important advantage of including ncRNA in a co-expression network is the chance to observe post-transcriptional groupings that result from adaptive responses, as well as the transcriptional responses. By focussing on the best connected transcripts in various modules, unexpected connections between genes of diverse pathways can be discovered. The work presented here confirms that ncRNA are important players in adaptation responses, and the existence of informative protein-coding co-expression networks can help to implicate these transcripts in adaptive responses and provide context for their activity.

2.6 CONCLUSION

This paper presents a large-scale network analysis of over 7000 transcripts expressed by *M. tuberculosis* under a variety of conditions. The modules group together clusters of co-expressed protein-coding genes, as well as ncRNA transcripts predicted from RNA-Seq signals. Several modules are statistically enriched for sRNAs, especially those modules positively correlated with hypoxia. The abundance of antisense RNA in conditions of stress has been widely observed, and it is therefore not a surprise to find them in the hubs of these modules. However, it is noticeable that the complementary ORF is usually excluded, which leads us to seriously consider antisense transcription as part of strategic regulation of protein production in response to environmental cues through mechanisms of divergent transcription, translational control or by regulating mRNA stability (Vargas-Blanco & Shell, 2020; Warman et al., 2021). If these strategies actually differ among the members of the MTBC, it may have implications for host specificity and virulence (Dinan et al., 2014). By the same logic, 3' UTR transcripts clustering in modules distinct from their upstream ORF implies independent function from the ORF. sRNAs generated from 3' UTRs have been reported in other prokaryotes and evidence points to widespread mRNA processing that could release independent

transcripts at the 3' end (Dar & Sorek, 2018; Desgranges et al., 2021; Updegrave et al., 2019; Wang et al., 2019). In compact bacterial genomes, 3' UTRs are also found to overlap other 3' UTRs in a converging transcription pattern which may provide a mechanism for regulating the expression or stability of either transcript (Ju et al., 2019; Vargas-Blanco & Shell, 2020).

The modules discussed in depth in this paper represent a limited snapshot of this extensive co-expression network. Modules of interest can be identified by correlations to experimental conditions, associated GO terms, functional categories, or gene group enrichment. The supplementary tables (A2.1 Supplemental Tables: Ch2_Supp_Table_4) have been organised into an easily-accessible spreadsheets for researchers to query particular genes or modules of interest and find associated protein-coding genes or ncRNA. These spreadsheets provide information about the module association, membership values, TSSs and for UTRs, the module membership of the adjacent ORFs for each predicted ncRNA. To facilitate further exploration of this extensive data, we have made a simple R Shiny app available at https://github.com/jenjane118/mtb_wgcna. Modules can be explored for hub members and individual transcripts can be queried for expression profiles and adjacent non-coding RNA. We anticipate this to be a useful resource for discovering ncRNA candidates for further investigation, add context to the circumstances of expression of previously identified ncRNAs, identify associations of genes with unknown functions and suggest roles for 'moonlighting' proteins that may be associated with unexpected gene groupings.

Chapter 3: Using transposon-insertion sequencing to identify the different essential gene requirements in vitro between human-adapted and animal-adapted members of the MTBC

*The data used in this chapter, and some analysis, was previously published in the following article. The chapter includes the original published TRANSIT analysis and additional complementary analysis. All of the text is original to this chapter.

Gibson, A. J., Passmore, I. J., Faulkner, V., Xia, D., Nobeli, I., Stiens, J., Willcocks, S., Clark, T. G., Sobkowiak, B., Werling, D., Villarreal-Ramos, B., Wren, B. W., & Kendall, S. L., 2021, "Probing Differences in Gene Essentiality Between the Human and Animal Adapted Lineages of the *Mycobacterium tuberculosis* Complex Using TnSeq", *Frontiers in Veterinary Science*, 8(December), 1-12 (Gibson et al., 2021).

3.1 ABSTRACT

The host-adapted species of the *Mycobacterium tuberculosis* complex share high degrees of sequence similarity but differ in pathology, virulence and host preference. Adapting to different host environments and immune systems causes different requirements for orthologous genes among the animal-adapted and human-adapted lineages. In this study, the essential gene requirements of *Mycobacterium bovis* and *Mycobacterium tuberculosis* are compared using parallel transposon insertion sequencing experiments in identical culture conditions. Libraries with similar insertion density levels were created for each species (55% for *M. bovis*, 40% for *M. tuberculosis*) and the essentiality of orthologous genes were compared using orthogonal statistical and quantitative analyses. Comparing essentiality predictions from independent analyses of each library using an HMM model, 363 (10%) of orthologous gene pairs were essential and 492 (13.7%) had some level of fitness defect in both species. Using a quantitative statistical comparison of the differences in the insertion frequency between orthologous gene pairs, 32 genes had statistically significant differences in mean insertions. Non-coding RNA predictions and annotations from *M. tuberculosis* RNA-seq data were also tested for differences in essentiality and 15 transcripts had different

essentiality predictions with the HMM analysis, however, none of these transcripts showed a statistically significant difference in mean insertions. This study provides a resource for mycobacterial researchers interested in characterising genes and non-coding RNA that may be involved in host adaptation.

3.2 AIMS

- Evaluate transposon insertion sequencing libraries from parallel cultures of *M. tuberculosis* and *M. bovis* grown in identical culture conditions to determine gene requirements
- Apply an additional quantitative 'resampling' analysis complementary to the qualitative analysis used in the published work
- Identify predicted non-coding RNAs required for survival in the *M. tuberculosis* or *M. bovis* genomes

3.3 INTRODUCTION

3.3.1 Host-adapted species may have different gene requirements

Determining which bacterial genes are essential for survival can inform researchers about the most important gene targets for future therapies. Zoonotic tuberculosis by *M. bovis* is an under-recognised human health problem, requiring different strategies for treatment and prevention than for human-adapted *M. tuberculosis* (Olea-Popelka et al., 2017). Members of the MTBC, including *M. bovis* and *M. tuberculosis*, are likely to have different gene requirements in order to meet the diverse challenges presented by different host immune systems. However, it is difficult to predict which genes are more, or less, required for survival among the closely related strains of the MTBC based on transcriptional differences. The correlation between gene expression and gene essentiality is low, as changes in expression under different conditions may not reveal the essential role of constitutively active genes, or gene regions (Carey et al., 2018; Griffin et al., 2011; Rengarajan et al., 2005). Genes can code for both essential and non-essential protein domains, and even non-coding regions of the genome may be essential.

3.3.2 Determining gene essentiality with transposon insertion sequencing

Transposon insertion sequencing (tn-seq) is a next-generation sequencing technique that can be used to identify required, or 'essential' regions of a genome for survival of a bacteria in a particular environment. Tn-seq has been used to characterise the essentiality of genes in *M. tuberculosis* cultured *in vitro*, using the reference strain, H37Rv, (Dejesus et al., 2017; Griffin et al., 2011; Minato et al., 2019; Patil et al., 2021; Sassetti & Rubin, 2003; Zhang et al., 2012), the attenuated vaccine strain *Mycobacterium bovis* BCG (Mendum et al., 2019), and in various other mycobacterial species (Budell et al., 2020; Lefrançois et al., 2024; Majumdar et al., 2017; Tateishi et al., 2020); but only one previous study has applied tn-seq to *M. bovis* (Butler et al., 2020). Experimental challenges and differences in culture media have resulted in significant variation among reports, which confounds simple comparisons between experiments. For *M. tuberculosis* experiments, many of these studies are compiled and re-standardised in an interactive database, https://www.mtbtndb.app/analyze_datasets (Jinich et al., 2021) which somewhat improves the situation for *M. tuberculosis*, but there exists a need to compare gene requirements between different mycobacterial species and strains in order to evaluate the applicability of vaccine and treatment regimens. For example, a quantitative approach using tn-seq libraries of 9 clinical isolates of *M. tuberculosis* identified significant differences in gene requirements between these strains and the reference strain for *M. tuberculosis* (H37Rv), leading to the identification of differences in antibiotic susceptibility among the clinical strains (Carey et al., 2018).

The tn-seq protocol used in this study begins with transducing a bacterial population with a phage (MycomarT7) that inserts a transposon (*Himar1* mariner) at regular positions throughout the target genome (at 'TA' dinucleotides, specifically) to create transposon libraries of mutants each with a single insertion (Griffin et al., 2011; van Opijnen et al., 2009; Zhang et al., 2012). These libraries, ideally, would include individual mutants with insertions at every possible insertion motif in the genome (and in multiple positions in each gene), though it is recognised that some insertion sites are relatively non-permissive for insertions (Dejesus et al., 2017). Insertions interrupt normal transcription and gene function depending on their location within the gene, with insertions inside the open reading frame less likely to be tolerated than insertions closer to the 5' and 3' ends (Griffin et al., 2011;

Hutchison et al., 1999; Zhang et al., 2012). After growth in culture, or in an organ or animal model, next generation sequencing of the pooled genomic DNA from the libraries allows the location of the insertions to be determined and quantified (Figure 3.1). It has been demonstrated that insertion counts are a reliable proxy for the number of unique mutants in the population (Griffin et al., 2011). The analysis is often applied using gene boundaries, and genes containing the expected number of mapped reads at the possible insertion sites based on the overall insertion density, are inferred to be 'non-essential' for the fitness of the mutant--indicating that inactivation of the gene does not result in a significant loss of fitness. If genes have fewer insertions than expected, it can be inferred that inactivation of these genes leads to a significant loss of fitness in these mutants and thus underrepresentation in the cultured library. Various statistical approaches are used to clarify whether a gene is 'essential', i.e. inactivation is fatal or causes a severe growth disadvantage, versus those genes that, when inactivated, cause a milder growth defect. The location of the insertion also impacts the severity of the growth defect as some proteins include both essential and less-essential/dispensable domains (Dejesus et al., 2017; Patil et al., 2021). However, inactivation of a gene with a redundant, but essential, function may result in a mild growth defect, or no phenotype at all, in which case it will be categorised as 'non-essential'. Less commonly, insertions that inactivate the gene can cause a growth *advantage* to the mutant--resulting in its overrepresentation in the mutant pool.

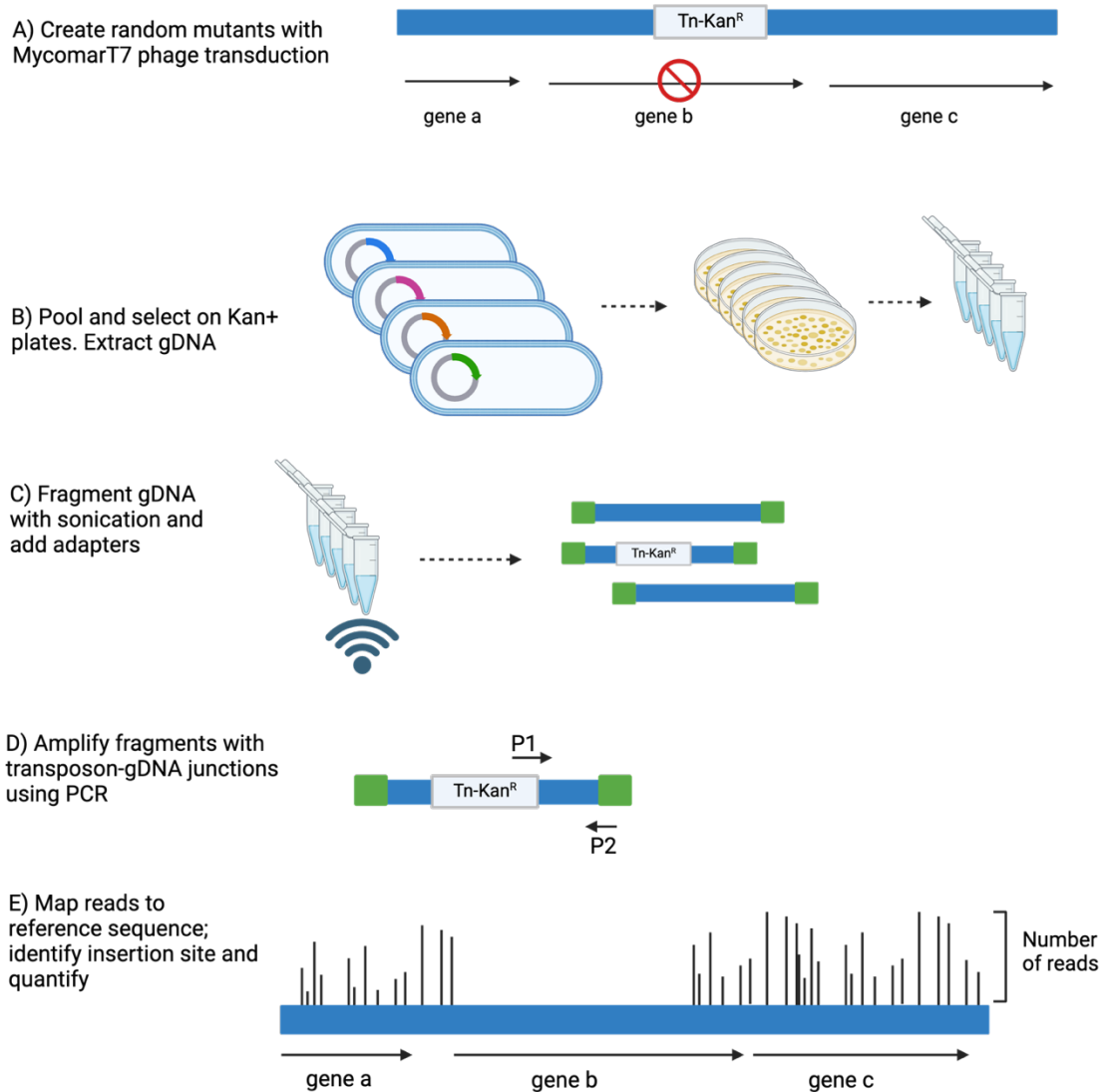


Figure 3.1. Overview of transposon insertion sequencing (tn-seq) experiments. A) Library of single insertion mutants created with transposon insertions (Tn-Kan^R) at 'TA' dinucleotides throughout the genome. B) Transduced bacteria were selected for on kanamycin-containing plates and genomic DNA extracted. C) Genomic DNA was fragmented using sonication and adapters (green boxes) ligated to gDNA fragments. D) PCR using primers (P1, P2) complementary to transposon sequence and adapter amplifies only transposon-containing fragments. E) Sequenced reads are mapped to reference genome and quantified; each read is a proxy for one insertion event. Figure made with Biorender.com

3.3.3 Statistical analysis of transposon insertion sequencing results

There are several different statistical models used to predict the essentiality status of a gene or gene region from the reads mapped to insertion sites (Cain et al., 2020; Chao et al., 2016; Long et al., 2015). Annotation-dependent methods include Bio-Tradis (Langridge et al., 2009), which uses an empirical method to compare the typically bimodal frequency distribution of normalised insertion indices for each gene with a likelihood ratio to determine whether or not a gene falls within the

'essential' or 'non-essential' peaks. This method is based on the use of the Tn5 transposon (TraDIS method) which inserts at random places in the genome, versus the *Himar1* transposon, used in most tn-seq protocols, which inserts at 'TA' sites, meaning the Bio-Tradis pipeline (Barquist et al., 2016) is not applicable to tn-seq/*Himar1* analysis without modification. Bayesian methods have also been used to calculate the posterior probability of a gene being essential. The saturation and potential number of insertion sites are used to predict the largest possible window of nucleotides that could have no insertions based on chance alone (DeJesus et al., 2013). If a gene contains a region with no insertions that is larger than this predicted size, it is determined to be essential. This method is easier to apply in tn-seq/*Himar1* transposon experiments where the number of possible insertion sites can be determined definitively.

Annotation-independent methods predict regions of essentiality across the genome, without being restricted to protein-coding gene annotations. These include 'sliding-window' methods that test small, overlapping windows of the genome to determine if insertions are underrepresented in the window compared to all other windows using a non-parametric test (Chao et al., 2016; Zhang et al., 2012). This also assumes the uniform insertion probability offered by the *Himar1* transposon to estimate library saturation, but its power is limited by a large multiple testing penalty (Chao et al., 2016). One of the most frequently utilised programs applies a Hidden Markov Model (HMM) to convert the number of reads mapping to insertion sites to a series of probabilities of moving from an 'essential' to a 'non-essential' state. Individual sites have a probability of being in a particular state depending on the state of the sites before and after them. The parameters of the HMM prediction algorithm were tested on datasets with various levels of saturation and thus the probability assignments are adjusted in response to different levels of saturation. This site-by-site analysis is more fine-grained than a gene-based approach but can be compounded to determine the essentiality status of a gene (or protein domain or any other defined region) containing multiple sites. Furthermore, the model incorporates two intermediate states, 'growth defect' (GD), which reflects a probability state where the number of reads at an insertion site is more likely to represent a mild negative impact on growth relative to wild-type and 'growth

advantage' (GA) where insertions at a specific 'TA' site are actually beneficial to survival of the mutant (DeJesus & Ioerger, 2013).

Due to technical constraints such as experimental bottlenecks, no single tn-seq library is likely to have every possible insertion site occupied by a transposon. However, the assumption of independence between independent tn-seq libraries (as each library will have an independent selection of which insertions sites are occupied by a transposon) means that individual libraries created under the same experimental conditions can be pooled to increase the number of unique insertion sites in an experiment (also known as 'insertion density' or 'saturation'). Increasing the insertion density before statistical analysis is important to reduce false positives resulting from stochastic processes (DeJesus et al., 2017; DeJesus & Ioerger, 2013; Mahmutovic et al., 2020; Patil et al., 2021).

In this chapter, data from parallel whole-genome transposon insertion sequencing experiments in the reference strains for *M. bovis* and *M. tuberculosis* was used to identify different requirements for survival in identical rich media culture conditions. As different culture conditions can impact the variety of mutants recovered from culture (Griffin et al., 2011), for purposes of comparison it is important to grow the libraries in controlled conditions rather than compare a new *M. bovis* tn-seq library with published *M. tuberculosis* results from diverse laboratories. To this end, *M. bovis* and *M. tuberculosis* transposon libraries were created to identify and compare genes essential for survival in identical *in vitro* culture conditions. I consider host-specific differences in requirements for the protein-coding genes between the species both by comparing the results of independent HMM analyses (published in Gibson et al, 2021) and a complementary, quantitative analysis, previously used by Carey et al to compare *M. tuberculosis* clinical strains to the reference strain, H37Rv (Carey et al., 2018). Expression of non-coding RNA shows species-specific differences between *M. bovis* and *M. tuberculosis* (Dinan, et al., 2014; Golby et al., 2013) and it is possible these differences can have a role in host-specific adaptation. Therefore, I have also evaluated the set of predicted, intergenic, non-coding RNA elements expressed in *M. tuberculosis* from Chapter 2 (Stiens et al., 2023) to determine whether these are essential in either species.

3.4 MATERIALS AND METHODS

3.4.1 Creation of Libraries

Transposon insertion libraries were created for *M. bovis* by transduction with the mariner-transposon containing MycoMarT7 phage, selection on kanamycin plates and genomic DNA extraction using bead-beating and enzymatic lysis as per (Gibson et al., 2021) by members of the Kendall lab at the Royal Veterinary College. *M. bovis* and *M. tuberculosis* libraries were grown on Middlebrook 7H11 solid medium containing 0.5% lysed defibrinated sheep blood, 10% heat inactivated foetal bovine serum, 10% OADC, 25µg/ml kanamycin and 0.05% Tween[®]80. Genomic DNA was extracted from three subsamples of the *M. bovis* library and two technical replicates of the *M. tuberculosis* libraries. Sequencing libraries for Illumina sequencing of gDNA enriched for transposon insertions were created by Ian Passmore in the lab of Brendan Wren at the London School of Hygiene and Tropical Medicine (Gibson et al., 2021) and sequenced on the Illumina HiSeq 3000 platform with 20% PhiX spike-in to increase the heterogeneity of signal in the early sequencing rounds. Paired-end fastq files of 150 base-pair length were generated and used by the author for all the subsequent data analysis steps included in this chapter.

3.4.2 Processing sequencing reads

Raw fastq reads were assessed for read quality using bash scripts and fastQC (Andrews, 2010) and pre-processed using the *TPP* utility from the TRANSIT package in single-end mode (DeJesus et al., 2015). Only the first read from each pair (containing the transposon-gDNA junction) was used for analysis, as the second read will not always include the transposon sequence and is not necessary for mapping. The first step identifies reads containing a transposon insertion by searching the read for the terminal sequence of the transposon (ending in '-GTTA') and trimming this transposon 'tag'. *BWA-mem* (Li, 2013) is then used to map the subsequent gDNA suffix to the appropriate genome: *M. tuberculosis* (H37Rv, NCBI Accession Number AL123456.3) or *M. bovis* (AF2122/97, NCBI Accession Number LT708304.1). The *TPP* program uses annotation tables in a specific format ('prot-tables') in order to map 'TA' insertion sites to gene regions. These were created for each genome using the appropriate genome annotations (gff file) with scripts

written in R. The resulting insertion files (in wig format) were analysed for insertion density, skew and indication of PCR jackpots or hotspots using TRANSIT *tnseq-stats* functions and custom R scripts.

3.4.3 Data Analysis

The HMM algorithm from the TRANSIT package was applied separately to the *M. bovis* and *M. tuberculosis* insertion files to identify the essentiality status of each coding gene. *TRANSIT-HMM* was run using a sum of reads per insertion site and TTR ("trimmed total reads") normalisation parameters. TTR normalisation trims the highest and lowest 5% of read counts before normalising to the mean read count. This method was used as it is most flexible with less saturated datasets.

Bio-Tradis (Barquist et al., 2016; Langridge et al., 2009) processing and analysis was also undertaken for purposes of comparison. The transposon tags were filtered and removed from quality-checked fastq files using *Bio-Tradis* scripts (<https://sanger-pathogens.github.io/Bio-Tradis>) and the reads were mapped to the respective genomes (as above) using *BWA-mem*. Reads were assigned to 'TA' sites using *Bio-Tradis* commands. However, when used with *Himar1* mariner transposon libraries, the scripts resulted in inaccurate insertion coordinates for reads on the lagging strand, and therefore artificially inflated the number of unique insertions for each sample.

For the TRANSIT *resampling* analysis, a 'bovis_on_tb' prot-table was made for the *M. bovis* libraries, mapping the coordinates of the orthologous *M. tuberculosis* genes to the *M. bovis* genome using custom R scripts. A null distribution was created by performing the permutation 100,000 times and p-values were assigned based on the difference between the observed distribution and the null. The returned 2-tail p-values were corrected for multiple testing using the Benjamini & Hochberg method (Benjamini & Hochberg, 1995). Resampling was run with the following parameters: TTR normalisation, 100000 permutations, winsorization and pseudocount = 5 to decrease the effect of individual sites with unusually high read counts. For non-coding RNA, separate prot-tables were created using the non-coding RNA *M. tuberculosis* coordinates and the corresponding genomic coordinates in *M. bovis*, and used with TRANSIT *resampling* (same parameters as above) to

evaluate log₂ fold-change of mean insertions within the genomic region of the ncRNA features.

3.4.4 Compiling set of orthologous genes

The essentiality calls were compared for orthologous genes in the *M. bovis* and *M. tuberculosis* genomes. Orthologous pairs of genes were identified by a previous study (Malone et al., 2018) based on amino-acid sequence of the protein products. As differences in the DNA sequence length could affect the number of 'TA' sites in the gene, and therefore the determination of essentiality, this is unsatisfactory for comparing essentiality between two orthologous genes with the HMM method. Therefore, reciprocal (or bidirectional) BLAST best hits between nucleotide sequences of coding genes in *M. bovis* and *M. tuberculosis* were inferred to be orthologs (Wolf & Koonin, 2012). A list of 'positionally orthologous' genes was also generated using the Progressive Mauve alignment tool (Darling et al., 2010) with corresponding genomic coordinates for protein coding genes in *M. bovis* and *M. tuberculosis* genomes which showed good agreement with the reciprocal BLAST orthologs. Some genes are split into two or more ORFs in one of the genomes. In these cases, the split genes in one genome were associated with a single gene in the other genome (example: Mb0074 and Mb0075 map to Rv0073). The number of 'TA' sites for each gene was calculated and only gene pairs with equal number +/- 1 'TA' sites were considered orthologs for comparison purposes. The gene set was annotated to show the existence of SNPs in the *M. bovis* genome relative to *M. tuberculosis*, and with a 'variable' or 'identical' label based on the amino acid sequence of the protein product as in (Malone et al., 2018).

3.4.5 Analysis of non-coding RNA

Intergenic regions expressed in *M. tuberculosis* were selected from previously compiled list of non-coding RNA expressed in *M. tuberculosis* using *baerhunter*, as described in Chapter 2. These were filtered to include only intergenic sRNA and 3' UTRs, resulting in 253 genomic features. All ncRNAs were trimmed to 200 bp to ensure the UTRs did not overlap any coding regions and prevent overlong transcripts, as expression of non-coding RNAs show indistinct 3' boundaries (Ju et al., 2024; Wade & Grainger, 2014). MUMMER v3.2 (Marçais et al., 2018) was used to align the query *M. tuberculosis* non-coding RNAs to the *M. bovis* genome reference

sequence (AF2122/97) and obtain coordinates of matching sequences within a threshold percent identity. This resulted in 157 *M. tuberculosis* ncRNA elements also present in *M. bovis*. The HMM predictions for each 'TA' site within the feature were used to determine a prediction for the entire feature, using the majority call and noting both calls in case of ties.

3.4.6 Functional enrichment

Overrepresentation of genes assigned to a particular functional category was determined using Mycobrowser annotation (Kapopoulou et al., 2011) and tested using Fisher's exact test with mid p adjustment and BH correction for multiple testing with the "exact2x2" R package (Fay, 2009). Gene set enrichment analysis (GSEA) (Subramanian et al., 2005) was performed using the "clusterProfiler" R package (Wu et al., 2021) to discover whether genes with similar log₂ fold-changes after treatment were enriched for any COG (clusters of orthologous genes), GO (gene ontology) terms or KEGG pathways (Ashburner et al., 2000; Galperin et al., 2021; Kanehisa et al., 2022). Analysis was performed using ranked signed-log-p-value (SLPV, the log₂ fold-change multiplied by the log of the p-value) with BH correction for multiple testing (adj. p-value).

All bioinformatic scripts are available at [https://github.com/jenjane118/thesis_work/tree/main/Chapter 3](https://github.com/jenjane118/thesis_work/tree/main/Chapter_3). Data manipulation and plots were created using R (version 4.3.1, 2023-06-16) with the following packages: dplyr, ggplot2, VennDiagram, eulerr and Circlize (Gu et al., 2014).

3.5 RESULTS

3.5.1 Libraries show good saturation of 'TA' sites

The saturation, or proportion of 'TA' sites that had reads mapped to them (also known as 'insertion density'), ranged between 13-49% for the *M. bovis* sub-samples and 32-36% for the *M. tuberculosis* replicates (Table 3.1). The cumulative number of 'TA' sites with mapped insertions was 40482 of 73,536 possible sites (55%) for *M. bovis* and 29,919 of 74,604 (40%) for *M. tuberculosis* (Figure 3.2). These numbers compare favourably with other tn-seq studies (Butler et al., 2020; Griffin et al., 2011;

Mendum et al., 2019; Minato et al., 2019; Zhang et al., 2012) but are significantly less than an assay by Dejesus et al (Dejesus et al., 2017) which used multiple composite *M. tuberculosis* libraries to achieve 84.3% saturation of 'TA' sites, which appears to be the maximum number of permissible insertions. However, the saturation levels of the individual libraries in the Dejesus et al. study was between 42-64%. Another published *M. tuberculosis* study by Patil et al., (Patil et al., 2021) created a single library with 67.1% of 'TA' sites with insertions.

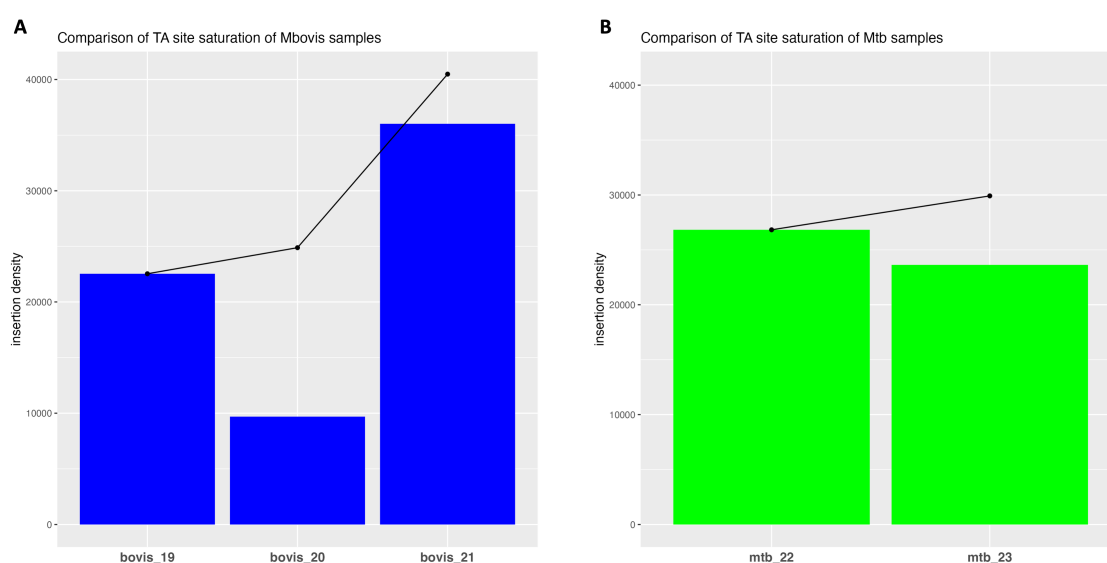


Figure 3.2. 'TA' site saturation for tn-seq libraries. Saturation measured in number of unique 'TA' sites with insertions. Black line indicates cumulative number of unique insertions. a) *M. bovis* sub-samples from single library (increasing CFU recovered). b) *M. tuberculosis* technical replicates.

The *M. bovis* samples represented sub-samples of increasing CFU plated from a single transposon library (Table 3.1). The location and number of insertion reads per gene in these sub-samples showed moderate correlation (Pearson's correlation coefficients of 0.57-0.68) with samples 19 and 21 better correlated than 20. This is likely due to the sparse coverage of 'TA' sites in sample 20. Samples 19 and 21 each had ~5M reads mapped with non-zero means of 168.2 and 215.5. Sample 20 had only ~34,000 reads mapped and a non-zero mean of 3.5 despite having double the number of colonies used for gDNA extraction as sample 19 (Table 3.1). The two technical replicates of the *M. tuberculosis* library were relatively well correlated (Pearson's correlation coefficient = 0.6, p-value < 2.2e-16). They both had over 1M reads mapped and non-zero means of 51.2 and 57.3. There was a positive trend for increase in saturation of 'TA' sites with the number of mapped reads (Figure 3.3).

Library	Sample	Colonies in sample	Reads mapped to 'TA' sites	Unique 'TA' site Insertions	Insertion density (saturation)
<i>M. bovis</i>	Mbovis_19	~5000	4856314	22539	30.7%
<i>M. bovis</i>	Mbovis_20	~10000	34035	9692	13.2%
<i>M. bovis</i>	Mbovis_21	~35000	6057051	36021	49.0%
<i>M. tb</i>	Mtb_22	~15000	1538393	26828	36.0%
<i>M. tb</i>	Mtb_23	~15000	1210855	23634	31.7%

Table 3.1. Sequencing statistics from *M. bovis* and *M. tuberculosis* *tn-seq* libraries.

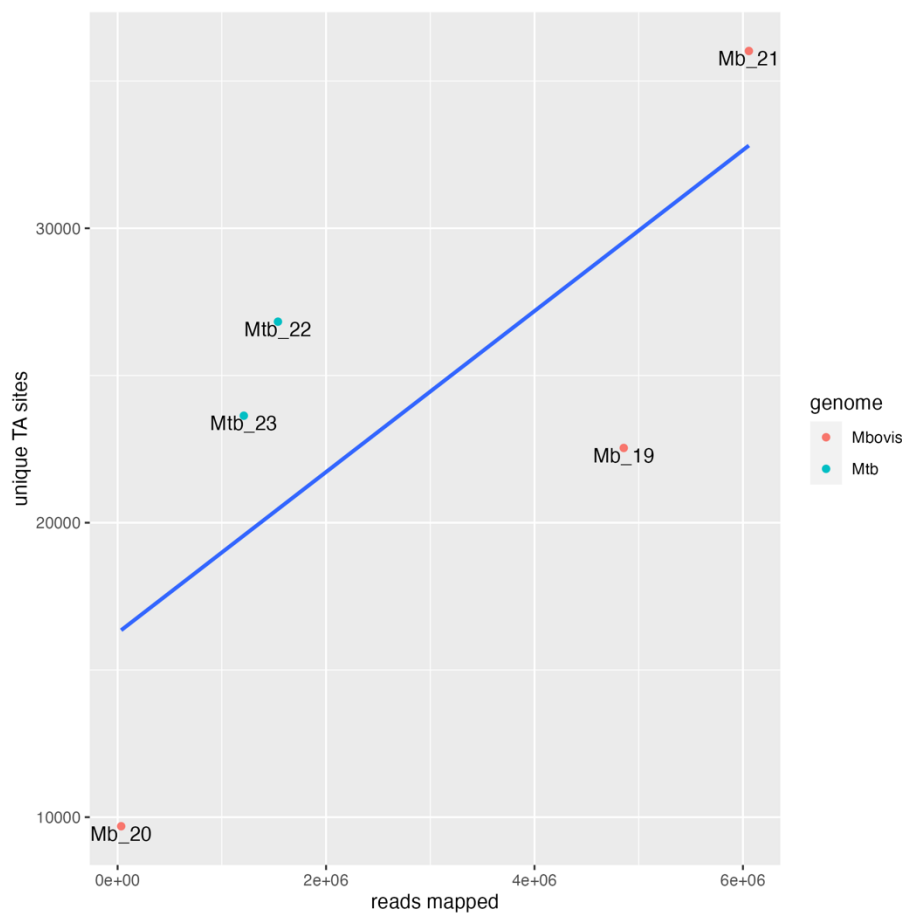


Figure 3.3. The number of mapped reads is loosely correlated to the number of unique 'TA' sites with insertions.

Histograms of non-zero reads resembled a geometric distribution but quartile-quartile plots, comparing the data distribution with the theoretical geometric distribution, show skew in all *M. bovis* samples, especially Mbovis_19 and 20 (Appendix A3.1, Figures S1, S2). This may be an indication of PCR duplication events

which cause outliers of large numbers of reads at a few sites, or technical issues with sampling or sequencing. A decision was made to exclude Mbovis_20 from subsequent analysis and sum the insertion site reads of the remaining two sub-samples as it appeared to be an outlier in respect to saturation and read distribution. This did not significantly reduce the number of unique insertions in the cumulative library and the final saturation was 54% with 39,987 unique insertions. The reads were summed for the technical *M. tuberculosis* replicates, resulting in 29,919 unique insertions and saturation of 40%. Insertions were evenly distributed throughout the genome and detected in 91% of coding sequences in *M. bovis* (3625 of 3990 genes) and 88% (3554 of 4019 genes) in *M. tuberculosis* (Figure 3.4). However, some genes will continue to function with insertions at the extremes of the transcript (Griffin et al., 2011), or in select domains, (Dejesus et al., 2017; Patil et al., 2021) and so some genes essential for survival may still be represented among genes with insertions.

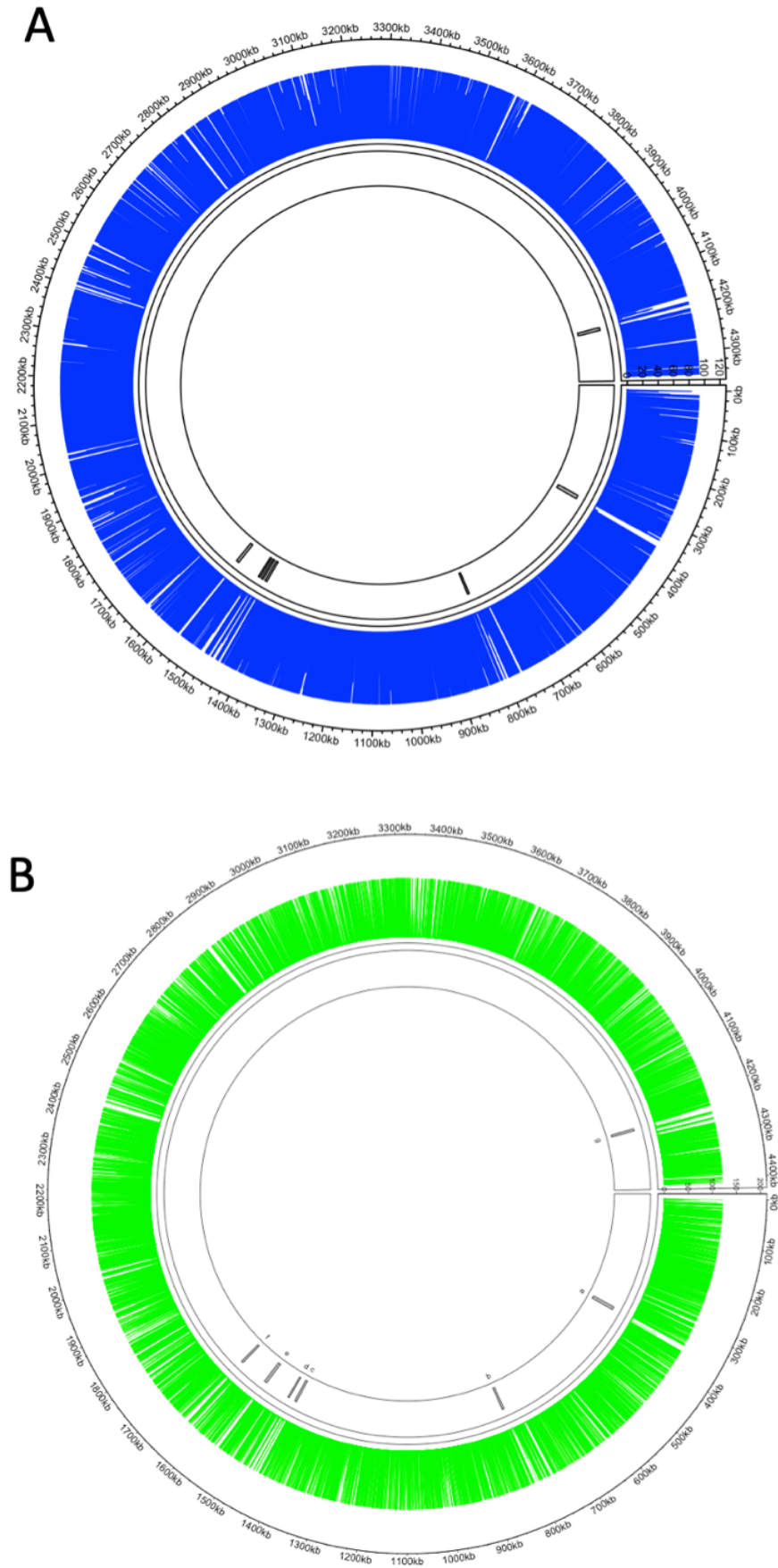


Figure 3.4. Insertion sites were well-distributed across the genomes. Length of coloured lines indicate read coverage. Blocks indicate gaps in coverage which correspond to known essential regions of genome. a) *M. bovis* b) *M. tuberculosis*. Plots made with the R package, Circlize (Gu et al., 2014).

3.5.2 Essential genes overlap with published datasets

The TRANSIT *HMM* method (DeJesus et al., 2015) was used to predict essentiality of each gene. In total, there were 530 genes predicted to be essential (ES) in *M. bovis* (13.2% of genes analysed) and 489 predicted ES in *M. tuberculosis* (12%) (Figure 3.5, A3.1 Supplemental Tables: Ch3_Supp_Table_3). The proportion of ES genes in our datasets (13.2% for *M. bovis*, 12% for *M. tuberculosis*) are comparable to the saturated *M. tuberculosis* dataset by DeJesus et al (11.5%). 326 (52.2%) of the essential genes observed for *M. tuberculosis* overlapped with the 461 ES calls from the DeJesus et al study (Figure 3.6)(DeJesus et al., 2017). The highly saturated single *M. tuberculosis* library by Patil et al., sequenced over 5 passages, identified 678 genes (close to 17% of coding genes) that had no fluctuations in insertions at 'TA' sites throughout the passages in over 60% of the gene body and were therefore considered essential for survival (Patil et al., 2021). 412 of these genes were called ES in this *M. tuberculosis* dataset (60.8% of them) (Figure 3.6C). The Butler et al *M. bovis* dataset predicted a lower proportion of genes to be ES (7.3%) but had a very similar level of saturation to this *M. bovis* dataset (58% 'TA' sites with insertions vs. 54% in this experiment). The *M. bovis* library shared 220 (35.2%) of ES calls with the Butler et al dataset (Figure 3.6A) (Butler et al., 2020). However, the Butler et al study indicated more genes with a growth defect (GD), (322 vs 176 in this study). Half of these (161) were called ES in this study and 30 GD genes in this study were called ES in Butler et al. Overall, there were 463 genes that showed some level of survival defect (predicted either ES or GD) in both *M. bovis* datasets (Figure 3.6A, B).

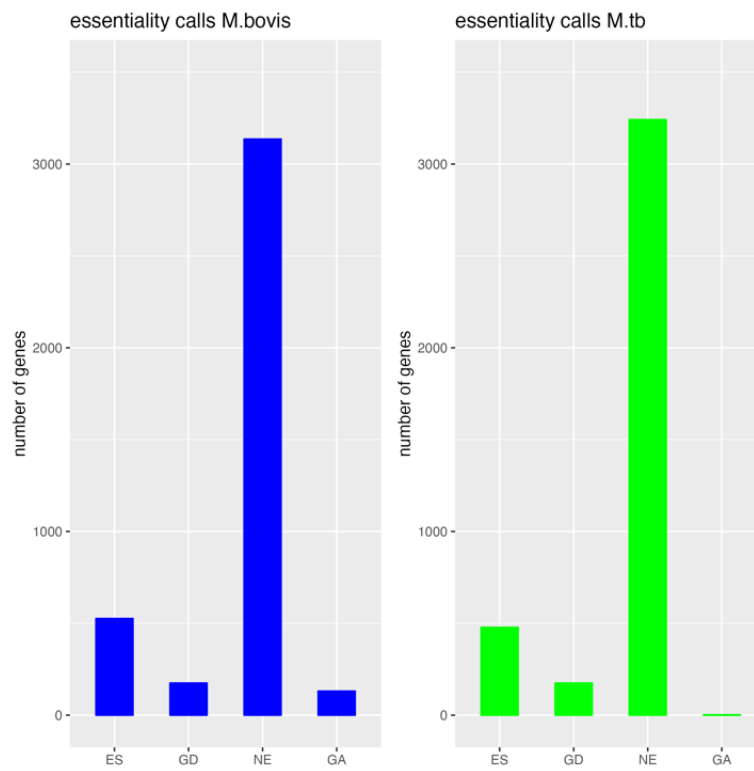


Figure 3.5. Essentiality calls for *M. bovis* and *M. tuberculosis* genomes with TRANSIT HMM. 'ES'=essential, 'GD'=growth defect, 'NE'=non-essential, 'GA'=growth advantage.

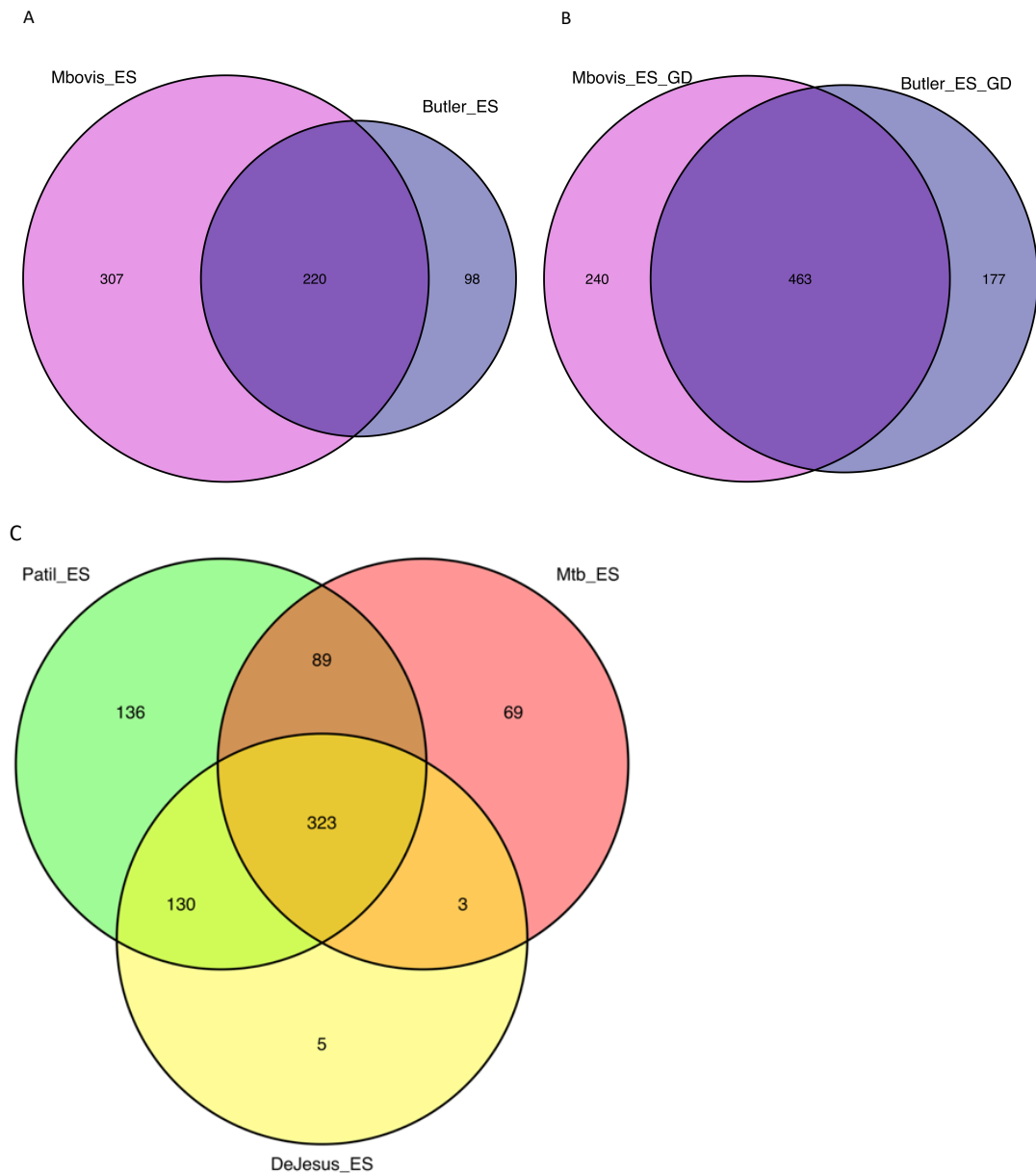


Figure 3.6. Essential genes in *M. bovis* and *M. tuberculosis* show a large overlap with published datasets (A3.1 Supplemental Tables: Ch3_Supp_Table_3). A) Overlap between *M. bovis* essential genes ('Mbovis_ES') and essential ('Butler_ES') genes in Butler et al, 2020. B) Including both essential and growth defect genes ('Mbovis_ES_GD') shows greater overlap (66% of ES/GD calls) with the Butler et al dataset. C) Overlap of *M. tuberculosis* essential genes ('Mtb_ES', 484 total) with DeJesus et al (461 total) and Patil et al (678 total). (DeJesus et al, 2017; Patil et al, 2021).

3.5.3 Comparing essentiality between *M. bovis* and *M. tuberculosis* orthologous genes with a qualitative approach

Despite the high level of sequence similarity between the *M. bovis* and *M. tuberculosis*, deletions, gene merges and SNPs can affect the essentiality of a gene. If an orthologous gene has a different essentiality prediction in the two datasets, it may be due to either a difference in the gene sequence, and therefore translated

protein product, or it could be a result of the identical gene product being more, or less, important to the survival of the particular host-adapted strain of MTBC. In addition, SNPs and differences in ORF length between orthologous genes can lead to a different number of possible 'TA' sites within the annotated gene boundaries which can change the probabilities used to calculate essentiality. To tease apart these possibilities, further analysis was made restricting the gene set to 3587 unique pairs of orthologous genes with a maximum difference of +/- 1 'TA' site in the gene sequence. Differences in essentiality among this set of genes should not be related to differences in number of possible insertion sites or large insertions or deletions; however, SNPs that do not create or destroy 'TA' sites may still be present. 363 of these orthologous gene pairs (10 %) were called essential in both species (Figure 3.7, A3.1 Supplemental Tables: Ch3_Supp_Table_1). The overlapping essential genes were enriched with the functional categories of 'information pathways' and 'intermediary metabolism and respiration' (adjusted p-values: 8.24×10^{-21} and 2.03×10^{-15} , respectively, using hypergeometric test). 298 orthologous genes were designated 'growth defect' (GD) genes in one or both species (Figure 3.7). Insertions in these genes presumably have a less deleterious effect on survival than essential genes. 61 of the ES orthologous genes in *M. bovis* were called GD in *M. tuberculosis* and 39 ES genes in *M. tuberculosis* were GD in *M. bovis*. In total, 492 (13.7%) of orthologous gene pairs showed a fitness defect of some degree (called ES or GD) in both the *M. bovis* and *M. tuberculosis* tn-seq libraries.

M. bovis had more orthologous genes than *M. tuberculosis* with a 'growth advantage' (GA) prediction (115 vs 1) which indicates a relative advantage to inactivating the gene product. These genes were non-essential (NE) in *M. tuberculosis*. In contrast, the composite saturated library from DeJesus, et al, called 244 of the orthologous *M. tuberculosis* genes GA, while the *M. bovis* library from Butler et al only predicted 2 GA genes in *M. bovis*. The 115 *M. bovis* GA genes were not found to be enriched for any functional category (using hypergeometric test). Stochastic processes make it difficult to discern a true difference between levels of insertions that indicate an advantage to growth versus having no effect and, therefore, the difference in the number of GA genes in the two libraries could be due to the lower saturation level of the *M. tuberculosis* library relative to *M. bovis*.

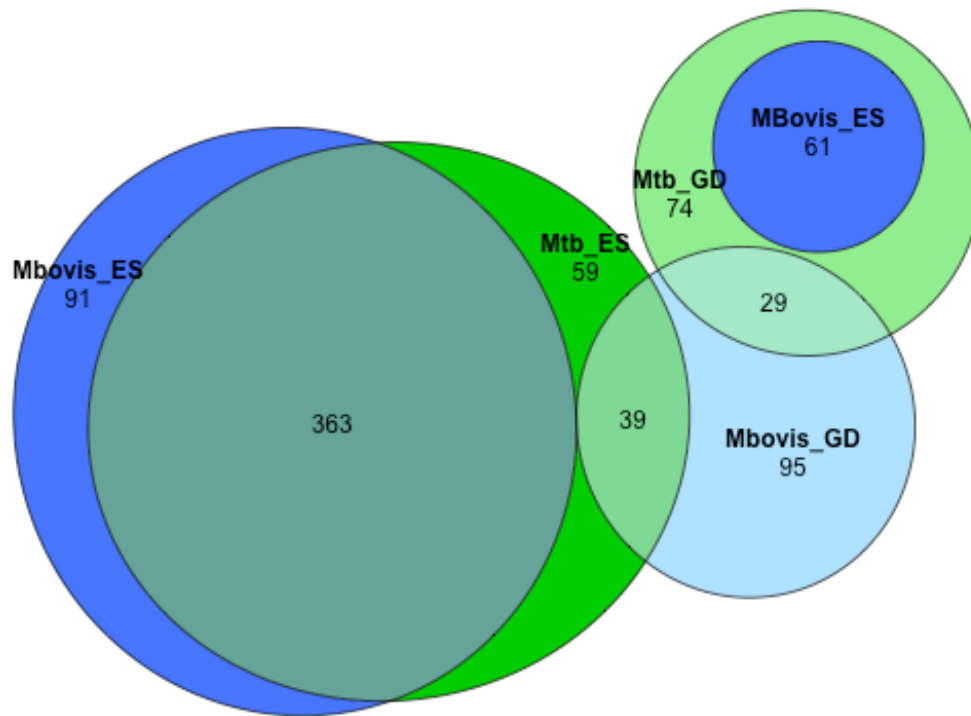


Figure 3.7. There is significant overlap of essential and growth defect calls among orthologous protein-coding genes in the two genomes. 363 genes were ES in both *M. bovis* and *M. tuberculosis*. A further 61 *M. bovis* ES genes overlap *M. tuberculosis* GD genes (smaller royal blue circle) and 39 *M. tuberculosis* ES genes overlap *M. bovis* GD genes. ES = essential, GD = growth defect. Diagram made using eulerr package in R.

3.5.4 Determination of different non-coding RNA requirements

It is possible that there are non-coding RNA features that are required for survival in the MTBC, and the importance of these elements may differ between *M. tuberculosis* and *M. bovis*. Therefore, an attempt was made to compare the effects of transposon insertions on annotated and predicted non-coding RNA between the species. Annotated and predicted intergenic short RNAs and 3' UTRs that did not overlap any coding regions on the opposite strand were selected for comparison (see Materials and Methods). 5' UTRs were not included in the analysis, as these often include promoter elements that, when disrupted, directly affect the transcription of a possibly essential downstream coding gene (polar effects) rather than comprising an independently-regulated transcript. Sequence comparison verified the presence and coordinates of the non-coding features in both *M. bovis* and *M. tuberculosis* genomes and comparisons were made between the HMM predictions for the entire region. The limited saturation in the libraries, combined with the low number of 'TA' sites in characteristically short ncRNA features, makes it difficult to be confident in what may be arbitrary differences in insertions across

a low number of sites. Limiting the results to ncRNA transcripts with 3 or more 'TA' sites, 15 had different HMM predictions in *M. bovis* versus *M. tuberculosis* (Table 3.2, A3.1 Supplemental Tables: Ch3_Supp_Table_2), including two annotated RNAs: G2/ncRv11689c and MTS0858/ncRv1092c, both predicted to have an increased requirement in *M. tuberculosis* than in *M. bovis*.

Table 3.2. Non-coding RNAs with different essentiality calls in *M. bovis* versus *M. tuberculosis* with >2 'TA' sites. *Indicates a tie between number of 'TA' sites with a particular call (see Methods and Materials 3.4.5)

Mtb ID	start Mb	end Mb	width Mb	strand Mb	TAs Mb	call Mb	start Mtb	end Mtb	width Mtb	strand Mtb	TAs Mtb	call Mtb
putative_UTR:p731365_731493	732601	732729	128	+	4	NE	731365	731493	128	+	4	GD/NE*
putative_UTR:p733326_733523	734562	734759	197	+	3	GA	733326	733523	197	+	3	NE
MTS0858/ncRv1092c	1221788	1221888	100	-	3	NE	1220388	1220487	99	-	3	GD
G2/ncRv11689c	1904739	1904940	201	-	4	NE	1914990	1915190	200	-	4	ES
putative_sRNA:p2038697_2038955	2033561	2033761	200	+	9	GD	2038697	2038897	200	+	9	ES
putative_sRNA:m2207133_2207499	2203343	2203543	200	-	4	ES	2207299	2207499	200	-	4	GD
putative_UTR:p2265039_2265188	2248365	2248514	149	+	4	NE	2265039	2265188	149	+	4	ES
putative_sRNA:p2500386_2500819	2483944	2484144	200	+	4	ES	2500386	2500586	200	+	4	GD
putative_sRNA:m2500131_2500820	2484178	2484378	200	-	3	ES	2500620	2500820	200	-	3	GD
putative_UTR:m3359212_3359584	3320411	3320611	200	-	4	NE	3359384	3359584	200	-	4	GD
putative_sRNA:p3467971_3468335	3428930	3429130	200	+	4	ES	3467971	3468171	200	+	4	NE
putative_sRNA:m3834596_3834735	3791637	3791776	139	-	8	GD	3834596	3834735	139	-	8	NE
putative_sRNA:p3907294_3907456	3857330	3857492	162	+	4	GA	3907294	3907456	162	+	4	NE
putative_sRNA:m3939324_3939465	3887974	3888115	141	-	3	GA	3939324	3939465	141	-	3	NE
putative_sRNA:m4386818_4387144	4325317	4325517	200	-	4	GA	4386944	4387144	200	-	4	NE

3.5.5 Quantitative comparison of gene requirements between orthologous genes in *M. tuberculosis* and *M. bovis* libraries

The TRANSIT *resampling* algorithm was also used to identify statistically significant differences in mean read counts between the orthologous gene pairs in the two genomes (DeJesus et al., 2015) as an orthogonal approach to the qualitative comparison of the HMM predictions. This permutation method calculates a p-value by shuffling the observed reads among all possible 'TA' sites in a designated region (typically, within protein-coding gene boundaries) and therefore should only be

applied to gene regions of roughly the same length and number of 'TA' sites. 32 genes showed statistically significant differences (adj p-value < 0.05) in the distribution of insertions between the two species, (Table 3.3, Figure 3.8, A3.1 Supplemental Tables: Ch3_Supp_Table_1). These genes showed an absolute \log_2 fold-difference in mean insertion counts greater than 2 (representing a 4-fold difference). Gene set enrichment analysis (GSEA) (Subramanian et al., 2005), using the ranked \log_2 fold-differences, showed enrichment of genes involved in sulfate assimilation, folate-containing compounds, quinone binding and amino acid metabolism, as well as KEGG pathways for 'sulfur metabolism' and 'glycine, serine and threonine metabolism' (adj. p-values < 0.05), indicating these functions show the most divergence in requirements between the human and animal-adapted strains.

Corresponding feature coordinates from the *M. tuberculosis* ncRNA annotations were used with the TRANSIT *resampling* method. Analysis was restricted to features that contained 3 or more 'TA' sites, for a total of 157. There were no statistically significant \log_2 fold differences between the species.

Table 3.3. Orthologous genes showing statistically significant \log_2 fold-difference between *M. tuberculosis* and *M. bovis* insertions using the TRANSIT resampling method ($p_{adj} < 0.05$). Genes with positive \log_2 fold-differences (green) are more important for *M. tuberculosis* survival relative to *M. bovis* and genes showing a negative \log_2 fold-difference (red) are more important for survival in *M. bovis* relative to *M. tuberculosis*.

Mtb Orf	Name	Mbovis Orf	Mtb call	Mbovis call	snps	Product	Description	TA sites	log2 FD	Adj p-value	Functional Category
Rv2454c	korB	Mb2481c	GD	NE	N	Identical	2-oxoglutarate oxidoreductase subunit KorB	15	7.99	0.0000	intermediary metabolism and respiration
Rv3579c	rlmB	Mb3610c	NE	GA	Y	Identical	23S rRNA (guanosine(2251)-2'-O)-methyltransferase	12	7.09	0.0429	intermediary metabolism and respiration
Rv1828		Mb1859	ES	NE	N	Identical	HTH-type transcriptional regulator	14	7.02	0.0060	conserved hypotheticals
Rv0812		Mb0835	NE	NE	Y	Variable	4-amino-4-deoxychorismate lyase	17	5.89	0.0000	intermediary metabolism and respiration
Rv0337c	aspC	Mb0344c	GD	NE	N	Identical	aspartate aminotransferase	22	5.88	0.0000	intermediary metabolism and respiration
Rv0467	icl1	Mb0476	NE	GA	N	Identical	isocitrate lyase	18	5.43	0.0026	intermediary metabolism and respiration
Rv1832	gcvB	Mb1863	ES	GD	Y	Variable	glycine dehydrogenase	57	4.19	0.0000	intermediary metabolism and respiration
Rv2455c	korA	Mb2482c	GD	NE	N	Identical	2-oxoglutarate oxidoreductase subunit KorA	37	3.9	0.0000	intermediary metabolism and respiration
Rv1239c	corA	Mb1271c	NE	NE	Y	Variable	Mg and Co transport transmembrane protein CorA	25	3.66	0.0153	cell wall and cell processes
Rv3497c	mce4C	Mb3527c	NE	GA	N	Identical	Mce family protein Mce4C	18	3.11	0.0060	virulence/ detoxification/ adaptation
Rv0485		Mb0495	NE	GA	N	Identical	transcriptional regulator	23	2.53	0.0225	regulatory proteins
Rv2940c	mas	Mb2965c	NE	GA	Y	Identical	multifunctional mycrocercic acid synthase	82	2.28	0.0000	lipid metabolism
Rv0642c	mmaA4	Mb0661c	NE	GA	Y	Variable	hydroxymycolate synthase MmaA4	18	2.22	0.0225	lipid metabolism
Rv0643c	mmaA3	Mb0662c	NE	GA	N	Identical	methoxy mycolic acid synthase MmaA3	22	2.2	0.0308	lipid metabolism
Rv1638	uvrA	Mb1664	NE	NE	N	Identical	excinuclease ABC subunit UvrA	39	-2.17	0.0268	information pathways
Rv1006		Mb1033	NE	NE	Y	Variable	hypothetical protein	41	-2.28	0.0209	conserved hypotheticals
Rv0016c	pbpA	Mb0016c	NE	NE	N	Identical	penicillin-binding protein PbpA	37	-2.58	0.0143	cell wall and cell processes
Rv1937		Mb1972	NE	NE	Y	Identical	oxygenase	48	-2.69	0.0225	intermediary metabolism and respiration
Rv2942	mmpL7	Mb2967	NE	NE	Y	Identical	transmembrane transport protein MmpL7	43	-2.85	0.0000	cell wall and cell processes
Rv3158	nuoN	Mb3182	NE	NE	Y	Identical	NADH-quinone oxidoreductase subunit N	32	-3.31	0.0000	intermediary metabolism and respiration
Rv3156	nuoL	Mb3180	NE	NE	Y	Variable	NADH-quinone oxidoreductase subunit L	35	-3.39	0.0060	intermediary metabolism and respiration
Rv2945c	lppX	Mb2970c	NE	NE	N	Identical	lipoprotein LppX	10	-3.56	0.0163	cell wall and cell processes
Rv1270c	lprA	Mb1301c	NE	NE	Y	Variable	lipoprotein LprA	11	-3.59	0.0473	cell wall and cell processes
Rv3682	ponA2	Mb3707	NE	GD	N	Identical	bifunctional transglycosylase/transpeptidase	37	-4.1	0.0094	cell wall and cell processes
Rv3859c	glbB	Mb3889c	GD	ES	Y	Variable	glutamate synthase large subunit	80	-4.22	0.0000	intermediary metabolism and respiration
Rv2400c	subI	Mb2422c	NE	ES	N	Identical	sulfate ABC transporter substrate-binding lipoprotein	19	-5.45	0.0187	cell wall and cell processes
Rv3490	otsA	Mb3520	NE	ES	Y	Variable	trehalose-phosphate synthase	30	-5.69	0.0060	virulence/ detoxification/ adaptation
Rv1286		Mb1317	NE	ES	Y	Variable	adenyllyl-sulfate kinase	34	-6.19	0.0000	intermediary metabolism and respiration
Rv3680		Mb3705	NE	GD	N	Identical	anion transporter ATPase	20	-6.22	0.0026	cell wall and cell processes
Rv0455c		Mb0463c	NE	ES	N	Identical	hypothetical protein	11	-6.53	0.0492	conserved hypotheticals
Rv2222c	glnA2	Mb2246c	NE	GD	N	Identical	glutamine synthetase	20	-7.39	0.0000	intermediary metabolism and respiration
Rv0244c	fadE5	Mb0250c	NE	ES	Y	Variable	acyl-CoA dehydrogenase	23	-7.49	0.0000	lipid metabolism

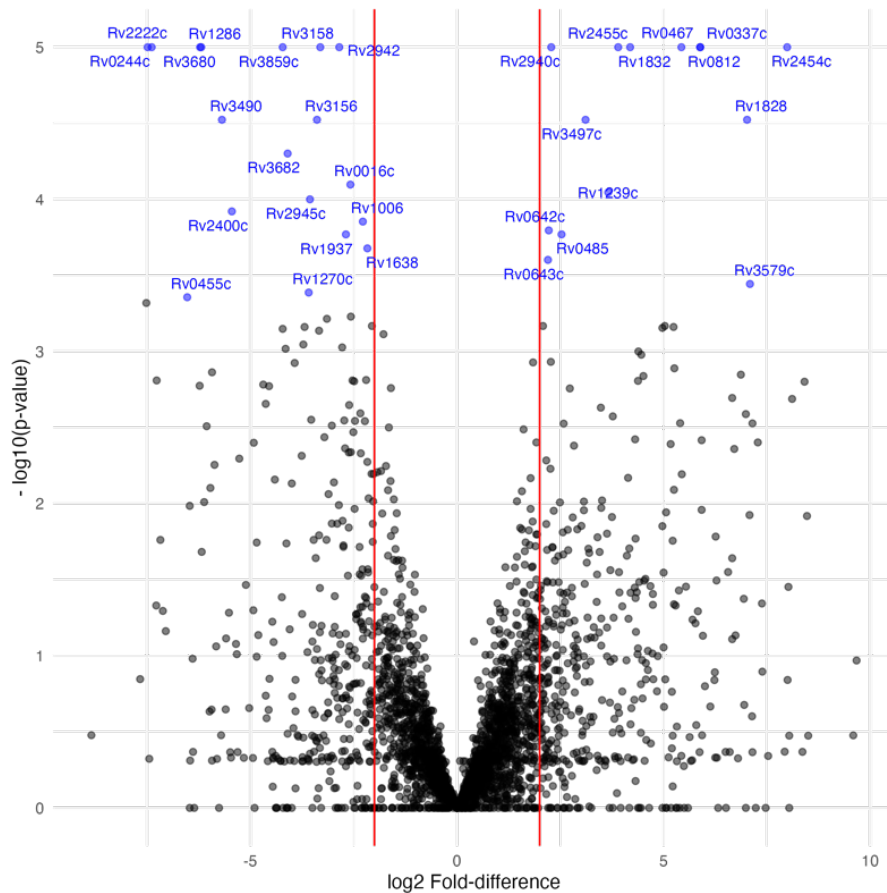


Figure 3.8. Volcano plot showing genes with statistically significant ($p_{adj} < 0.05$) \log_2 fold-difference in *M. tuberculosis* vs *M. bovis* mean insertion counts in blue determined with TRANSIT resampling. Red bars indicate \log_2 fold-difference of ± 2 (4-fold difference). Genes to the left of 0 are more required for *M. bovis* survival and genes to the right are more required for *M. tuberculosis* survival..

3.6 DISCUSSION

3.6.1 Insertions in identical metabolism and respiration genes show different fitness effects in *M. bovis* versus *M. tuberculosis*

14 orthologous genes showed a statistically significant higher mean insertion count in the *M. bovis* library, indicating these are less important for *in vitro* survival in *M. bovis* than in *M. tuberculosis* (Table 3.3). This result is in contrast to the large number of GA calls for *M. bovis* versus *M. tuberculosis* using the HMM method (115 vs 1). Half of these genes are involved in 'intermediary metabolism and respiration'. For example, despite having identical protein products in both species, *aspC* (Mb0344c/Rv0337c) (aspartate aminotransferase) and the *korAB* operon (Mb2482-1c/Rv2455-4c) (2-oxoglutarate oxidoreductase subunits) are predicted to be less important for survival of *M. bovis* than for *M. tuberculosis*. Another

example, Icl1, isocitrate lyase, is an enzyme involved in the assimilation of fatty acids through the glyoxylate shunt and is identical in both species. In previous *M. tuberculosis* studies, *icl1* (Rv0467) showed conditionally-dependent essentiality based on carbon source (Dejesus et al., 2017; Griffin et al., 2011; Minato et al., 2019; Serafini et al., 2019). It has also been observed that Icl1 is expressed at a constitutively lower level in *M. bovis* and *Mycobacterium bovis* BCG than in *M. tuberculosis* (Lee et al., 2018; Malone et al., 2018). In this study, this gene was not predicted to be essential for either species (and was predicted GA in *M. bovis*), but *M. bovis icl1* contained significantly more insertions in the gene versus *M. tuberculosis*, perhaps indicating an advantage to deactivating the gene in *M. bovis*.

A potential HTH-type, MerR-family, transcriptional-regulator gene (Mb1859/Rv1828), appears dispensable in *M. bovis* in comparison with *M. tuberculosis*, with a higher mean number of insertions in the *M. bovis* gene. The orthologs are identical in gene sequence and protein product. Rv1828 has been shown to both activate and repress gene expression at specific promoter sequences in *M. tuberculosis* and may also be implicated in metabolism/respiration functions. Rv1828 binds its own promoter ahead of *garA* (Rv1827), a gene linked to regulation of the TCA-cycle and glutamate synthesis (Singh et al., 2018).

Several genes involved in nitrogen metabolism and sulfate assimilation were found to have a greater requirement in *M. bovis* than in *M. tuberculosis*, showing significant decreases in mean number of insertions in *M. bovis* versus *M. tuberculosis* (Table 3.3). L-glutamine synthesis from L-glutamate and ammonia is important for formation of the peptidoglycan layer of the cell wall in pathogenic mycobacteria (Shaku et al., 2023) and for nitrogen assimilation (Harth et al., 1994; Tripathi et al., 2013; Viljoen et al., 2013). *GlnA2* (Mb2246c/Rv2222c) is a 'possible glutamine synthetase' and codes for identical protein products in both genomes but appears to be more crucial for survival of *M. bovis*. The secreted paralog, *glnA1*, is essential for both species and recognised as a virulence factor involved in host-pathogen response (Harth et al., 1994; Parveen et al., 2023). The gene for the large subunit of glutamine oxoglutarate aminotransferase, *gltB* (Mb3889c/Rv3859c), has fewer mean insertions in *M. bovis* than *M. tuberculosis*. This enzyme is responsible for converting L-glutamine to L-glutamate and differs between the species by a single amino acid substitution

(glutamic acid to the amide in *M. bovis*) at position 550. The small subunit, *gltD* also had decreased mean number of insertions in *M. bovis*, but this difference was not statistically significant (possibly because the small subunit has only 1/4 of the number of 'TA' sites as the large subunit). GltB contains an iron-sulfur cluster, and two genes involved in the sulfate assimilation pathway, *cysN* (Rv1286/ Mb1317) and *subI* (Rv2400c/Mb2422c), appear to be more required in *M. bovis* than in *M. tuberculosis*. CysN has a V/L amino acid substitution at residue 245, but SubI is identical in both species. It may be fruitful to explore whether the observed difference in *in vitro* requirements for these genes in *M. bovis* indicate a different priority for regulating nitrogen assimilation (Tripathi et al., 2013; Viljoen et al., 2013).

3.6.2 Differences in requirements for cell wall-associated genes may reflect species-specific cell-wall lipid repertoires

The cell walls of the MTBC have distinct lipid repertoires that most likely reflect adaptation to host immune systems, including differential production of sulfolipids (SL), diacyltrehaloses (DAT) and polyacyltrehaloses (PAT), and phthiocerol dimycocerosates (PDIM) (Malone et al., 2018; Malone & Gordon, 2017). The differences in the essentiality of related genes between the species may have some relation to differences in the composition of the cell walls which may relate to host-specific interactions. Insertions in several genes associated with synthesis and catabolism of lipids were *overrepresented* in *M. bovis* relative to *M. tuberculosis*, indicating these genes are more dispensable for survival of *M. bovis*. The genes *mas*, *mce4C*, *mmaA3* and *mmaA4* are involved with synthesis of lipids present in the cell walls. All of these genes produce identical protein products in the two species. A gene shown to be involved in peptidoglycan biosynthesis, Rv0812/Mb0835, also had significantly higher mean insertion count in *M. bovis* (Black et al., 2021) with the *M. bovis* protein product having a single amino acid substitution M/I at position 145.

Several of the genes showing significantly lower mean insertion counts in *M. bovis* relative to *M. tuberculosis*, and therefore more required for *in vitro* survival, are associated with cell wall lipid synthesis in the MTBC. For instance, insertions in *fadE5* (Rv0244c/ Mb0205c), an acyl-CoA dehydrogenase involved in lipid metabolism induced a severe fitness cost in *M. bovis* relative to *M. tuberculosis*. This

acyl-CoA dehydrogenase appears to be involved in two processes in the mycobacterial cell: catabolism of fatty acids for use as a carbon source and processing fatty acids to feed into cell wall lipid biosynthesis. Disruption of *fadE5* causes increased susceptibility to antibiotics in *M. tuberculosis*, presumably due to membrane disruption (Chen et al., 2020), however, SNPs have been recorded from hyper-virulent strains that increase drug resistance (Bellerose et al., 2020; Rajwani et al., 2022). The *M. bovis* gene, Mb0205c, has 2 SNPs relative to *M. tuberculosis*: a K/R substitution at residue 394 and an E/A at 479 (Brenner & Sreevatsan, 2023). In other tn-seq studies, *fadE5* was predicted to be ES for *M. tuberculosis* on cholesterol and in rich media (Griffin et al., 2011; Minato et al., 2019).

Three genes related to peptidoglycan biogenesis in the cell wall were also more necessary for *M. bovis* survival relative to *M. tuberculosis*. These genes include: *ponA2* (Rv2490/Mb3707), coding for a bifunctional enzyme (and penicillin-binding protein) which can both polymerise the glycan strands and cross-link the strands together (Kieser et al., 2015) which is identical in DNA and peptide sequence in *M. tuberculosis* and *M. bovis*. Previous studies have shown this gene to be conditionally-essential for *M. tuberculosis* in rich media (Minato et al., 2019) and required for mouse infection (Vandal et al., 2009) but insertions were advantageous for survival in the saturated *M. tuberculosis in vitro* tn-seq study (Dejesus et al., 2017). In a tn-seq study in Δ *ponA1* (a redundant, but non-identical enzyme) and Δ *ponA2* backgrounds, *otsA* (Rv3490/Mb3520), alpha-trehalose-phosphate synthase, was found to be essential in both deletion backgrounds (Kieser et al., 2015). There is a V/L substitution at position 334 in *M. bovis otsA*, and in this study, this gene had fewer insertions (and more importance for survival) in *M. bovis* relative to *M. tuberculosis*. Finally, the penicillin-binding protein (PBP), *pbpA* (Rv0016c/Mb0016c) is less important for *M. bovis* survival. It is involved in cross-linking peptidoglycan strands, contributing to membrane shape and rigidity (Birhanu et al., 2019).

3.6.3 Other membrane-associated genes important for in vitro survival in *M. bovis*

Rv3680/Mb3705 is part of a two-gene operon of ATP-ases linked to protection against nitric oxide and glycerol toxicity in *M. tuberculosis* (Whitaker et al., 2020).

These genes are identical in both genomes but showed a severe growth penalty in *M. bovis* compared to *M. tuberculosis*. Whitaker et al found an interaction with the *M. tuberculosis* ortholog of this gene and *glpK*, a glycerol kinase involved in the first step of glycerol metabolism; increased susceptibility of a Δ rv3679-3680 mutant to glycerol toxicity was reversed with mutation of *glpK*, (Whitaker et al., 2020). In *M. bovis*, *glpK* is annotated as two overlapping transcripts (Mb3721c/*glpKa* and Mb3722c/*glpKb*) with a frameshift mutation in *M. bovis* resulting in a truncated peptide product. Subsequent to this analysis, genomic sequencing of the lab strain of *M. bovis* AF2122/97 used to make these transposon libraries was shown to have a mutation that corrects the frameshift. *M. bovis* is unable to utilise carbohydrates such as glycerol as carbon sources due to another mutation in *pykA* which codes for the enzyme in the final stage of glycolysis (Keating et al., 2005). Perhaps an active glycerol kinase with an inactive *pykA* means this lab strain of *M. bovis* is more vulnerable to glycerol toxicity from phosphorylated intermediates and, therefore, more reliant on intact Mb3705/Rv3680 than *M. tuberculosis*.

Siderophores, such as mycobactin, are secreted by the bacteria to scavenge iron from the host in iron-poor environments, like that found inside of host macrophages. Rv0455c/Mb0463c has been shown to be required for siderophore secretion through the MmpL4 and MmpL5 efflux pumps in *M. tuberculosis* (Zhang et al., 2022). It is identical in sequence in *M. tuberculosis* and *M. bovis*, however, in this study, insertions in this gene caused a more severe impact on fitness in *M. bovis* versus *M. tuberculosis*, which is in agreement with other studies that found the gene to be non-essential for *M. tuberculosis* and essential for *M. bovis in vitro* (Butler et al., 2020; Dejesus et al., 2017). In *M. tuberculosis*, it has been found to be an essential gene for survival in mice and low-iron conditions (Zhang et al., 2022) and in a comparative tn-seq study using different iron sources, there was an increased requirement for the gene in high mycobactin concentrations versus heme (Zhang et al., 2020). The media used in this study includes defibrinated sheep blood and foetal bovine serum which contain heme, in addition to the iron-containing Middlebrook 7H11 media. As *M. bovis* is not able to utilise heme as an iron source due to a deletion in *ppe37* (Tullius et al., 2019), it therefore has a stronger reliance on siderophore secretion for scavenging iron than *M. tuberculosis*.

3.6.4 Comparing orthologous genes with different numbers of 'TA' sites

422 additional gene pairs are considered orthologous between the two strains but do not have the same number of 'TA' sites, due to gene fusions or deletions, and therefore application of the quantitative resampling method would not be appropriate. Despite the nucleotide differences, 90% of these pairs have the same prediction of essentiality with only 44 pairs with different essentiality predictions. The PE/PPE genes have variable numbers of proline-glutamate repeats that have been shown to differ between orthologs in clinical isolates, especially the PE_PGRS family (Fishbein et al., 2015), meaning orthologs will have different lengths and number of 'TA' sites. This mostly uncharacterised family of genes are generally associated with the outer cell envelope and some are involved in forming porin-like complexes for import of nutrients (Mitra et al., 2017; Wang et al., 2020). Three PE/PPE genes were found to be more required in *M. bovis* than in *M. tuberculosis*: PE_PGRS13, PE_PGRS54, and PPE50 (Rv3135). Three others, PPE_47, PE_PGRS55 and PE_PGRS57, were predicted to be ES in *M. tuberculosis* and NE in *M. bovis*.

3.6.5 Comparing results from HMM and Resampling Methods

There are biological reasons relating to culture conditions, such as the number of divisions before harvesting mutants⁵, and media composition that can cause shifting essentiality status and make direct comparisons between datasets difficult⁶ (Mahmutovic et al., 2020; Minato et al., 2019). This study attempts to address these challenges by standardising the culture conditions and comparing tn-seq libraries selected on identical culture media. However, differences in the saturation and stochastic distribution of insertions in the creation and sequencing of the different libraries will lead to some type-1 (falsely predicting a gene is essential) and type-2 (predicting a gene is non-essential, when it truly is essential) errors in the calls, despite the sophistication of the HMM algorithm. Direct comparison of the predicted essentiality calls from two different tn-seq analyses may risk compounding the errors and confounding the conclusions. Resampling is based entirely on indicated

⁵ Any apparent difference in doubling time between libraries was not recorded in this experiment but could be used to estimate any differences in the net division rate between the mutant libraries which could affect the survival of mutants where insertions have only a moderate effect on fitness (Mahmutovic et al., 2020).

⁶ Interestingly, the 14 different *M. tuberculosis* libraries consolidated to create a 'saturated' dataset in DeJesus et al were not all grown on the same media (DeJesus et al., 2017).

genome coordinate boundaries, and normalises the mean insertion counts for a specified genomic region only, before determining if there is a statistically significant difference in the insertion distributions between two datasets. For short genetic features, such as ncRNA, the reshuffling of insertions among a limited number of 'TA' sites limits its utility for generating p-values. In general, it is more conservative than comparing differences in essentiality calls, but it can highlight differences in distributions between genes that have the same essentiality prediction and can quantify the statistical significance of this difference. This information can also be used to add another layer of interpretation when considering those genes which also have different essentiality predictions with HMM.

For example, the gene *wag31* (Rv2145/Mb2169) was shown to be required for normal cell morphology in *M. tuberculosis* (Kang et al., 2008). In this study, it is predicted to be NE in *M. tuberculosis* but ES in *M. bovis*; however, resampling analysis based on the *wag31* gene boundaries shows no difference in the means between the orthologs (A3.1 Supplemental Tables: Ch3_Supp_Table_1). Previous tn-seq studies present disparate results: *wag31* has been shown to be ES in several *M. tuberculosis* studies (DeJesus et al., 2017; Griffin et al., 2011; Minato et al., 2019) and was predicted to be NE in *M. bovis* (Butler et al., 2020). A closer look at the insertions in the two datasets analysed here reveal that *wag31* has no insertions beyond the first possible 'TA' site in both *M. tuberculosis* and *M. bovis* (Figure 3.9). Due to low numbers of insertions in the adjacent genes, and in the first and last 'TA' sites within *wag31*, the HMM probability model makes a conservative calculation that the probability of the next 'TA' site in *wag31* changing state from NE to ES is very low in the *M. tuberculosis* dataset (as the lower number of insertions could be sparse due to stochastic factors rather than impairment of survival). Conversely, in *M. bovis*, the higher number of insertions present in Rv2144c and in the first 'TA' site of *wag31*, makes the probability of the state changing from GA to ES at the next insertion site much higher. This adjustment to different saturation levels is by design in order to prevent high numbers of false positives in unsaturated datasets (DeJesus et al., 2015) but in this case, may result in a false negative call. The final word is perhaps the phenotypic analysis of CRISPRi strains with knocked-down expression of *wag31*

which showed equally severe growth defects in both *M. tuberculosis* and *M. bovis* (Gibson et al., 2021).

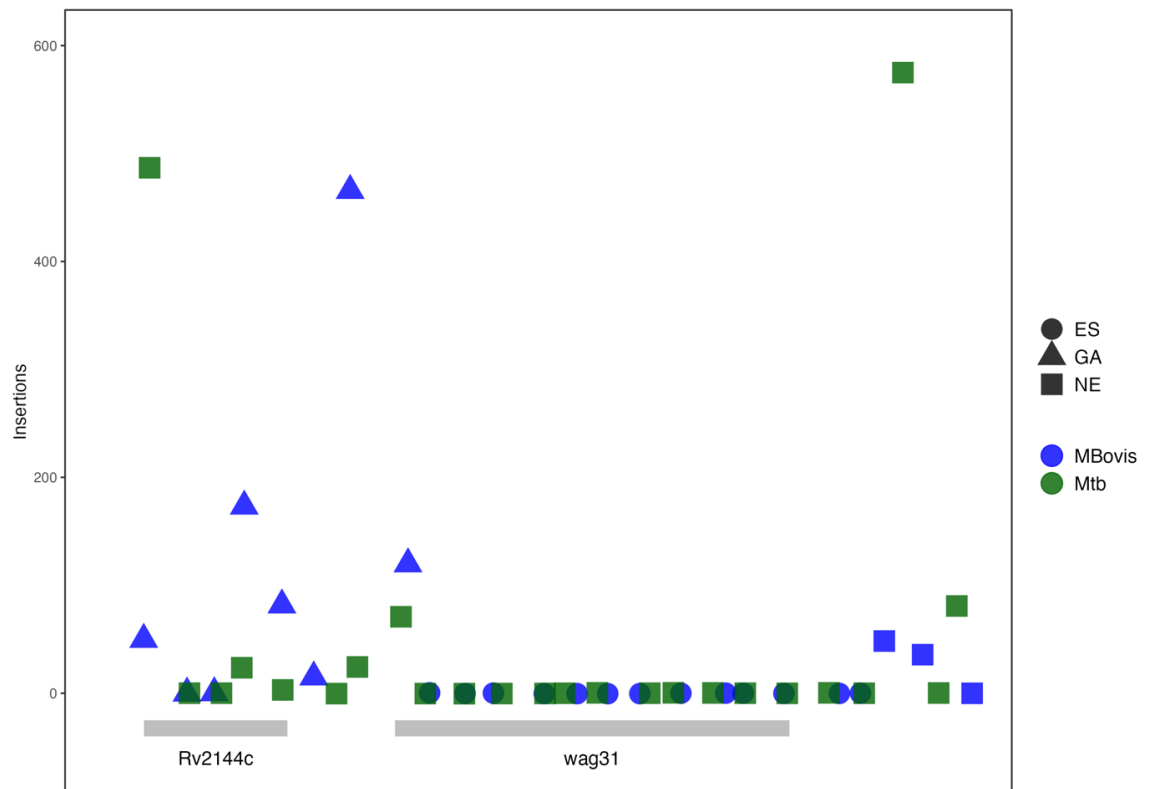


Figure 3.9. The normalised insertion counts for wag31 are similar in the *M. bovis* and *M. tuberculosis* libraries though the essentiality calls differ. Equivalent insertion sites for the positions relative to the two genes in the two genomes plotted on x-axis.

3.6.6 Limitations of the study

There are several points at which a bottleneck effect can influence the diversity of the insertion library, such as random sampling at several stages during the culturing and sequencing steps. In addition, depending on the birth/death rate of a population and the relative fitness of the other mutants in a population, over time, mutants with relatively mild growth defects can become outcompeted in the population and disappear from the mutant pool altogether--and thus be predicted as 'essential' (DeJesus & Ioerger, 2013; Mahmutovic et al., 2020). The resulting reduction in the saturation of the library can lead to false positives, especially when comparing libraries grown in different culture conditions which can directly affect the birth/death rates of a population and the time it takes to reach an equivalent CFU. These considerations can be accommodated by considering time spent at log-growth, number of replication cycles and establishing an excess of mutants per

insertion site (Chao & Vogel, 2016; Mahmutovic et al., 2020; van Opijnen et al., 2009). In this study, the gene requirements for two different members of the MTBC are directly compared, but there are differences in media preferences and growth rates between the species, evident in the differences in number of colonies used to make the libraries. These samples likely have differences in the amount of time spent in log-growth stage and this may exaggerate more subtle effects.

3.7 CONCLUSIONS

In this chapter transposon insertion sequencing libraries made from the reference strains of animal-adapted (*M. bovis*, AF2122/97) and human-adapted (*M. tuberculosis*, H37Rv) members of the MTBC were used to determine the differences in gene requirements between the species. Comparing the total number of orthologous genes predicted as 'essential' or 'growth defect' in either species makes apparent that both genomes contain genes that, when mutated, have greater impacts on one species or the other. These uniquely-essential genes are interesting because they indicate adaptations to a specific host and can provide clues about host-pathogen interactions. The immune systems of human and animal hosts present different challenges to mycobacterial pathogens and adaptation to these differences are reflected in the relative importance of homologous gene products on survival, even in identical culture conditions. Though these genes may be genetically indistinguishable, their function and impact may be altered by deletions or SNPs in other genes in their protein-protein interaction networks. Furthermore, differences in post-transcriptional regulation of protein expression may play an unknown role in modifying the activity of the final protein product.

Genes involved in amino acid metabolism and sulfur assimilation were more required in *M. bovis* which also showed a stronger reliance on siderophore secretion for iron scavenging. Differences in the requirements for various genes involved in lipid degradation and synthesis reflect the necessity of dispensing with potentially toxic intermediates formed by utilisation of different carbon sources. As well as genes for metabolic functions that are more or less important in the different genomes, there are many possible candidates for further study, especially in the different requirements for genes involved in nitrogen assimilation (*gltBD*, *glnA2*).

These genes have potential importance in the response to oxidative and acid stress, which is an important factor in host-pathogen interactions. Further work comparing *M. tuberculosis* and *M. bovis* libraries in similar stress conditions may help to expand our understanding of the ways these species have adjusted to their host niche.

Chapter 4: Using transposon insertion sequencing to identify the gene requirements for adjustment to redox stress in *Mycobacterium bovis*

4.1 ABSTRACT

Oxidative stress is used by the host immune system to restrict the growth of pathogenic mycobacteria such as *Mycobacterium bovis*, an animal-adapted member of the MTBC which incurs serious economic costs in the UK and is responsible for zoonotic tuberculosis in countries with high TB burdens. To discover which genes are essential for survival of *M. bovis* in conditions of oxidative stress, transposon insertion sequencing libraries were grown both in the presence and absence of menadione, a menaquinone analogue known to generate reactive oxygen species (ROS) in bacterial cells. The libraries were well-saturated, with saturation levels of 63% and 53% for the untreated and treated libraries, respectively. Quantitative statistical methods identified 18 genes with statistically significant differences in the mean insertion counts between the conditions, including several oxidoreductases which have a role in combating ROS generation in diverse reactions. Cell-wall associated proteins, and those involved in cell wall integrity, had increased requirements in the menadione treated library.

4.2 AIMS

- To identify genes that are conditionally essential for *M. bovis* growth in oxidative stress
- Implement an improved primer strategy in tn-seq sequencing library creation that can identify PCR duplicates with barcodes
- Create data pipeline and custom scripts to process tn-seq reads and remove PCR duplicates before analysis

4.3 INTRODUCTION

Pathogenic mycobacteria have evolved to withstand host immune responses to infection by triggering pathways to counteract, and in some cases, exploit, the effects of the host immune system and survive in a hostile environment. Upon infection by mycobacteria, host immune cells engulf the invaders and attempt to limit growth by restricting nutrients, lowering pH, increasing levels of toxic metals and increasing redox stress. The ability of the bacteria to survive and disseminate in the host depends on flexibly adjusting its utilisation of specific gene pathways and protein products in order to respond to changing environmental conditions. For example, *M. tuberculosis* must respond metabolically to changes in carbon sources in order to maintain redox homeostasis (Pacl et al., 2018; Pawełczyk et al., 2021); and necessary metals and cofactors, such as iron, must be scavenged from the host environment in restricted conditions (Zhang et al., 2020), while simultaneously upregulating efflux pumps to avoid toxic levels of copper and zinc (Neyrolles et al., 2015).

Global phenotypic analysis, such as with transposon insertion sequencing (tn-seq), has been used to identify genes that are required for growth by bacteria in culture and infection settings (See Introduction, Chapter 3). The creation of transposon insertion sequencing libraries allows the evaluation of the fitness cost of inactivating gene insertions throughout the entire genome. By manipulating the conditions for bacterial survival by altering the culture conditions of the tn-seq library to create a selective pressure, it is possible to identify genes that are essential in a relevant physiological or pathological context (Figure 4.1), for example, with *M. tuberculosis* in media that includes cholesterol (Griffin et al., 2011), in rich versus minimal media (Minato et al., 2019) and for *M. smegmatis* in low-iron (Dragset et al., 2019). Passaging tn-seq libraries through macrophages (Rengarajan et al., 2005) or using them to infect mammalian hosts (Gibson et al., 2022; Mendum et al., 2019; Smith, C.M. et al., 2022) can uncover the virulence factors required for infection.

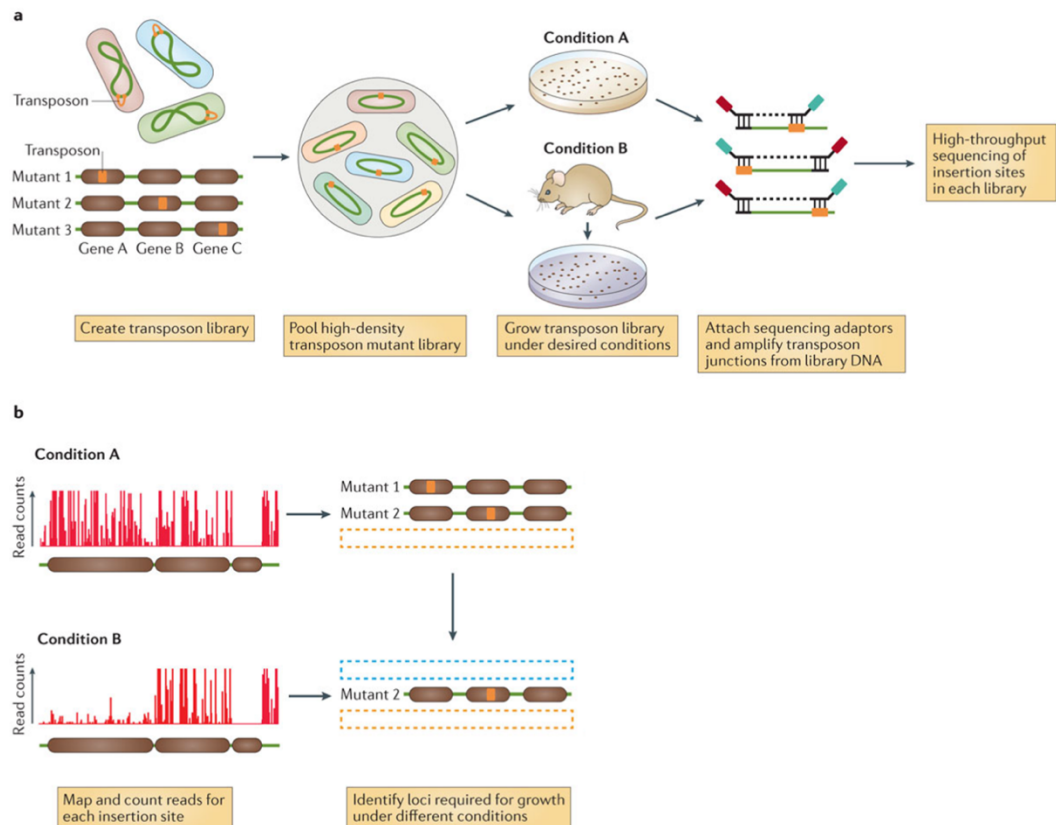


Figure 4.1. Transposon insertion sequencing analysis (tn-seq) can be used with selective pressures to identify genes that are more, or less, essential for survival under the selective condition. A) Creation of tn-seq libraries and selection on plates or in an infection model B) Using high-throughput sequencing to identify genes with fewer insertions in condition B than condition A. From (Chao et al., 2016), figure 1. Reproduced with permission from Springer Nature.

This direct phenotypic analysis differs from transcriptomic and proteomic assays that measure a transient shift in gene and protein expression in response to a change in the extracellular environment. Transcriptomic responses to stress conditions have been studied extensively for the MTBC and have revealed several large gene regulons under the control of transcriptional regulators, such as DosR, KstR and PhoP (Baker et al., 2014; Gonzalo-Asensio et al., 2008; Kendall et al., 2010; Rustad et al., 2008, 2014). However, differential expression shows minimal correlation with gene requirements and essentiality (Carey et al., 2018; Rengarajan et al., 2005). Genes may be constitutively active, and though their expression levels may not be changed in a particular stress condition, they may nevertheless be essential and/or regulated post-transcriptionally. Conversely, a gene that is differentially expressed in certain conditions may be redundant and not necessarily essential for survival. In unrestricted growth conditions, it has been found that approximately 13% of *M.*

bovis genes are essential (Chapter 3), however, additional genes will be required for growth in more challenging conditions.

A bacterial cell must maintain 'redox homeostasis' in order to use the energy from catabolising carbon sources to synthesise nucleic acids and proteins. This state requires an environment where oxidizing and reducing reactions are balanced by coupled reactions between electron acceptor/donor molecules which respond to the reduction potential of the cell, such as NAD⁺/NADH. The host immune system creates redox stress for the invading mycobacteria through changes in intracellular pH, carbon sources, nitrogen intermediates, metal and cofactor availability and oxygen levels. These must be detected and controlled by the bacteria by multiple mechanisms including regulating energy metabolism, the composition of the membrane and secretion systems.

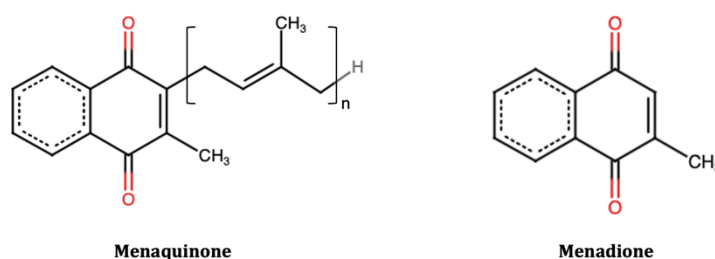


Figure 4.2. Chemical structures of menaquinone and menaquinone analogue, menadione.

Menadione is a menaquinone analogue that has been shown to have antimicrobial properties in high concentrations (Schlievert et al., 2013)(Figure 4.2). Menadione generates reactive oxygen species (ROS) as it is reduced in a single-electron reaction to semiquinone and the subsequent oxidisation releases a superoxide anion (Singh & Husain, 2018; Yao et al., 2021). An excess of ROS can lead to molecular damage of proteins, DNA and lipids in the cell. However, as the endogenous analogue, menaquinone, is not antimicrobial in excess, and menadione has been shown to be toxic for both aerobic and anaerobic bacteria, it is possible there is an additional mechanism for interference with microbial survival, perhaps interference with cell envelope or membrane integrity (Schlievert et al., 2013).

In this study, a sub-antimicrobial concentration of menadione was used to create oxidative stress on the plates used for selecting viable tn-seq mutants. Quantitative analysis was used to determine any change in the insertion frequency within genes between the two conditions to provide insight on which genes are essential for *M. bovis* to maintain redox balance when faced with oxidative stress. The *resampling* method from TRANSIT (DeJesus et al., 2015) was used to determine differences in the mean log₂ fold-changes of insertions in genes between the menadione-treated and untreated conditions (described in Chapter 3). Other methods have been used to determine if there are differences in the insertion distributions within genes, including the Mann-Whitney U-test (Santa Maria et al., 2014) and zero-inflated normal binomial regression methods (Subramaniyam et al., 2019). The Mann-Whitney U-test is a non-parametric ranked-based test that compares the distribution of the ranked reads within an ORF between two independent samples (menadione-treated and untreated) and determines if the distributions are significantly different. If two samples do not differ, the ranked insertion reads will be randomly distributed, however, if they are different, they will cluster at each end of the ranking. It is meant to be less sensitive to outliers but loses power with small sample sizes (such as with small genes with few insertion sites). The zero-inflation normal binomial regression model is especially useful when comparing multiple conditions but has been shown to be of similar effectiveness as *resampling* with a simple pairwise comparison (Subramaniyam et al., 2019).

In order to apply quantitative models to determine essentiality, the distribution of the sequenced reads at insertion sites is assumed to approximate a geometric distribution, where sites with a large number of insertions are relatively rare (DeJesus & Ioerger, 2013). Large numbers of PCR duplicates, which do not represent independent insertion mutants, can skew the distribution and invalidate this assumption, requiring more extensive normalisation and potentially biasing the output (Alkam et al., 2021; Chao et al., 2016; DeJesus & Ioerger, 2015). In previous tn-seq work in Chapter 3, there was evidence of significant PCR duplication, despite restricting the number of PCR cycles. To assess the extent of this duplication, and its impact on the distribution of insertions, an enhanced primer strategy was used that allowed the identification of the PCR duplicates before analysis, based on a strategy used by Dr. A. Smith in the lab of Dr. G. Stewart (Smith, 2017; Smith et al., 2020).

This new strategy required the creation of a new data pipeline, including scripts to identify and remove duplicates.

In this chapter, the effect of oxidative stress on the gene requirements of *M. bovis* was assayed by selection of a tn-seq library on menadione. Sequencing reads were reduced to unique template reads, eliminating PCR duplicates, by utilising molecular barcodes in the adapters. Custom bioinformatics scripts were written to create a pipeline for processing the sequencing reads, including removing PCR duplicates.

4.4 MATERIALS AND METHODS

4.4.1 Transposon Library Construction

Two independent transposon insertion libraries were created by members of the lab of Sharon Kendall at the RVC for *M. bovis*, by transduction with MycoMarT7 phage, as before (Chapter 3, (Gibson et al., 2021)). The pelleted cells were washed with PBS-Tween80™, resuspended in 10 mL PBS-Tween80™ and split into two 5mL aliquots. These were plated onto 7H11 plates + 10% OADC + Tween80™ + Kanamycin (20 µg/mL) + fetal calf serum + DMSO and +/- 50 µM menadione (10 plates for each condition). Two replicate samples of genomic DNA were extracted from each of the two independent libraries in each condition (8 samples in total) using bead-beating and enzymatic lysis as in (Gibson et al., 2021) and were used by the author for sequencing library creation.

4.4.2 Sequencing adapter and primer design

Custom adapters (Table 4.1) were designed that included the 5' Illumina flow-cell attachment P7 sequence, followed by a 8-nucleotide random molecular barcode and the Read 2 Illumina sequencing primer, as indicated in previous work by (Mendum et al., 2019; Smith, 2017; Smith et al., 2020). Adapter2 terminates with a trailing thymine nucleotide to create a sticky end to anneal to A-tailed DNA (Figure 4.3A). Adapter1 has a 3' modification to prevent extension. The random barcode will be captured in the i7 index read generated by the sequencer (Figure 4.3B) and can be used to reduce the reads to unique template counts in order to eliminate the effect of any 'jackpot' PCR amplification events.

Table 4.1. Oligonucleotides used in this chapter. Blue bases are Illumina™ read flow-cell attachment sequence. Purple bases represent random bases for PCR template barcoding. Red bases are the sample index sequence. Green bases are the Illumina™ sequencing primers.

Adapters		
Name	Full adapter sequence	Len
Adapter1	GATCGGAAGAGCACAC (5'Phos, 3'ddC)	32
Adapter2	CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	66
gDNA-junction enrichment/flowcell-binding PCR primers		
Name	Full Primer sequence	Len
P7 primer	CAAGCAGAAGACGGCATACG	20
P5_A	AATGATACGGCGACCACCGAGATCTACACATCACGTTACACTCTTCCCTACACGACGCTCTCCGATCTGTCT AGAGACCGGGGACTTATCAGC	95
P5_B	AATGATACGGCGACCACCGAGATCTACACCGATGTTACACTCTTCCCTACACGACGCTCTCCGATCTGTCT TAGAGACCGGGGACTTATCAGC	96
P5_C	AATGATACGGCGACCACCGAGATCTACACTTAGGCATACACTCTTCCCTACACGACGCTCTCCGATCTGATA GTCTAGAGACCGGGGACTTATCAGC	99
P5_D	AATGATACGGCGACCACCGAGATCTACACTGACCACTACACTCTTCCCTACACGACGCTCTCCGATCTATC TAGTCTAGAGACCGGGGACTTATCAGC	101
P5_E	AATGATACGGCGACCACCGAGATCTACACACAGTGGTACACTCTTCCCTACACGACGCTCTCCGATCTGTCT AGAGACCGGGGACTTATCAGC	95
P5_F	AATGATACGGCGACCACCGAGATCTACACGCCAATGTACACTCTTCCCTACACGACGCTCTCCGATCTGTCT TAGAGACCGGGGACTTATCAGC	96
P5_G	AATGATACGGCGACCACCGAGATCTACACAGATCTGACACTCTTCCCTACACGACGCTCTCCGATCTGATA GTCTAGAGACCGGGGACTTATCAGC	99
P5_H	AATGATACGGCGACCACCGAGATCTACACACTTGATGACACTCTTCCCTACACGACGCTCTCCGATCTATC TAGTCTAGAGACCGGGGACTTATCAGC	101
P5_I	AATGATACGGCGACCACCGAGATCTACACGATCAGCGACACTCTTCCCTACACGACGCTCTCCGATCTGTCT AGAGACCGGGGACTTATCAGC	95
P5_J	AATGATACGGCGACCACCGAGATCTACACTAGCTTGTACACTCTTCCCTACACGACGCTCTCCGATCTGATA GTCTAGAGACCGGGGACTTATCAGC	99
PCR Primers		
Name	Primer sequence	Len
KAPA_P1	AATGATACGGCGACCACCGA	20
KAPA_P2	CAAGCAGAAGACGGCATACGA	21
adapter_screen F	CAAGCAGAAGACGGCATA	18
adapter_screen R	GTGTGCTCTCCGATCT	17

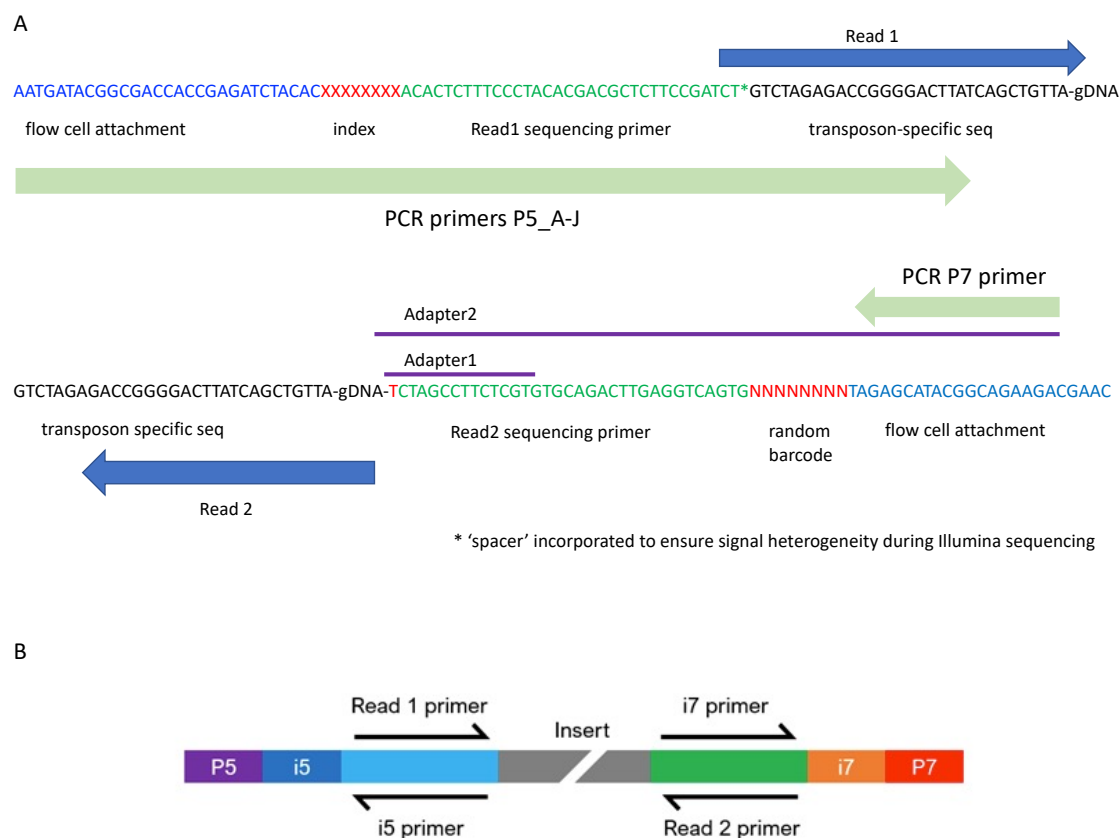


Figure 4.3. Strategy for *tn-seq* primer design based on (Smith, 2017). A) P5 PCR primers (P5_A-J) include an 8-bp index sequence to allow for de-multiplexing of samples, as well as the flow-cell attachment and Read 1 sequencing primer and is complementary to the transposon sequence. A spacer of 0-6 bp is incorporated to ensure signal heterogeneity during the calibration rounds of Illumina sequencing. The P7 PCR primer is complementary to the long adapter (adapter2). Adapter2 includes an 8-nt random oligonucleotide barcode, as well as the P7 flow cell attachment and Read 2 sequencing primer. The PCR primers are designed to amplify only gDNA fragments that contain an adapter and the transposon sequence. Figure inspired by Fig.1 in (Long et al., 2015). B) Diagram of dual-indexed sequencing reads generated by Illumina™ paired-end sequencing method. The i5 index read is used for de-multiplexing. The i7 index read will contain a random 8nt barcode. Figure reproduced with permission from Genewiz (<https://web.genewiz.com/seq-only-faq>).

Custom PCR primers were designed for the amplification and enrichment of transposon-gDNA junctions (P5_A-J, P7 primer, Table 4.1). The P5 PCR primers (Figure 4.3A) contain the P5 Illumina™ flow-cell binding sequence, followed by a library-specific 8-nucleotide index sequence, the Read 1 sequencing primer and a transposon-specific sequence. The P7 primer is complementary to the Illumina™ P7 flow-cell binding region in the adapters. PCR with these primers will amplify only those fragments of ligated DNA that include both the transposon and adapter sequences and will incorporate the P5 flow-cell binding and Read 1 primer Illumina™ sequences. The i5 index read is used by the sequencing service provider

(Genewiz/Azenta) to de-multiplex the pooled libraries. The i7 index read will contain the molecular barcode for each template sequenced.

4.4.3 Sequencing Library Preparation

Library preparation for Illumina™ sequencing was undertaken by the author as a guest in the lab of Dr. Brendan Wren at the London School of Tropical Medicine under the supervision of Dr. Ian Passmore. Library preparation was based on previous work by this lab (Chapter 3) (Gibson et al., 2021; Gibson et al., 2022) with alterations in the primer strategy to identify PCR duplicates.

4.4.3.1 Fragmentation and repair of gDNA samples

Concentrations of individual gDNA library extractions were assessed using Qubit™ fluorometer. 2 µg of each gDNA library sample was added to nuclease-free ddH₂O to a total volume of 50 µL and sonicated in a Covaris M220 Focused Ultrasonicator in order to fragment the gDNA to 550 bp target fragment size, using the following settings:

Peak incident power = 75W

Duty factor = 10%

200 cycles/burst

treatment time = 40s

Sonicated gDNA was transferred to PCR tube for one-step blunt end repair and A-tailing using the NEB Next End Prep kit (#E7645S). DNA was transferred to a PCR tube and the following reagents were added and mixed with repeated pipetting:

Enzyme Prep mix	3 µL
Reaction Buffer	7µL
fragmented DNA	50µL
	—
final volume	60 µL

Tubes were spun and incubated in thermocycler for 60 min at 20°C, 30 min at 65°C and held at 4°C. A-tailed DNA was column-purified with Monarch PCR/DNA Clean-up kit (#T1030C) was used according to manufacturer protocol except with an extra 1-minute spin and 3-minute evaporation step to ensure all ethanol had evaporated. DNA was eluted in 25µL nuclease-free ddH₂O and concentration assayed by fluorometer.

4.4.3.2 Adapter ligation

Adapter1 and adapter2 were added in equimolar concentrations to a final concentration of 10 μ M in 100 μ L nuclease-free H₂O, incubated at 95°C for 7 min in a thermocycler and then allowed to cool to RT over 2 hours to anneal and stored at 4°C. 300 ng of A-tailed gDNA is added to PCR tubes with 30 μ L NEBNext Ligation Mix, 1 μ L Ligation Enhancer, 4 μ L 10 μ M annealed adapters and nuclease-free water up to 93.5 μ L. The mix was pipetted thoroughly, and tubes were spun before incubation for 1 hour in thermocycler at 20°C. DNA was column purified as before and quantified with fluorometer. Adapter-screening PCR was performed on 1:10 dilutions of the ligated DNA (approximately 1-2ng) to confirm adapter ligation, with A-tailed, non-ligated DNA as a negative control for each sample. Using primers complementary to the adapter sequences (adapter_screen F/R), PCR was performed with Q5 High-fidelity (NEB kit #E055L) and ligation confirmed with agarose gel electrophoresis (Figure 4.4). PCR cycles were as follows: 98°C 8m; 30 cycles of 98°C 20s, 61°C 20s, 72°C 30s; 72°C 8m; hold 4°C.

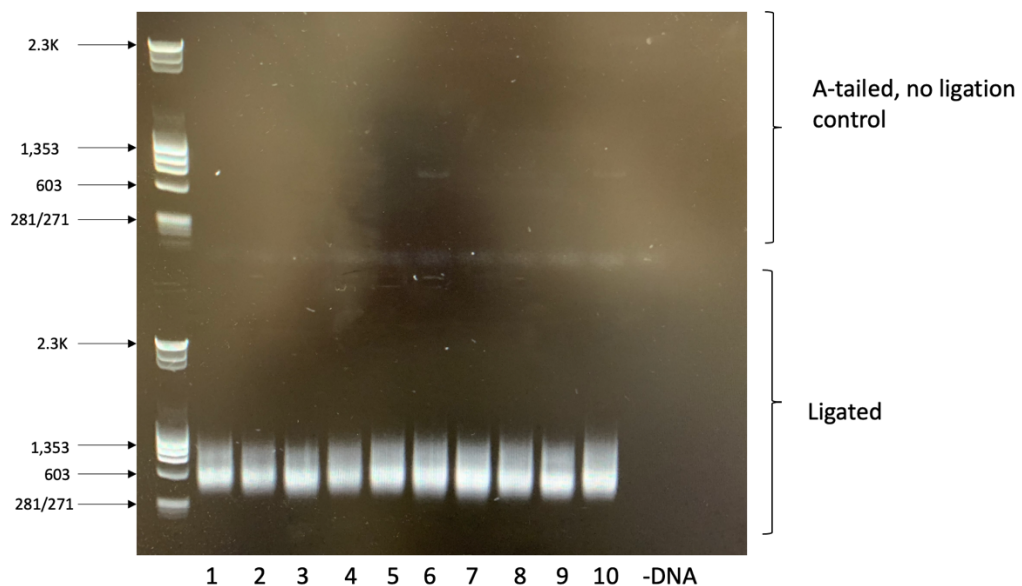


Figure 4.4. Adapter-ligation screen using adapter_screen F/R primers. In adapter-ligated samples (bottom half of gel), amplified smear visible at around 550bp. 0.8% agarose, 100V for 30m.

4.4.3.3 Transposon-junction enrichment

PCR reactions with Phusion High Fidelity DNA Polymerase were optimised for DNA concentration, primer concentration, buffer composition and annealing

temperature. Final conditions used for transposon junction enrichment were as follows:

	<u>Volume</u>	<u>Final Concentration</u>
5X Phusion HF buffer	10 μ L	1X
10 mM dNTPs	1 μ L	200 nM
5 μ M P7 primer	2 μ L	200 nM
5 μ M P5 primer (A-J)	2 μ L	200 nM
Phusion	0.3 μ L	
ligated gDNA sample	3 μ L	70-75 ng
nuclease-free dH ₂ O to 50 μ L		

Thermocycler program:

98°C 3m;

4 cycles: 98°C 20s, 70°C 20s, 72°C 1m;

20 cycles: 98°C 20s, 62°C 20s, 72°C 1m, 72°C 3m;

4°C hold

Three samples (B1, C1, and D2) had suboptimal number of transposon junctions (<0.2 nmol estimated with qPCR) and the amplification step was repeated using the remainder of the ligated DNA for these three samples under the same PCR cycling conditions. The amount of DNA in the reactions was increased to ~150-200ng. This marginally increased the yield for sample C1 but gave lower yields for B1 and D2.

4.4.3.4 Purification and library quantification

PCR products were bead-purified using Agencourt AMPureXP PCR Purification protocol with some modifications. Beads were added at 1.6x PCR reaction volume and incubated for 5 minutes off of the magnetic plate. Incubated 2 minutes back on magnetic stand and supernatant cleared and discarded. The beads were washed 3 times with fresh 70% ethanol to remove all ethanol and air-dried for 1 minute before the plate was removed from magnetic stand and the DNA eluted for 3 minutes with either 23 μ L or 30 μ L nuclease-free H₂O. The eluate was removed to new tubes and quantified with Qubit™ fluorometer.

qRT-PCR library quantification was performed with serial dilutions of 1:100 and 1:1000 of the bead-purified amplified libraries and KAPA primers (KAPA_P1 and KAPA_P2) which detect the Illumina flow-cell binding sequences (and therefore only those molecules with transposon insertions as P5 primer was complementary to

transposon sequence). The PowerUp™ SYBR Green (#A25741) system was used in an ABI 7500 Real-Time PCR System with standard cycling mode with NEBNext library quantification standards of 100pM-0.001pM (#E7642S). The reactions with the highest yield of amplified transposon junctions were pooled in relative amounts to achieve an approximation of equimolar concentration but samples B1, C1 and D2 remained underrepresented in the pool. A further round of Ampure bead purification was performed to concentrate the pooled sample, with elution in 20µL of nuclease-free H2O followed by Qubit™ fluorometer quantification and qRT-PCR library quantification. The pooled concentration was estimated at between 14-36 nM.

4.4.4 Sequencing and read processing

Paired-end sequencing was performed on the pooled libraries using the Illumina Nova-Seq platform. The samples were demultiplexed by the sequencing provider according to the index sequence in the P5 primer. Three fastq files were delivered per sample which included the Read1, Read2 and i7 index reads. Only the forward read from each pair was retained (Read 1) and the second read is discarded as it often does not include the insertion site and is unnecessary for accurate mapping. Quality checks were performed to assess read quality using bash scripts and *fastQC* (Andrews, 2010).

The sequencing reads were processed with a custom pipeline utilising Snakemake (Mölder et al., 2021) and scripts written in Bash and Python. Briefly, the pipeline is as follows (Figure 4.5): after trimming with *fastp* (Chen et al., 2018), the molecular barcode is extracted from the i7 index reads and added to the header of the corresponding Read1 reads. The reads are then searched for the 20 nucleotide-long transposon sequence tag starting within the first 22 nucleotides, and those that contain this sequence are retained and transposon tag removed (Figure 4.6). The remaining portion of the read (the gDNA fragment) is mapped using *BWA-mem* (Li, 2013) for single-end reads to the *M. bovis* genome (AF2122/97, NCBI Accession Number LT708304.1). Successfully mapped reads are de-duplicated by removing reads with identical molecular barcodes and mapped start coordinate. The number of reads mapping to each 'TA' dinucleotide in the genome is quantified and an

insertion file (in wig format) is generated for each sample. All scripts for tn-seq read processing are available on Github (https://github.com/jenjane118/tnseq_pro).

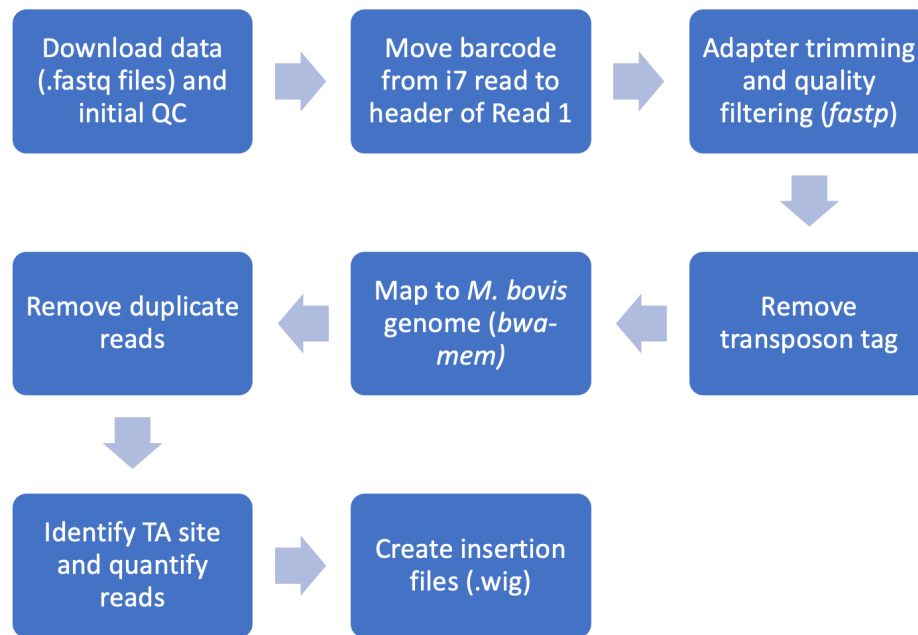


Figure 4.5. Flowchart of read processing pipeline. Process starts with de-multiplexed fastq files and ends with wig files that include count of unique template reads per 'TA' dinucleotide. Code available at https://github.com/jenjane118/tnseq_pro.

Read 1

TGTNTAGAGACCGGGGACTTATCAGCCAACCTTGTTAAGCACGCGGGTATGATAGGACCACCGGCTGGCAGG
TTGAGCCGTTTACCCAGCTTGGGCAGGTCGGCTCGGGACAGGTTGGGGTCCACCANTTGACTCCACGTCTGC
ATCACC

Read 2

CNCCAAGGGACAGCGCTGGGTGATGCAGACGTGGAGTCAAGTGGTGGACCCCAACCTGTCCCGAGCCGACCT
GCCCAAGCTGGGTGAACGGCTCAACCTGCCAGCCGGGTGGTCTATCATACCCGCGTGCTTAACAGGTTGGC
TGATAA

Figure 4.6. Representative paired-end sequencing reads after adapter trimming. Underlined bases are part of the transposon sequence that is trimmed off from Read1 during processing. Blue bases indicate 'TGTTA' sequence that indicates insertion site. Read 1 is 'forward' read and will contain transposon tag (~25-30 bp ending with 'TGTTA') followed by gDNA fragment. Read 2 is 'reverse' read and will include gDNA, sometimes followed by transposon sequence, and is not used in this analysis.

4.4.5 Data analysis

In order to apply the trimmed-total-reads (TTR) normalisation used with TRANSIT, the histograms of the insertions should resemble a geometric distribution. These

were generated for each sample using R before proceeding with data analysis (Appendix A4.1). Skew and kurtosis were tested with *moments* R package. TRANSIT 'HMM' was run on each condition summing the insertion counts for the replicates and using TTR normalisation and default settings. Prior to this, 6605 'non-permissive' sites were removed based on a sequence motif identified by (Dejesus et al., 2017) which appears non-permissive for *himar1* transposon insertions using custom scripts. TRANSIT 'resampling' was run on original insertion files (without removing non-permissive sites) to compare fold-change between the conditions using the following parameters: TTR normalisation, winsorization (to reduce effect of outliers), pseudocount = 5 (to reduce effect of high log₂-fold changes), and 100000 permutations (to resolve p-values).

Gene set enrichment analysis (GSEA) (Mootha et al., 2003; Subramanian et al., 2005) was performed using the *clusterProfiler* R package to discover whether genes with similar log₂ fold-changes after treatment were enriched for any COG (clusters of orthologous genes), GO (gene ontology) terms or KEGG pathways (Ashburner et al., 2000; Galperin et al., 2021; Kanehisa et al., 2022). There were no direct COG or GO associations available for *M. bovis* so the orthologous *M. tuberculosis* loci were retrieved from the DAVID web service as described in Chapter 3 (D. W. Huang et al., 2009b, 2009a; Jiao et al., 2012). KEGG associations for *M. bovis* were downloaded using the KEGG API (<https://www.genome.jp/kegg/rest/keggapi.html>). Analysis was performed using two different methods: with the signed-log-p-value (SLPV, the log₂ fold-change multiplied by the log of the p-value) and with the log₂ fold-change, and results were compiled.

Data handling and plots were generated in R using *dplyr* and *ggplot2* (Wickham, 2016; Wickham et al., 2022). All scripts are available at:

https://github.com/jenjane118/tnseq_pro

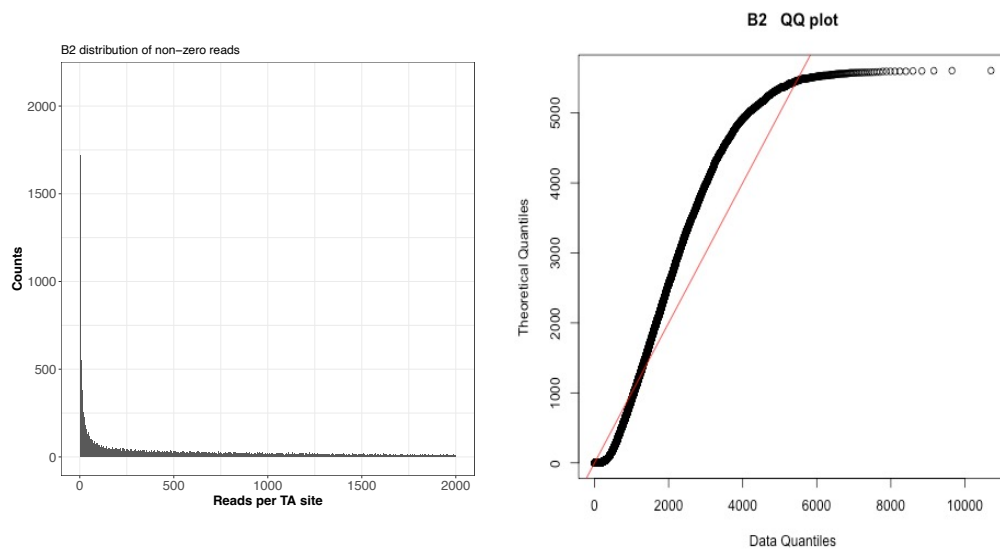
https://github.com/jenjane118/thesis_work/tree/main/Chapter_4.

4.5 RESULTS

4.5.1 Sequencing of independent libraries and technical replicates

The sequenced libraries produced between 3.3-64.8M raw reads each for each sample, with an average read length of 151 base pairs. All samples had more than 96% of reads with the correct transposon tag and mapping percentage of greater than 82%. Samples had between 2.6-16.1% duplicates (Table 4.2). Histograms of the insertion frequencies of all samples resembled a geometric distribution and distributions were free of skew (Figure 4.7).

A. Untreated



B. Treated

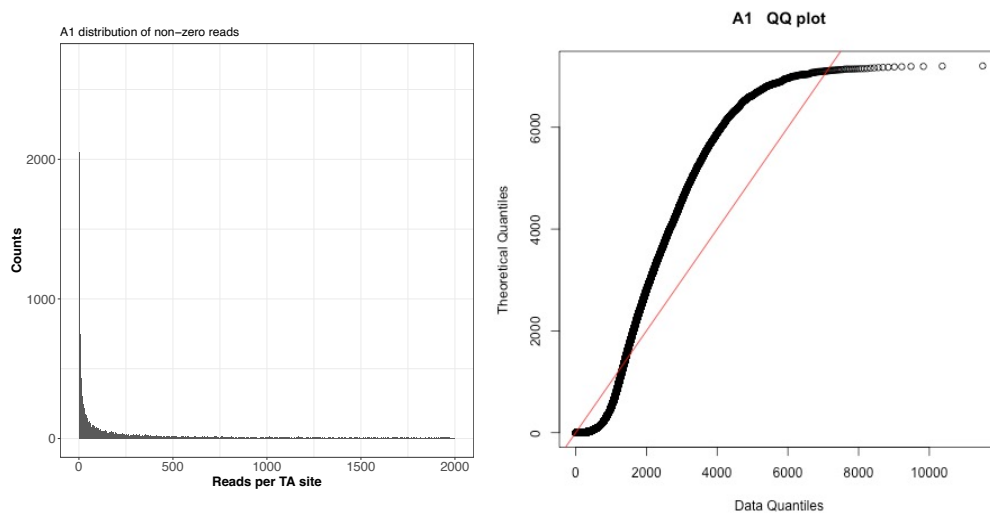


Figure 4.7. Distribution histograms and quantile-quantile plots from representative samples with similar insertion densities. Distributions resemble a geometric distribution

Table 4.2. Sequencing and processing statistics. Tagged reads are reads that contained the transposon insertion indicated by 'TGGTA' tag. Unique reads have a unique barcode and mapping coordinate combination. Non-zero mean is the mean of reads at 'TA' dinucleotide sites with at least one insertion.

Sample	Menadione	Total Reads	Tagged Reads	Mapped Reads	Unique Reads	Unique TAs hit	Insertion Density	Non-zero mean
A1	Treated	64796903	99%	62078208	84%	30384	41%	1704.5
A2	Treated	27327693	95%	23315608	91%	14864	20%	1414.2
B1	Untreated	3312969	97%	2894890	91%	12614	17%	206.7
B2	Untreated	59668345	98%	54511011	93%	35024	48%	1433.4
C1	Treated	5240906	96%	4170196	97%	23713	32%	168.7
C2	Treated	14854011	97%	12320144	95%	15555	21%	745.5
D1	Untreated	46601908	99%	44185634	92%	35024	48%	1151.3
D2	Untreated	5057770	97%	4276826	97%	25385	35%	162.9

Correlation between technical replicates, using the mean of insertions calculated for each gene, ranged from 0.7-0.96 (Spearman's rank correlation) (Figure 4.8). Lower levels of technical replicate correlation are likely a result of sampling effects that could occur either when scraping and homogenizing the cells from the library plates to extract gDNA, or when pipetting a sample of the total gDNA for sequencing. Correlation between independent experiments/libraries is not expected to be high as each library represents an independent subset of all possible insertions and dependent on the efficiency of transduction. Library 2 showed better correlation between technical replicates versus library 1.

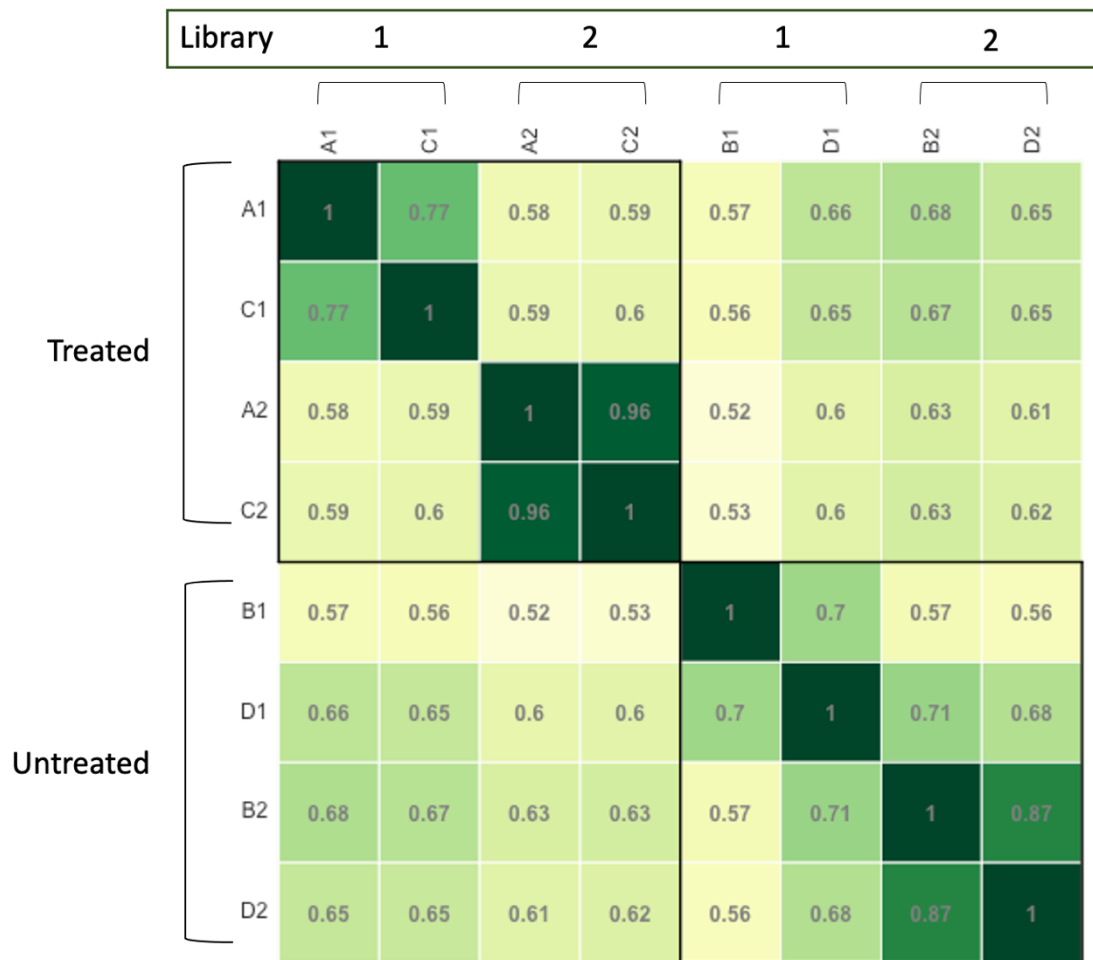


Figure 4.8. There is a higher degree of correlation between technical replicates (A and C, B and D, bracketed on top of plot) than between independent libraries (1 and 2). Top = Treated, bottom = Untreated. Degree of correlation determined comparing the mean insertion counts per gene using Spearman's correlation coefficient. All p -values $< 10e-39$. Visualised with Corrplot package in R.

There was good correlation between the total number of reads and the number of duplicates per sample⁷ ($\rho^2 = 0.93$, p -value = 0.002, Spearman's rank correlation) which shows that the number of PCR duplicates is roughly proportional to total reads (Figure 4.9A). The relationship between the number of unique TA insertions ('Unique TAs Hit' in Table 4.2 and Figure 4.9B) and number of unique (de-duplicated) reads was well-correlated in the untreated condition ($\rho^2 = 0.95$, p -value = 0.05, Spearman's rank correlation) where more reads led to a higher insertion density, but not correlated in the menadione-treated, where increasing the number of reads had minimal positive effect on the number of insertions.

⁷ A1, which had the greatest number of reads mapped (>62M) and proportion of duplicate reads (> 16%), had suboptimal amount of ligated DNA in the transposon junction amplification reactions (~25ng vs 75-100ng for other samples). After PCR amplification, sample A1 had the highest concentration of amplified transposon junctions (~2 nmol based on qPCR).

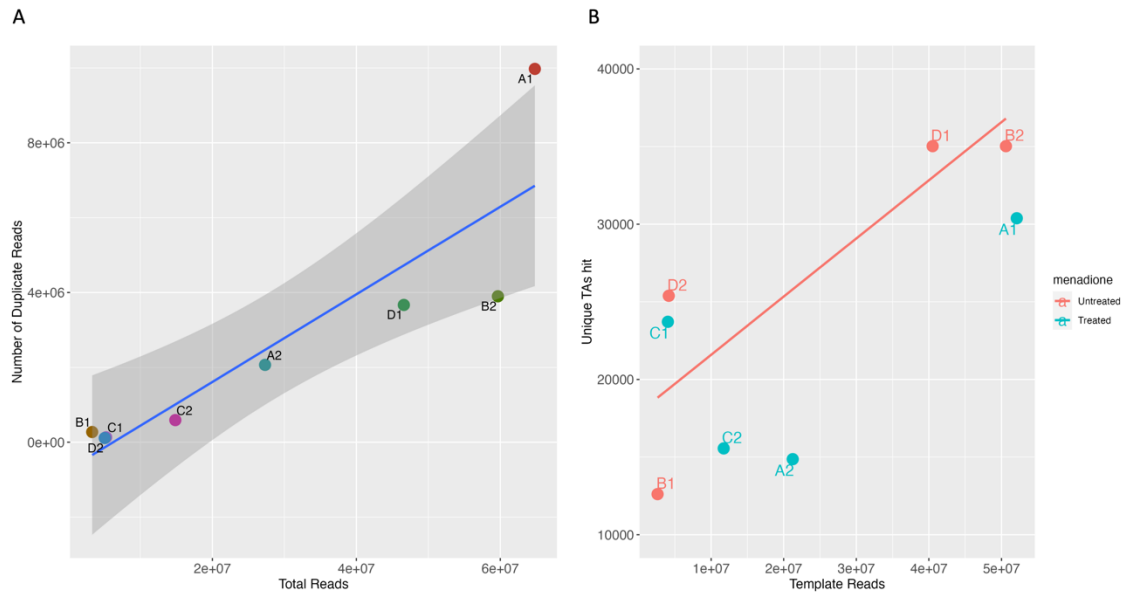


Figure 4.9. Read correlation plots. A) The number of duplicate reads is correlated to the number of total reads ($\rho^2 = 0.93$, p -value= 0.002). B) Unique TAs hit (insertion density) is positively correlated to number of template reads for untreated samples (red line, $\rho^2 = 0.95$, p -value= 0.05) but not for treated samples. Spearman's rank correlation test.

4.5.2 Calculation of the insertion density of menadione-treated and untreated samples

Summing all insertions across datasets, the control samples had a total of 3823 of 4200 genes (91%) with at least five reads mapped to insertions within the gene while the treated samples had 2.7% fewer genes with at least five reads (3720 genes). The saturation, or insertion density, of the individual samples ranged from 17% to nearly 48% (Table 4.2). Cumulative insertion density, found by counting the number of unique sites among all replicates, is 63% for the untreated and 53% for treated samples, showing less diversity of insertion sites with treatment⁸ (Figure 4.10). This could be due to biological effects, as more genes are essential for survival in a stress condition and insertions are less tolerated, or a technical bottleneck at some stage in the process where the sample size did not accurately represent the full range of mutants (M. C. Chao et al., 2016; Mahmutovic et al., 2020). Genetic drift due to menadione treatment is less likely, as previous work in the lab had shown no growth inhibition with 50 μ M menadione (unpublished results, Kendall lab).

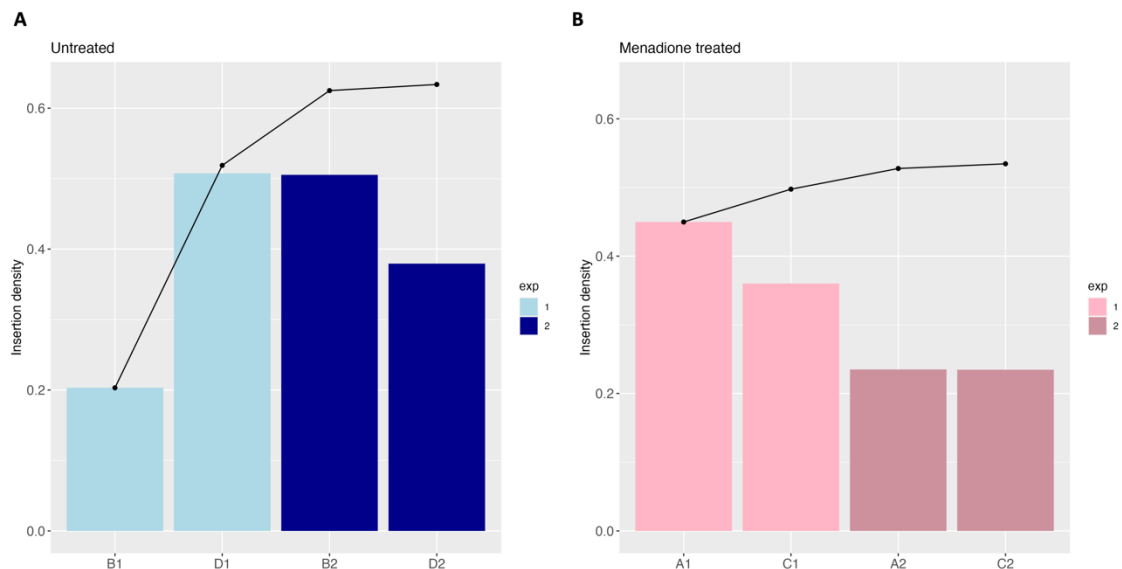


Figure 4.10. Barplots showing insertion density (number of unique 'TA' sites hit / total 'TA' sites) and the cumulative insertion density for each condition using the union of the unique sites. (A) Untreated samples have a cumulative insertion density of 63%. (B) Menadione treated samples have a cumulative insertion density of 53%.

⁸ 3829 'TA' sites had insertions *exclusively* in the treated condition compared to 11,000 'TA' sites with insertions *only* in the untreated condition (nearly 3 times as many).

4.5.3 Determination of conditional essentiality with menadione treatment

To make a quantitative determination of which genes have increased or decreased importance for *M. bovis* survival in menadione, the TRANSIT *resampling* method (DeJesus et al., 2015) was used to indicate a change in the tolerance of insertions in a particular gene (A4.1 Supplemental Tables: Ch4_Supp_Table_1). This is measured using a \log_2 fold-change in the mean number of insertions within a gene after treatment. Statistical significance is established using a permutation test on the distribution of reads along 'TA' insertion sites within the gene boundaries to calculate a p-value (as before, see Chapter 3). 18 protein-coding genes were found to have \log_2 fold-changes in menadione treatment that met the adjusted p-value cut-off (<0.05), (Figure 4.11, Table 4.3). Two of the statistically significant gene hits had positive \log_2 fold-changes, indicating that these genes had more insertions in the treated condition, implying that inactivation of these genes was advantageous for survival in menadione. The remaining 16 genes had fewer insertions sequenced in the menadione-treated samples than in the untreated--implying that inactivation leads to a survival defect. GSEA analysis (Mootha et al., 2003; Subramanian et al., 2005) was applied to the ranked \log_2 fold-changes to identify any gene groups that together had a small, but statistically significant difference in insertions between the two conditions. GO, COG and KEGG associations with *M. bovis* genes were queried but no highly ranked significant associations were discovered.

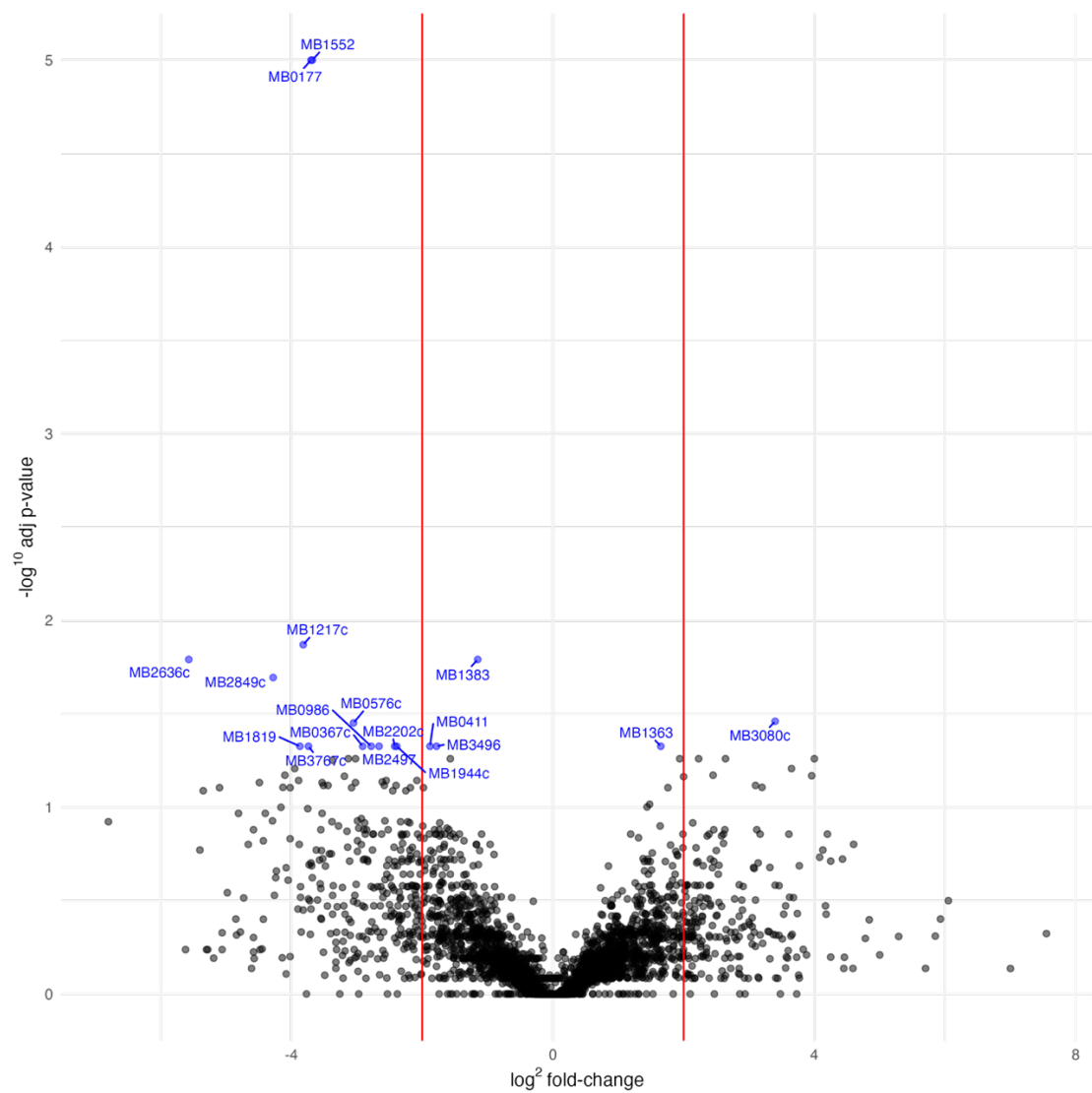


Figure 4.11. Volcano plot of resampling results comparing menadione treated and untreated *tn-seq* datasets. Genes with statistically significant ($p_{adj} < 0.05$) \log_2 fold-changes are in blue and labelled. Red bars indicate \log_2 fold-difference of ± 2 (4-fold difference). P-values corrected for multiple testing with BH method (Benjamini & Hochberg, 1995).

Table 4.3. Resampling results for genes with statistically significant \log_2 fold-changes ($p_{adj} < 0.05$). Genes with negative \log_2 fold-changes (in red) have a lower mean insertion count in the treated condition and therefore are more required for survival than in untreated. Genes with a positive \log_2 fold-change (in green) have a greater mean insertion count in the treated condition, i.e. mutations may give a selective advantage in treated condition.

Locus	<i>M.tb</i> ortholog	Functional Category	Name	Description	TA sites	Mean Untreated	Mean Treated	\log_2 Fold-change	Adj. p-value
MB0177	Rv0171	virulence, detoxification, adaptation	mce1C	MCE-FAMILY PROTEIN MCE1C	24	96.30	2.80	-3.70	0.00
MB0367c	Rv0360c	conserved hypotheticals	NA	conserved protein	8	217.10	24.50	-2.91	0.05
MB0411	Rv0404	lipid metabolism	fadD30	fadd30 (fatty-acid-amp synthetase) (fatty-acid-amp synthase)	85	37.70	6.60	-1.88	0.05
MB0576c	Rv0561c	intermediary metabolism and respiration	NA	POSSIBLE OXIDOREDUCTASE	17	139.30	12.50	-3.05	0.04
MB0986	Rv0961	cell wall and cell processes	NA	PROBABLE INTEGRAL MEMBRANE PROTEIN	5	41.10	1.70	-2.78	0.05
MB1217c	Rv1185c	lipid metabolism	fadD21	ligase fadd21 (fatty-acid-amp synthetase) (fatty-acid-amp synthase)	30	94.30	2.10	-3.82	0.01
MB1363	Rv1328	intermediary metabolism and respiration	glgP	PROBABLE GLYCOGEN PHOSPHORYLASE GLGP	42	63.50	210.30	1.65	0.05
MB1383	Rv1348	cell wall and cell processes	irta	iron-regulated transporter irta	34	6.30	0.10	-1.15	0.02
MB1552	Rv1525	virulence, detoxification, adaptation	wbbL2	POSSIBLE RHAMNOSYL TRANSFERASE WBBL2	20	64.60	0.40	-3.68	0.00
MB1819	Rv1791	PE/PPE	PE19	pe family protein pe19	6	106.20	2.60	-3.87	0.05
MB1944c	Rv1909c	regulatory proteins	furA	Ferric uptake regulation protein FurA (fur)	9	352.30	63.10	-2.39	0.05
MB2202c	Rv2180c	cell wall and cell processes	NA	integral membrane protein	11	354.60	62.10	-2.42	0.05
MB2497	Rv2470	intermediary metabolism and respiration	glbO	globin (oxygen-binding protein) glbo	9	217.30	30.10	-2.66	0.05
MB2636c	Rv2604c	intermediary metabolism and respiration	snop	probable glutamine amidotransferase snop	3	248.30	0.30	-5.57	0.02
MB2849c	Rv2825c	conserved hypotheticals	NA	CONSERVED HYPOTHETICAL PROTEIN	4	97.90	0.30	-4.28	0.02
MB3080c	Rv3054c	conserved hypotheticals	NA	NADPH:quinone oxidoreductase	6	143.30	1561.50	3.40	0.03
MB3496	Rv3467	insertion seqs and phages	NA	13E12 repeat family protein	14	192.50	52.70	-1.78	0.05
MB3767c	Rv3741c	intermediary metabolism and respiration	NA	POSSIBLE OXIDOREDUCTASE	7	223.80	12.20	-3.74	0.05

4.6 DISCUSSION

4.6.1 Menadione treated libraries have increased requirements for genes involved in cell wall integrity

Menadione has been observed to cause oxidative stress in bacteria and fungus through an excess of reactive oxidative species (ROS) and is bactericidal at high concentrations (Castro et al., 2008; Negri et al., 2023; Schlievert et al., 2013; Singh & Husain, 2018; Yao et al., 2021). It is also hypothesised to interfere with membrane permeability (Schlievert et al., 2013). The results here indicate several genes associated with cell-wall integrity had an increased requirement for survival in menadione-treated culture (Table 4.3). For example, the glycoprotein *mce1C* (Mb0177/Rv0171) is a subunit of the Mce1 complex that manages mycobacterial response to stress by translocating lipids through the cell wall envelope and maintaining the mycolic outer membrane (Chen et al., 2023; Forrellad et al., 2014; Klepp et al., 2022). It is interesting that insertions in the other subunits of the Mce1 complex were not as disadvantageous, given they together form an elongated transporter for lipids (Chen et al., 2023). In fact, insertions in subunit *mce1B* were *more* frequent in the menadione treated samples (\log_2 fold-change = 2.13) but this change was not statistically significant.

There was an increased requirement for Mb1819/Rv1791/*pe19*, which has been implicated in membrane permeability, possibly by forming 'porin' complexes along with PPE51, which appear to be required for the synthesis of the waxy phthiocerol dimycocerosate (PDIM) layer in the cell wall (Wang et al., 2020). This gene has been shown to have a role in adapting to oxidative stress in *M. tuberculosis*, where, paradoxically, overexpression of *pe19* led to higher sensitivity to membrane and oxidative stress (Ramakrishnan et al., 2016). In *M. bovis*, it was found to be required for virulence in a bovine infection tn-seq study (Gibson et al., 2022).

Genes involved in cell-wall architecture showed different requirements with menadione treatment. Mb1552/Rv1525/*wbbL2* is involved in the insertion of rhamnosyl residues in the cell wall which is integrative to its architecture (Deng et al., 2014; Grzegorzewicz et al., 2008). Insertions in Mb1363/Rv1328, which codes for a glycogen phosphorylase, GlgP, were more frequent in menadione-treated

samples, indicating mutations in this gene are advantageous for survival in menadione. The GlgP enzyme degrades α -glucan, which is the most abundant polysaccharide composing the outer capsule of the cell wall (Kalscheuer et al., 2019; Sambou et al., 2008).

4.6.2 The requirement for oxidoreductases with menadione treatment varies

Insertions in a NADPH:quinone oxidoreductase gene, Mb3080c/Rv3054c, were advantageous to survival in menadione compared to the untreated condition. In the fungal species *Aspergillus nidulans*, it was shown that a NADPH reductase was necessary for ROS generation by menadione. (Yao et al., 2021). Quinones, such as menadione, can be reduced by NADPH:quinone oxidoreductases to hydroquinone which are then auto-oxidized, releasing two superoxide molecules (Figure 4.12) (Singh & Husain, 2018). Inactivation of this pathway may reduce levels of superoxide generation. Interestingly, in a saturated *M. tuberculosis* tn-seq study, the identical ortholog was found to have a growth advantage when disrupted (Dejesus et al., 2017). In the *in vitro* tn-seq study performed by this lab comparing essentiality in *M. tuberculosis* and *M. bovis* (Chapter 3), the orthologs were both found to be non-essential in normal *in vitro* growth conditions, in agreement with other *M. tuberculosis* (Griffin et al., 2011) and *M. bovis* studies (Butler et al., 2020). The discrepant results could be a result of differences in growth media or library saturation.

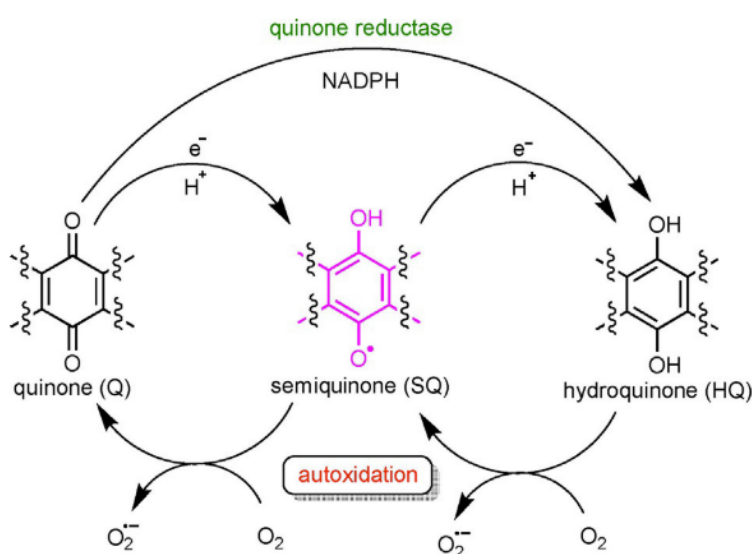


Figure 4.12. Redox cycling of quinones. From Scheme 1, (S. K. Singh & Husain, 2018), reproduced with permission from John Wiley and Sons.

In contrast, insertions in two other proposed oxidoreductases Mb0576c/Rv0561c and Mb3767c/Rv3741c caused an increased fitness cost in menadione-treated samples. Rv0561c/*menJ* catalyses the reduction of the menaquinone side chain and deletions of *menJ* in *M. tuberculosis* and *M. smegmatis* lead to significant disruption of the electron transport chain and loss of virulence (Upadhyay et al., 2015). The gene has been shown to be required for *M. tuberculosis* survival in macrophages (Rengarajan et al., 2005) and in mouse infection (Bellerose et al., 2020). Menaquinone reduction may contribute to redox balance in the cell and insertions that disrupt this gene in *M. bovis* may increase the bacteria's sensitivity to oxidative stress.

Increased requirement for some of the known pathways for combating excess ROS were not observed with menadione treatment, most notably, the redox buffer NADPH-dependent mycothiol reductase, *mtr* (Mb2880/Rv2855). This gene reduces the mycothiol disulfide (MSSM) back to mycothiol (MSH) to maintain redox homeostasis (Pacl et al., 2018). There was a reduction in the number of insertions in this gene in the treated samples (\log_2 fold-change = -1.66) but it was not statistically significant. The gene includes two conserved domains, a N-terminal cofactor-binding/pyridine nucleotide-disulphide oxidoreductase domain and a C-terminal dimerisation domain (<https://www.ebi.ac.uk/interpro/protein/UniProt/P9WHH3/>) (Mistry et al., 2021; Paysan-Lafosse et al., 2023). Looking more closely at the locations of insertions by domain, it would appear that insertions in the N-terminal domain were more deleterious for survival in menadione but tolerated in the dimerisation domain (Figure 4.13). Repeating resampling using domain regions instead of the entire ORF showed the N-terminal domain (residues 4-316) had a \log_2 fold-change of -3.63 (decrease in tolerated insertions with menadione treatment) while the dimerisation domain (residues 345-454) had an increase in insertions (+2.81 \log_2 fold-change). Neither change was statistically significant, but nevertheless indicates a difference in the domains that may be meaningful. TRANSIT HMM analysis of essentiality found the gene was non-essential in the untreated library but was essential in the treated library (A4.1 Supplemental Tables: Ch4_Supp_Table_1) and repeating this for domain-specific regions indicated the cofactor domain was essential in the treated while the dimerisation domain was non-essential. In the untreated library, the

cofactor binding domain was non-essential but insertions in the dimerisation domain caused a growth defect. MSH is necessary to resist oxidative stress in mycobacteria, but menadione has also been shown to be a 'subversive substrate' for *mtr*, generating ROS in a similar reaction to NADPH:quinone reductases (Figure 4.12) (Mahapatra et al., 2007). Perhaps there is an advantage to cofactor binding without reduction in the presence of menadione. It is possible that the domains of this enzyme may have independent functions depending on the redox state of the cell which would be an interesting avenue to explore further.

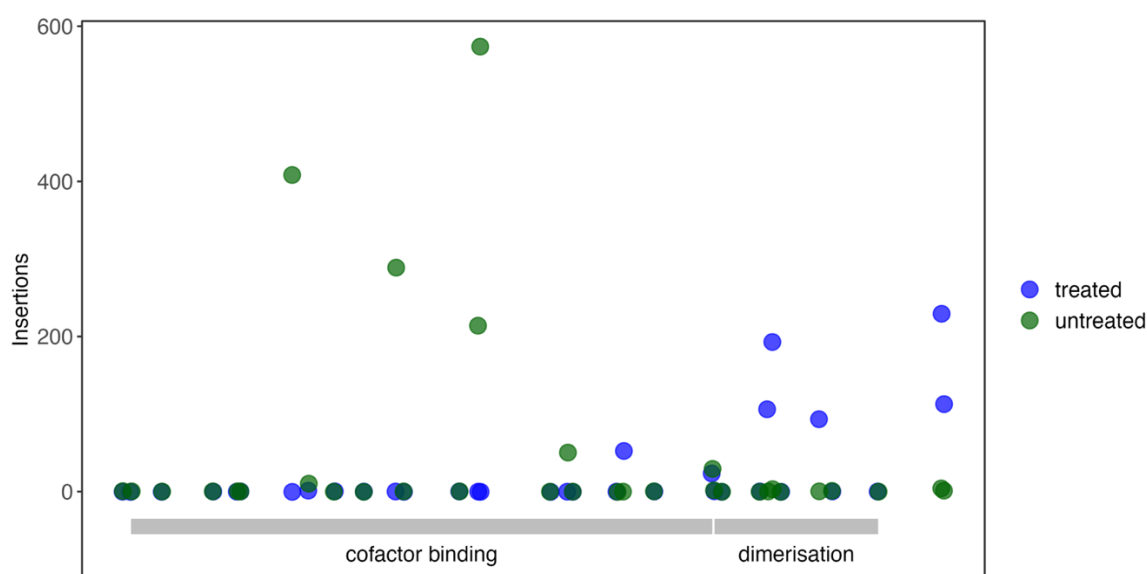


Figure 4.13. Relative position and quantity of normalised transposon insertions in the known domains of *mtr* (Mb2880). Insertions are more frequent in cofactor binding domain in the untreated condition (\log_2 fold-change -3.63, $p_{adj} = 0.13$) and in the dimerisation domain in the treated condition (\log_2 fold-change +2.81, $p_{adj}=0.50$). Each circle represents a possible 'TA' insertion site.

Another reductase, Rv2466c, a mycothiol-dependent nitroreductase, has been found to partially protect *M. tuberculosis* from menadione stress (Negri et al., 2018). A transposon mutant of the *M. bovis* ortholog (Mb2493) did not show any difference versus wild type in menadione sensitivity at concentrations of 100-500 μ M (unpublished results from Kendall laboratory). In this study, there was a modest difference in mean insertions in Mb2493c with 50 μ M menadione treatment (\log_2 fold-change -1.13), however, this was not statistically significant.

4.6.3 Oxidative stress response genes required for survival in menadione

Mb1944c/Rv1909c/*furA* codes for a transcriptional regulator that is regulated by RbpA in response to oxidative stress (Hu et al., 2016). It is transcribed with *katG*,

which codes for a hydrogen peroxide scavenger (Zahrt et al., 2001). In this study, insertions in *furA* were detrimental to fitness in menadione but were tolerated in *katG* in both the treated and untreated conditions. *furA* has also been shown to autoregulate in response to oxidative stress, and may be involved in regulation of other genes in the oxidative stress response pathway (Xin et al., 2023; Sala et al., 2003).

Regulation of fatty acid metabolism is another mechanism mycobacteria use to maintain redox homeostasis. Mutations in fatty acid synthases Mb1217c/Rv1185c/*fadD21* and Mb0411/Rv0404/*fadD30* were less tolerated in the menadione-treated samples. *fadD21* has been shown to be positively regulated by PhoP, a transcription factor activated by redox stress (Cimino et al., 2012).

Mutations in a well-conserved oxygen-binding truncated hemoglobin, Mb2497/Rv2470/*glbO* were deleterious to survival in menadione. It associates with cell membrane lipids and is thought to scavenge O₂, supporting survival in hypoxia (C. Liu et al., 2004; Pawaria et al., 2007). *glbO* promoter constructs were upregulated in oxidative stress and in metal-induced hypoxia in *M. smegmatis* and *M. tuberculosis*, and in *M. tuberculosis* infected macrophages, indicating the molecule may be involved in detoxifying ROS (Pawaria et al., 2008). This gene was markedly attenuated in a *M. bovis* tn-seq infection study, but the changes were not found to be statistically significant (Gibson et al., 2022).

Insertions in Mb1383/Rv1348/*irtA* were less tolerated in the menadione treated samples. IrtA is one subunit of a membrane-bound, iron-regulated siderophore importer. The other subunit, *irtB*, did not show a statistically significant difference in number of insertions in either condition with very few insertions sequenced in either library (A4.1 Supplemental Tables: Ch4_SupTable1). The two components are similar in their transmembrane and ABC transporter domains and are both required for normal iron uptake in mycobacteria but *irtA* contains an N-terminal siderophore interaction domain that reduces the imported siderophores (mycobactin and carboxymycobactin) through a FAD-binding domain (Ryndak et al., 2010). Insertions in the first half of the peptide, including the siderophore interaction domain, were not tolerated in either treated or untreated conditions,

however, in the ABC transporter region, insertions were tolerated in the untreated condition but not in the presence of menadione (Figure 4.14). To confirm this, resampling was repeated using the ABC transporter domain coordinates instead of the entire ORF. This region alone (residues 293-843) tolerated fewer insertions in the treated condition (\log_2 fold-change = -1.49, $p_{\text{adj}} = 0.04$) while the N-terminal portion (residues 1-267) had no insertions sequenced in either condition. Regulation of iron levels in the cell is critical for survival in oxidative stress as free iron can cause an increase in ROS (Rodriguez et al., 2002). In a previous study comparing *M. bovis* and *M. tuberculosis* tn-seq libraries in unrestricted *in vitro* growth, *irtA* was found to be non-essential for *M. bovis*, but the identical ortholog was found to be essential in *M. tuberculosis* (Chapter 3) which may suggest a difference in the sensitivity to oxidative stress or iron assimilation between the animal and human-adapted lineages.

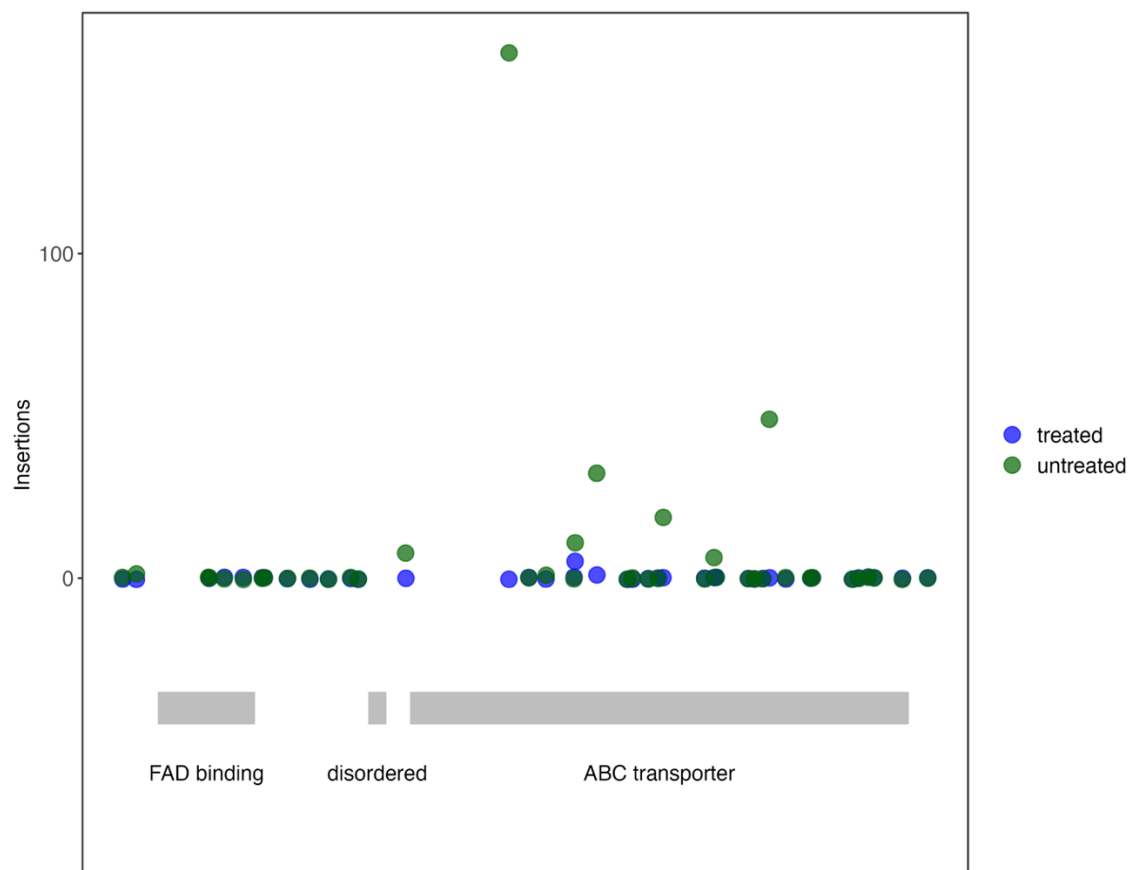


Figure 4.14. Relative position and quantity of normalised transposon insertions in the known domains of Mb1383, *irtA*. Insertions in N-terminal domains of the peptide (including the FAD binding domain) were not tolerated in either condition, however, in the treated condition, insertions in the ABC transporter domain were more deleterious to survival (\log_2 fold-change = -1.49, $p_{\text{adj}} = 0.04$). Each circle represents a possible 'TA' insertion site.

4.6.4 Limitations of the study and further work

All transposon insertion sequencing assays will be limited by the insertion density of the libraries, i.e. the number of insertions at every possible non-essential 'TA' site. By making two independent libraries, this study increased the insertion density to 63% in the untreated condition, which is somewhat short of the 83% reached by aggregating 14 different libraries in (Dejesus et al., 2017) but close to the 69.6% reached for a single library by (Patil et al., 2021). I assume the 10% loss of diversity in unique sites in the menadione treated condition is a result of an increased requirement for certain genes and gene regions in the face of oxidative stress but cannot rule out technical factors.

'Domain'-level tn-seq analysis has been investigated by several groups, based on a sliding window that looks for genomic regions within an annotated gene that have fewer than expected insertions (Dejesus et al., 2017; Patil et al., 2021; Y. J. Zhang et al., 2012). However, these studies do not consider the protein domain annotations available from sources such as the Interpro or PFAM databases (Blum et al., 2020; Mistry et al., 2021; Paysan-Lafosse et al., 2023). Work has begun in our group to incorporate protein domain annotations and streamline the analysis and visualisation of tn-seq results to improve the granularity of the analysis. Re-analysing this data with resampling using corresponding annotations for protein domains, instead of the entire ORF, may uncover more examples of domain-specific requirements and expose proteins that may 'moonlight' with additional functions in different environmental conditions.

To complement the tn-seq results presented here, differential expression analysis of menadione treated and untreated *M. bovis* with RNA-seq would be useful to evaluate transcriptomic responses to oxidative stress among the essential genes. Phenotyping deletion mutants of some of the less characterised protein candidates discussed here would be fruitful, especially to unravel the interplay between membrane integrity and oxidative stress. Four uncharacterised conserved proteins were more required in the menadione treated libraries, two of these are predicted integral membrane proteins (Mb0986/Rv0961 and Mb2202c/Rv2180) and two are conserved hypothetical proteins (Mb2849c/Rv2825c and Mb0367c/Rv0360c). The

results presented here indicate that they likely have a role in maintenance of redox homeostasis and/or membrane integrity.

4.7 CONCLUSIONS

In this study a transposon insertion sequencing library was constructed and grown with and without menadione, a menaquinone analogue that causes oxidative stress. An improved sequencing strategy was successful in allowing the removal of PCR duplicates, which can skew the distribution of reads. Mutants with insertions in genes crucial to adapting to oxidative stress were less represented in the mutant pool, and with next-generation sequencing, 16 genes were identified by the decrease in the number reads mapped within the gene coordinates compared to the untreated library pool. Two genes (*glgP*, Mb3080c) that had more insertions in the treated condition are presumed to be less essential, and mutations in these genes provide a survival advantage. Genes that known to be involved in membrane integrity (*mceC1*, *wbbL2*, *PE19*) were shown to have an increased requirement in menadione, hinting that menadione may have direct effects on the membrane. Several enzymes involved in electron transfer (oxidoreductases) were differently required with menadione treatment, as well as several genes known to be involved in redox balance such as fatty acid ligases *fadD21* and *fadD30*, *furA* and the truncated hemoglobin, *glbO*. Protein domain level analysis of the iron-regulated transporter, *irtA*, suggests that the N-terminal cofactor-binding domain may be essential in both treated and untreated conditions while the ABC-transporter domain is essential only with menadione-generated oxidative stress.

Chapter 5: Exploring the regulation of *phoR* expression by antisense RNA

5.1 ABSTRACT

Host specificity in the MTBC may involve different post-transcriptional regulation and use of antisense and other non-coding transcripts. Antisense transcription is pervasive in the MTBC and has been shown to be active in post-transcriptional regulation of protein expression. An antisense transcript found opposite the *phoR* gene in *Mycobacterium tuberculosis* was identified using computational methods with publicly-available RNA-seq data which is highly expressed in acid and stationary growth conditions. The PhoR sensor-kinase is part of a two-component system that controls the cell response to acid stress by activating the PhoP transcription factor. The system is essential for virulence in both *M. tuberculosis* and *M. bovis* despite a potentially deleterious SNP in *M. bovis phoR*. Using a CRISPR-interference system to create a *M. tuberculosis* strain with inhibited expression of the *phoR*-antisense, the effects of this inhibition on the transcriptome were evaluated. Silencing the antisense by 95% resulted in a 50% reduction in *phoR* expression but none of the genes related to the PhoP regulon were differentially expressed. The antisense transcript may be involved in regulating the translation and stability of *phoR* mRNA.

5.2 AIMS

- Validate the transcription of a predicted antisense RNA, as_*phoR*, expressed opposite the *phoR* gene in *M. tuberculosis* using RT-qPCR
- Silence the expression of the as_*phoR* transcript using CRISPRi in exponential growth conditions
- Use RNA-seq to evaluate transcriptome-wide changes in gene expression with as_*phoR* silencing

5.3 INTRODUCTION

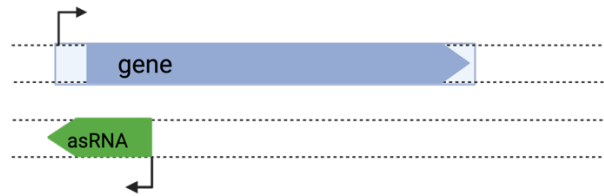
5.3.1 Antisense RNA and bacterial gene expression

With the advent of genome-wide microarray and strand-specific RNA-seq strategies, researchers began to observe an abundance of antisense transcription in the *Mycobacterium tuberculosis* complex (MTBC) and its increase in stress conditions (Arnvig et al., 2011; Dinan et al., 2014; Golby et al., 2013; Miotto et al., 2012; Pellin et al., 2012). Antisense transcripts are transcribed opposite coding genes, non-coding RNA and other annotated gene elements and are thought to act locally on adjacent gene targets (see Chapter 2). Recent studies have confirmed that 50% of coding genes in *M. tuberculosis* have been found to have an antisense TSS overlapping the coding region (Ju et al., 2024). Lineage-specific differences in antisense expression have been reported between *M. tuberculosis* and *M. bovis*, which may account for host-specific differences in gene essentiality and protein expression (Golby et al., 2013; Malone et al., 2018). However, to date, most characterisation of non-coding RNA in the MTBC has focussed on the role of sRNA--short, structured RNA transcripts that are thought to bind and regulate distant mRNA transcripts. Potentially, this was fuelled by a bias from model bacterial systems where chaperone proteins, not expressed in Mycobacteria, moderate RNA-RNA interactions and make it easier to identify and characterise them. The sheer pervasiveness of antisense transcription across the bacterial genome has also been a hurdle to functional characterisation and has led to debates of its biological relevance in the MTBC and in other bacteria (Adams et al., 2020; Dinan et al., 2014; Georg & Hess, 2018; Lloréns-Rico et al., 2016; Lybecker et al., 2014).

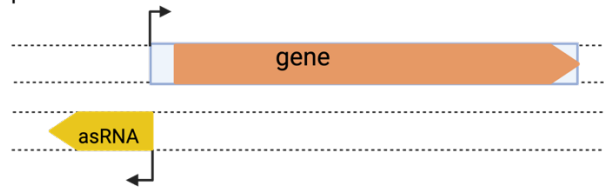
Despite these challenges, an increasing number of bacterial antisense RNA have been characterised and found to regulate gene and protein expression. Regulation of transcripts via antisense transcripts can function through several mechanisms, reviewed in (Georg & Hess, 2018; Lejars et al., 2019; Sesto et al., 2013), (Figure 5.1). Transcriptional regulation by antisense RNA can involve transcriptional interference, including competition for bi-directional promoters and transcriptional termination (Ju et al., 2019; Lybecker et al., 2014; Warman et al., 2021) and post-transcriptional mechanisms, such as obstructing regulatory binding sites for ribosomes and sRNAs, or by creating or masking RNase sites (Aiso et al., 2014; Lei et al., 2018; Lejars et al., 2022; Morra et al., 2023). Antisense transcripts located

internal to genes can function to isolate or 'decouple' the transcription or translation of individual genes in an operon (Dawson et al., 2022; DeLoughery et al., 2018).

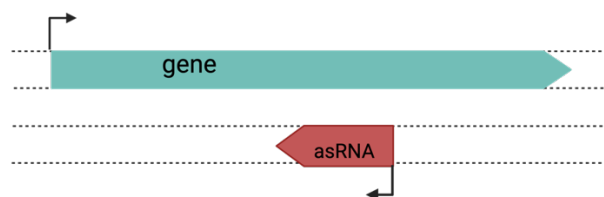
Head-to-head: inhibition of RBS, transcription termination, blocking RNase sites, generating specific RNase sites



Diverging: transcriptional competition at bi-directional promoters



Internal antisense RNAs: co-degradation, generation of specific RNase sites



Intra-operon antisense RNAs: transcription termination (transcriptional decoupling), generation of specific RNase sites (post-transcriptional decoupling), inhibition/stabilisation of TIS (post-transcriptional decoupling)

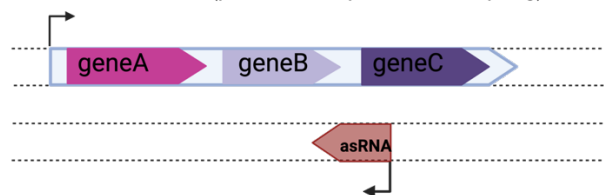


Figure 5.1. Models of regulation by antisense RNAs in bacteria. Arrows show direction of transcription. Figure adapted from Figure 1, (Georg & Hess, 2018) and created with BioRender.com.

5.3.2 The PhoPR two-component system

This chapter investigates an antisense RNA expressed on the non-coding strand opposite PhoR, part of an important two-component system (TCS) and virulence factor of the MTBC. PhoPR is a well-studied two-component system of the MTBC that is involved in response to acid stress, hypoxia and other stress conditions. Two-component systems include a sensor kinase, which responds to an environmental stimulus or condition, and an effector molecule which acts as a transcriptional

regulator on a host of downstream gene targets. There are 12 known TCS in pathogenic mycobacteria and these are implicated in various virulence systems. There is evidence for 'cross-talk' between the sensors and transcriptional regulators of the different TCSs and heterodimers of transcriptional regulators from different systems (Agrawal et al., 2015; Stupar et al., 2022; Vashist et al., 2018).

The PhoR sensor kinase component is a homodimeric transmembrane protein with an N-terminal extra-cytoplasmic sensing domain, a histidine kinase domain, dimerisation domain, and cytoplasmic ATP-binding and catalytic domains (Figure 5.2). The N-terminal sensory domain is composed of an external loop of about 120 residues flanked by 2 transmembrane helices. The sensor domain is stimulated by an external signal and undergoes a conformational change, transferring the signal through the membrane and activating the cytosolic flexible dimerisation/kinase domain. A conserved histidine is autophosphorylated, which in turn phosphorylates the regulatory transcription factor, PhoP. The exact sensor stimulus for PhoR is unknown, though the system has been shown to be responsive to changes in redox potential including acid stress in both *M. tuberculosis* (PhoPR_{Mtb}) and *M. bovis* (PhoPR_{Mb}) (Baker et al., 2014; Bansal et al., 2017; Feng et al., 2018; García et al., 2021; Goar et al., 2022) and hypoxia in PhoPR_{Mtb} (Singh et al., 2020).

PhoR_{Mb} contains a SNP, relative to PhoR_{Mtb}, that has been linked to lower virulence in macrophage and mouse infection models (Gonzalo-Asensio et al., 2014) (Figure 5.2). However, survival of *M. bovis* in bovine macrophages is dependent on PhoP_{Mb} (García et al., 2018) and a tn-seq study of *M. bovis* survival in the bovine host, showed that inactivating transposon insertions were attenuating in both *phoP* and *phoR* (log₂ fold-change < -7.1 for both genes) (Gibson et al., 2022).

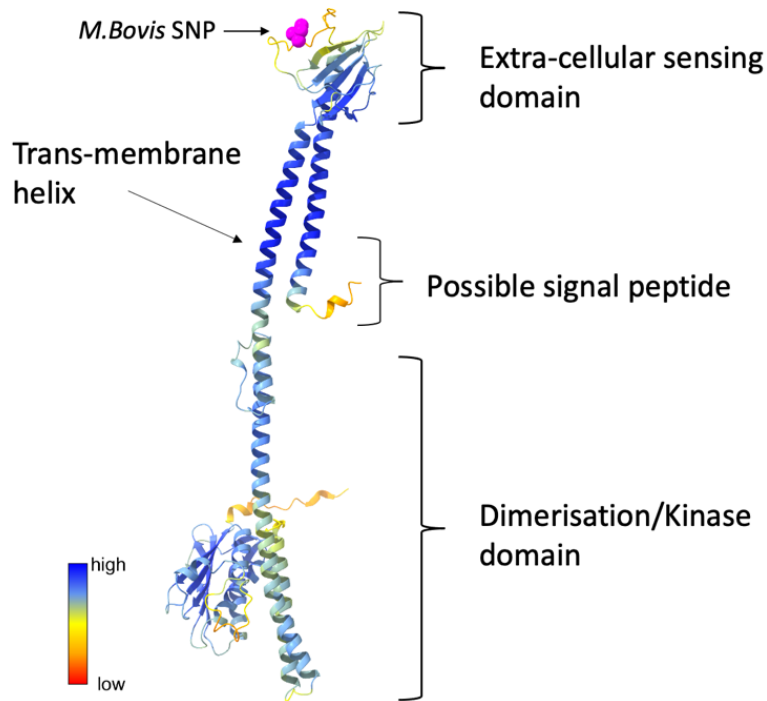


Figure 5.2. AlphaFold prediction of monomer of *PhoP_{Mtb}* with predicted domains (InterPro/UniProtKB P71815_MYCTU) (Blum et al., 2020; Jumper et al., 2021; Meng et al., 2023; Paysan-Lafosse et al., 2023; Varadi et al., 2022; Xing et al., 2017). Legend reflects AlphaFold per-residue model confidence score (pLDDT). The *M. bovis* SNP at codon 71 (G to I) is indicated by magenta balls. Figure created with ChimeraX (Meng et al., 2023).

PhoP_{Mtb} activates a large gene regulon by binding of the C-terminal DNA-binding domain to upstream promoter regions (Figure 5.3). Phosphorylation of *PhoP_{Mtb}* by *PhoR* increases binding strength, but not specificity, and is not strictly necessary for binding to target genes (He & Wang, 2014; Xing et al., 2017). Over 30 genes are directly regulated by *PhoP_{Mtb}* through binding, but knockouts have shown up to 2% of the genome differentially expressed in Δ *phoP_{Mtb}* mutants, with lipid metabolism, central carbon metabolism and PE/PPE/PE_PGRS genes overrepresented (Cimino et al., 2012; Goar et al., 2022; He & Wang, 2014; Ryndak et al., 2008; Solans et al., 2014; Walters et al., 2006). Regulators involved in stress-response and regulation of redox homeostasis such as transcription factors, *WhiB3* and *DosR*, and alternative sigma factors, *SigE* and *SigH*, are targets of *PhoP_{Mtb}* and they regulate many downstream targets (Figure 5.3). *PhoP_{Mtb}* knockout mutants have cell envelope defects and are more sensitive to oxidative stress and low pH (Bansal et al., 2017; Goar et al., 2022; Khan et al., 2024; Walters et al., 2006). In *M. bovis*, *PhoP_{Mb}* has been

shown to regulate genes involved in ammonia production in response to acidic pH (García et al., 2018, 2021).

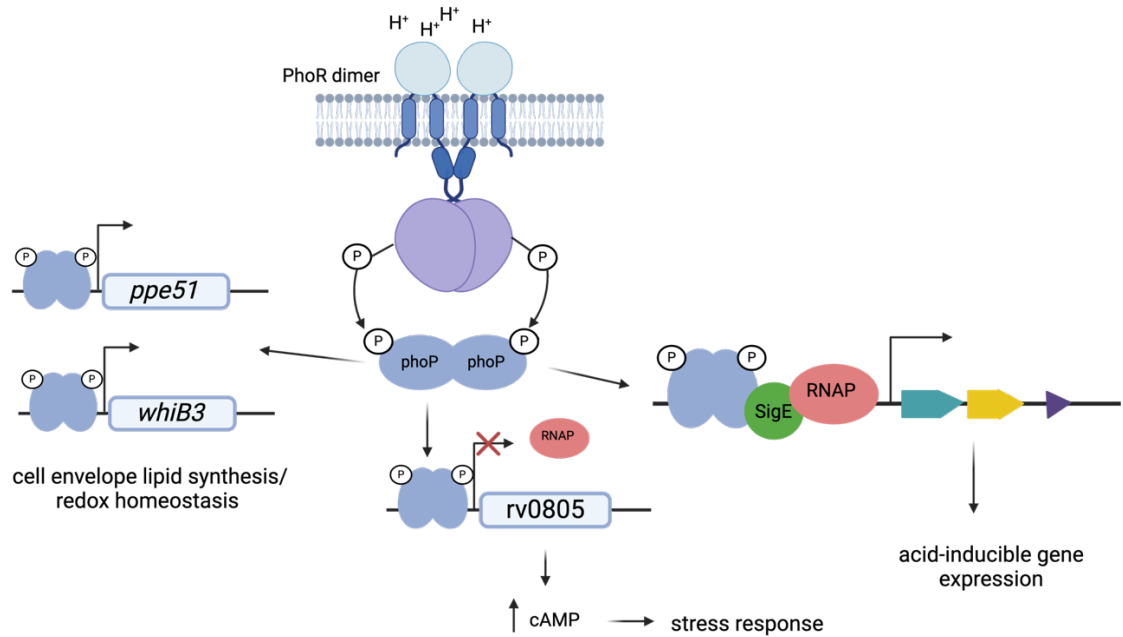


Figure 5.3. Several transcriptional regulators involved in acid and stress responses are regulated by PhoP. Figure adapted from (Baker et al., 2019; Bansal et al., 2017; Dechow et al., 2021; Feng et al., 2018; Khan et al., 2024) and created with BioRender.com.

5.3.3 An antisense RNA transcribed opposite phoR

Previous work (Chapter 2) identified an expressed antisense RNA transcript (as-phoR) opposite the coding gene for *phoR* in *M. tuberculosis*. This transcript was upregulated in certain conditions, including stationary growth and low pH and coexpressed with PE/PPE genes, recognised as virulence factors (De Maio et al., 2020; Wang et al., 2020). Transcription of as-phoR begins within the *phoR* ORF and overlaps the location of the SNP in PhoR_{Mb}. The existence of this antisense transcript, expressed differentially across many culture conditions, hints at a more complex regulation of the gene operon than has been explored in the literature. As-phoR could be involved in regulating *phoR* expression and perhaps plays a role in the independent regulation of genes in the *phoPR* operon (Figure 5.1, 'intra-operon antisense') through transcriptional interference or post-transcriptional pathways, such as regulating transcript stability. Furthermore, expression of as-phoR could differ between *M. tuberculosis* and *M. bovis* and contribute to lineage-specific differences in the regulation of lipid biosynthesis and transport. I propose to investigate how as-phoR, specifically, may be involved in transcriptional or post-

transcriptional gene regulation of the PhoPR TCS in *M. tuberculosis* and, more generally, gain insight on the role of antisense regulation in the MTBC.

5.3.4 CRISPR inhibition as a strategy to silence antisense RNA

CRISPR inhibition is a technique that harnesses the technology of CRISPR gene editing to silence expression of a targeted RNA transcript by directing a catalytically-inactive Cas9 enzyme (dCas9) to a sequence-specific region of the genome where it sterically hinders the transcription of the targeted gene region (Figure 5.4). It has been shown to be positively dependent on the specific strand targeted and by the position of the targeted sequence within the transcript, and has been successfully applied to silence genes in *M. tuberculosis* (Choudhary et al., 2015; Qi et al., 2013; Rock et al., 2017; Singh et al., 2016). It is therefore theoretically possible to silence an antisense transcript independently of the sense expression; however, it has not yet been widely used to silence bacterial antisense RNA. In this study, CRISPRi strains were created that target as-phoR, an antisense transcript located opposite the 5' end of *phoR* to establish the feasibility of this approach and to evaluate the transcriptomic effects of antisense-phoR silencing on the *M. tuberculosis* transcriptome.

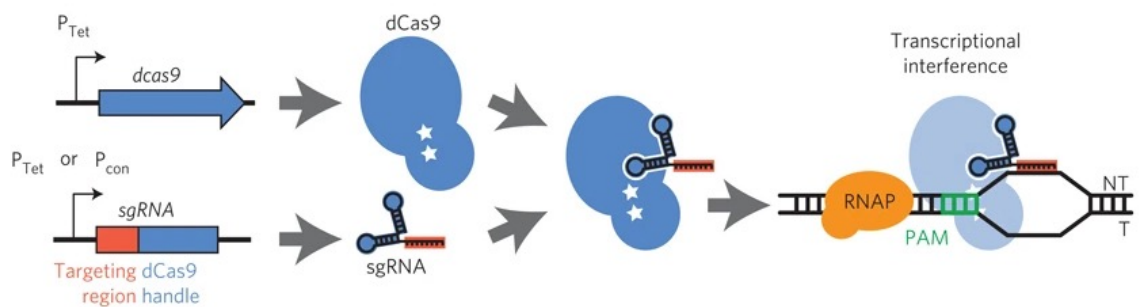


Figure 5.4. Schematic diagram of CRISPR-interference system. dCas9 and the sgRNA are expressed from Tet-inducible promoters and bind due to a dCas9 'handle' on the sgRNA, forming the sgRNA:dCas9 complex. The complex binds at the programmed sequence on the non-template strand of the target transcript, sterically inhibiting target transcription. A PAM (proto-spacer adjacent motif) is required for dCas9 binding. Figure from (Rock et al., 2017). Reproduced with permission from Springer Nature.

5.4 MATERIALS AND METHODS

5.4.1 RT-qPCR

RT-qPCR was used to quantify expression of *as-phoR* in *M. tuberculosis*. 100 ng DNase I (Invitrogen) treated RNA was reverse transcribed with Superscript III (Invitrogen) with 300 ng/μL random primers and 10 mM dNTPs to 13 μL total volume. RNA samples were denatured at 65°C for 5m followed by decreasing temperature to 4°C in the thermocycler. After brief centrifugation, 5X RT buffer, 1 μL 0.1M DTT, 1 μL RNaseOUT (Thermo Fisher Scientific) and 1 μL Superscript III were added to sample tubes (in addition to -RT control with no Superscript III). After careful mixing with pipette, reactions were incubated at 25°C for 5m for primer annealing, followed by first strand synthesis at 55°C for 50m and 70°C for 15m to terminate the reaction.

Serial dilutions of *M. tuberculosis* gDNA were used with primers to *as-phoR* and housekeeping gene (*sigA*) to create a standard curve of 10³ to 10⁷ copies. Reactions included 1μL gDNA standard or cDNA sample and 0.3 μM forward and reverse primers with PowerUp SYBR Green Master Mix (Thermo Fisher Scientific) in 20 μL total volume. RT-qPCR was run using a BioRad CFX96 Maestro analyser at 50 °C for 2m, 95 °C for 2m, followed by 40 cycles: 95°C for 15s, 60°C for 15s, 72°C for 1m and 85°C for 5s. Melt curve analysis was carried out after each run (65°C – 95°C in increments of 0.5°C, 5s each cycle). PCRs were run in duplicate or triplicate for each sample and gDNA dilution. Direct quantification was made using gDNA standard curves for each primer pair and relative quantification as a proportion of the housekeeping gene, *sigA*.

5.4.2 Design of sgRNAs to target antisense-phoR

Small guide RNAs (sgRNAs) were designed to target the non-template DNA strand as close as possible to the transcriptional start site of the antisense transcript according to published protocols (Larson et al., 2013). The sgRNAs are designed by searching the 5' end of the antisense transcript for the typical proto-spacer adjacent motifs (PAM site) specific for the *Streptococcus pyogenes* (*SPy*) CRISPR system (Singh et al., 2016) (in this case 'CCN') in the sequence of the antisense transcript (5' to 3') and including the following 20-25 nucleotides. The final 12 nucleotides of the reverse-complement of this target sequence (plus two nucleotides of the PAM

sequence) make up the seed region which is primarily responsible for target specificity. This sequence, plus the terminal two nucleotides of the PAM sequence was used with BLAST (Sayers et al., 2022) to identify any possible off-target hits in both sense and antisense strands of the *M. tuberculosis* genome proximal to the PAM site. Any sgRNAs with full-length hits other than the target transcript are not taken forward. The full length transcribed sgRNA plus the dCas9 handle and terminator sequence are used for secondary structure predictions using M-fold (Zuker, 2003) to rule out misfolding of the handle hairpin and formation of loops in the target sequence. Finally, the sequence of the sgRNA was checked to make sure there are no BbsI restriction enzyme sites which are necessary for cloning into the pRH2521 plasmid. 3 sgRNAs (sgASrna1, sgASrna2, sgASrna3) were taken forward, and forward and reverse oligos were synthesised including 4 added nucleotides for ligation at BbsI site in pRH2521 (Table 5.1).

Table 5.1. List of oligonucleotides used in this study

RT-qPCR Oligonucleotides		
Target	Forward Primer	Reverse Primer
<i>sigA</i>	CCTACGCTACGTGGTGGATT	TGGATTTCAGCACCTTCTC
<i>asrna</i>	GGCTGATCACCCGAACGTAGA	CTGGACTTGTGGCCTCGGGG
<i>phoR</i>	TTCGAGGAAGGCTGCCCC	GATCCCCGAGGCCACAAGTC
pRH2502-dCas9	AAGAAGTACAGCATCGGCCTGG	TTCTTGCGCCGCGTGTATCG
pRH2521-sgRNA	AATATTGGATCGTCGGCACC	TTGGAGAAGCAGCTGAAGTG
CRISPRi silencing sgRNA oligos (underlined bases added for ligation)		
Target	Forward Primer	Reverse Primer
sgASrna1	<u>GGG</u> ATCAACGACAACACTGCCATAC	<u>AAAC</u> GTATGGCAGTGTGTCGTTGA
sgASrna2	<u>GGG</u> ACCCCGACGGCCAGAGCTATA	<u>AAAC</u> TATAGCTCTGGCCGTCGGGG
sgASrna3	<u>GGG</u> AGTTCGGGTGATCAGCCCCGA	<u>AAAC</u> TCGGGGCTGATCACCCGAAC

5.4.3 Cloning of sgRNAs into pRH2521 plasmid

The synthesised oligo pairs were annealed at 95°C for 5m and allowed to return to RT on the bench. Annealed primers were phosphorylated with T4 PNK (New England Biolabs) in 10X reaction buffer A and 10mM ATP in final volume of 20 µL with incubations at 37° for 20m and 75°C for 10m. 1 µg pRH2521 (Table 5.2, Figure 5.5) was digested with 2 units BbsI (New England Biolabs) + 10X buffer G in 20 µL total volume at 37°C for 4h followed by enzyme inactivation at 65°C for 20m, and put on ice. Annealed and phosphorylated oligos were ligated at 16°C overnight into BbsI digested pRH2521. A ratio of 1:1 oligo:plasmid DNA was used in the ligation reactions: 100ng BbsI-digested pRH2521 + 25pmol annealed/phosphorylated oligonucleotides + 10X ligation buffer and 5units T4 DNA ligase (New England Biolabs). The ligation mixtures were transformed into *E. coli* DH5α competent cells (Table 5.2) by heat shock at 43°C for 30s followed by incubation on ice for 5m. Cells were recovered in 950µL warm LB broth and incubated in orbital shaker at 37°C with 200 rpm shaking for 1h. Successful transformants were selected on hygromycin-containing LB plates (200 µg/mL). Multiple colonies from each ligation were picked and grown in 10mL liquid LB + 200 µg/mL hygromycin. Plasmid DNA was purified with Monarch Plasmid Miniprep kit (New England Biolabs) and sent for sequencing with the corresponding forward sgRNA primer to confirm presence of insert. Positive clones were used to re-transform DH5α competent cells for creation of glycerol stocks for long-term storage at -80°C.

Table 5.2. List of bacterial strains and plasmids used in this study.

STRAINS	GENOTYPE	SOURCE
<i>E.Coli</i> DH5α	SupE44 1lacU169 (lacZ1M15) hsdR17 recA1 endA1 gyrA96 thi-1 relA1	New England Biolabs
Mtb _{dCas9}	<i>M.tuberculosis</i> H37Rv with integrative plasmid containing <i>dCas9_{SPy}</i> (pRH2502), kan ^R	Gibson et al, 2021
Mtb _{dCas9_ctrl}	<i>M.tuberculosis</i> H37Rv with pRH2502 and sgRNA -ve control plasmid (pRH2521), kan ^R , hyg ^R	Gibson et al, 2021
Mtb _{dCas9_sgRNA1}	<i>M.tuberculosis</i> H37Rv with pRH2502 and sgRNA-containing plasmid (pASphoR_1), kan ^R , hyg ^R	this study
Mtb _{dCas9_sgRNA2}	<i>M.tuberculosis</i> H37Rv with pRH2502 and sgRNA-containing plasmid, (pASphoR_2), kan ^R , hyg ^R	this study
Mtb _{dCas9_sgRNA3}	<i>M.tuberculosis</i> H37Rv with pRH2502 and sgRNA-containing plasmid, (pASphoR_3), kan ^R , hyg ^R	this study
PLASMIDS		
pRH2502	Integrative plasmid derived from pTC-0X- 1L, expressing <i>dCas9_{SPy}</i> from an inducible tetRO promoter (uv15tetO), kan ^R	pRH2502 was a gift from Robert Husson (Addgene plasmid # 84379)
pRH2521	Non-integrative plasmid derived from pTE-10M-0X, expression sgRNA from inducible tetRO promoter (Pmyc1tetO), hyg ^R	Singh et al, 2016
pASphoR_1	pRH2521 with sgRNA ASrna1, hyg ^R	this study
pASphoR_2	pRH2521 with sgRNA ASrna2, hyg ^R	this study
pASphoR_3	pRH2521 with sgRNA ASrna3, hyg ^R	this study

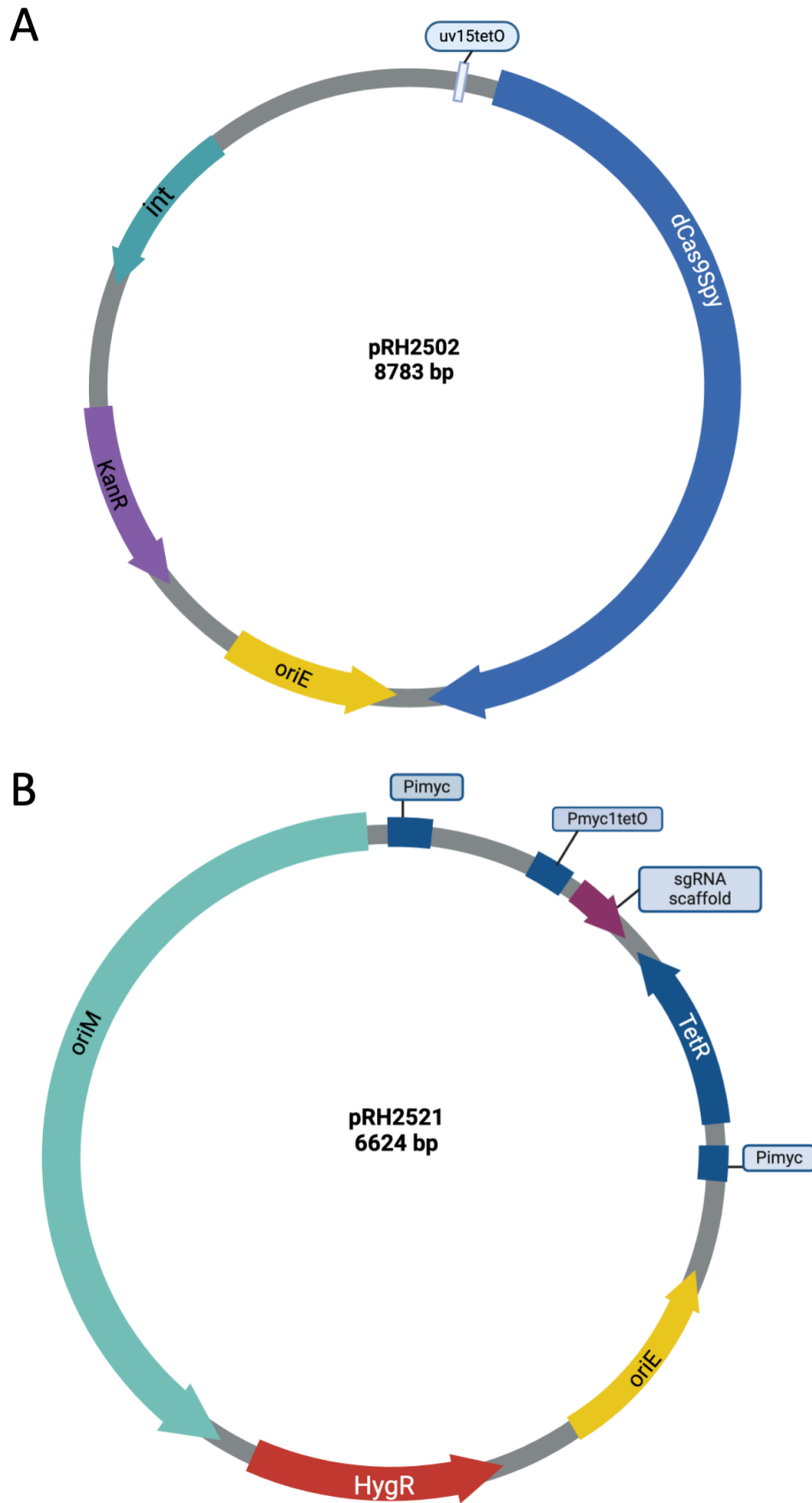


Figure 5.5. CRISPRi two-plasmid system (A. K. Singh et al., 2016) used in this study involves two plasmids: A) integrative plasmid (pRH2502) which expresses dCas9_{Spy}, and B) episomal plasmid (pRH2521) which expresses the sgRNA scaffold, both from Tet-inducible promoters. Figure made with Biorender.com

5.4.4 Transformation into *M. tuberculosis* and Induction of CRISPRi system

All work in the containment level 3 (CL-3) laboratory was performed by members of the Kendall lab. Plasmids pASphoR_1-3 and pRH2521 were transformed into *M. tuberculosis* strain, Mtb_{dCas9}, with electroporation. Recovered cells were plated onto large (140mm) Middlebrook 7H11 agar plates supplemented with hygromycin and kanamycin and incubated for 5-6 weeks at 37°C. Single colonies were selected and streaked on to 7H11 agar plates with antibiotics and incubated at 37°C for 4 weeks.

For colony PCR, 15 µL of heat-killed, clarified cell lysate from each transformed strain (Mtb_{dCas9}_sgRNA1-3, Mtb_{dCas9}_ctrl) was used immediately in PCR reactions with 10 µM primers: pRH2502-dCas9 for the integrated pRH2502 plasmid and pRH2521-sgRNA for the non-integrating pRH2521 plasmid. OneTaq[®] 2X master mix (New England Biolab) was used for PCR: 94°C for 1m; followed by 30 cycles of: 94°C for 30s, 55°C for 1m, 68°C for 1m; followed by final extension of 68°C for 5m. PCR products were visualised on 0.8% agarose gel with 1:10000 dilution SYBR-SAFE (Thermo Fisher Scientific) and Quickload[®] Purple DNA 100 bp Ladder (New England Biolab).

For initial CRISPRi strain evaluation, PCR-confirmed colonies from each strain (Table 5.2) were grown in CL-3 laboratory conditions by members of the Kendall lab to OD 0.5 in 10mL cultures in Middlebrook 7H9-ADC broth supplemented with 0.2% glycerol, 0.05% Tween 80 and hygromycin, in 490 cm² Corning Roller Bottles (Sigma Aldrich) rolling at 2rpm at 37°C. The cultures were divided, and the 'ATc+ve' group was treated with 200 ng/mL anhydrotetracycline (ATc) for 24 hours to induce expression of dCas9 and the sgRNA from the Tet-responsive promoters (Figure 5.6). Initial experiments were performed in duplicate. After initial RT-qPCR analysis, the engineered strain with the largest apparent knockdown of the antisense transcript, Mtb_{dCas9}_sgRNA2, was taken forward and the experiment repeated in triplicate with 30 mL cultures.

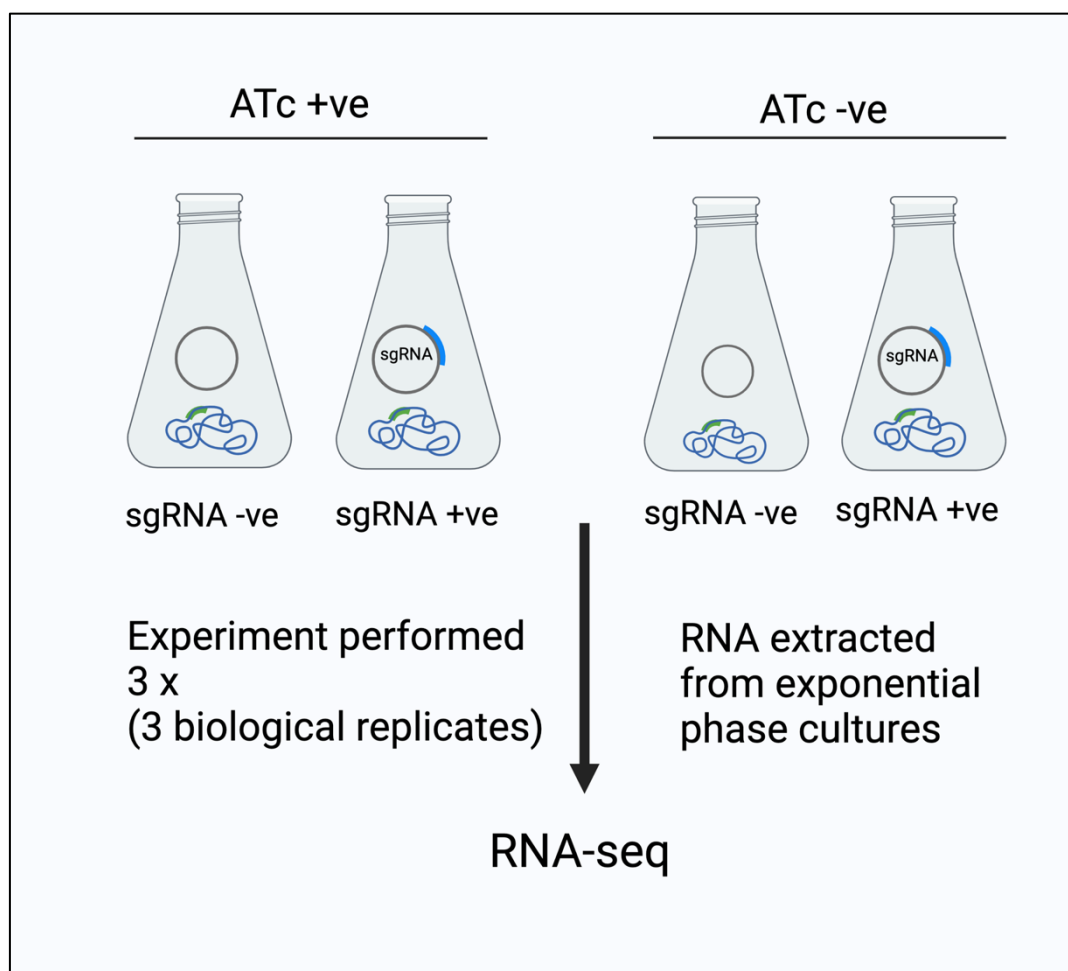


Figure 5.6. Design of CRISPR-inhibition experiment. Parallel cultures of sgRNA-containing (sgRNA+) and control plasmid without the sgRNA insert (sgRNA-) are grown and one set is treated with anhydrotetracycline (ATc) to induce transcription of dCas9 and the sgRNA (if present) from the Tet-activated promoters in both plasmids. Figure made with BioRender.com

5.4.5 RNA extraction and sequencing

RNA extraction was performed in the CL3 lab by members of Sharon Kendall's group at RVC according to established protocols (Rustad et al., 2009).

Extracted nucleic acids were then purified with RNeasy® columns (New England Biolabs) according to manufacturer instructions, and 400ng of each sample was DNase treated with 1 u DNaseI (Invitrogen) and 10X buffer in 20µL total volume at 37°C for 30m. 1 µL 25 mM EDTA was added to each sample and heated for 10m at 65°C to inactivate the enzyme. RNA was quantified using Qubit™ fluorometer and DeNovix™ spectrophotometer. 2-6 µg RNA from each sample was sent for paired-end sequencing on an Illumina NovaSeq platform. Between 74-148M reads were sequenced per sample with mean read quality score > 38 and length of 150 bp.

5.4.6 Quantification and Data Analysis

Reads were downloaded from sequencing provider server and quality control checked for read length and appropriate headers. Remaining adapters were trimmed and reads filtered for quality with *fastp* (Chen et al., 2018). Reads were mapped to the *M. tuberculosis* H37Rv genome (AL123456.3) using *bwa-mem* (Li 2013) for paired-end reads with default parameters. Unmapped reads (102K-160K per sample) were mapped to the episomal pRH2502 plasmid sequence with > 97% mapped reads. Mapped reads with length 20-21 bp and mapping to location of sgRNA target sequence were generated from sequencing of the expressed sgRNA on the episomal pASphoR_2 and were therefore removed from further analysis using bash scripts with samtools (Danecek et al., 2021). Base coverage files were created for each sample using deepTools *bamCoverage* (Ramírez et al., 2016), normalising with RPKM and a bin size of 10bp. DeepTools *coverageBed* was used to calculate strand-specific, mean base-pair read coverage for select transcripts (*phoP*, *phoR*, *as-phoR*) and create bedgraph files for ATc-treated and untreated Mtb_{dCas9}-sgRNA2 RNA-seq samples.

Reads mapping to protein-coding and non-coding annotations in *M. tuberculosis* H37Rv (using custom annotation file including predicted ncRNA from Chapter 2) were quantified using *count_features* from the baerhunter R-package (Ozuna et al., 2019). This generates a 'counts matrix' of raw read counts per feature.

The counts matrix was used with DESeq2 (Love et al., 2014) and sva (Leek et al., 2012) to test for surrogate variables. Data was explored using DESeq2 and PCATools (Blighe & Lun 2024), with and without rRNA transcripts included in the counts matrix, to identify batch effects. Batch correction was made using limma (Ritchie et al., 2015). Differential expression analysis was performed with DESeq2 using a model design that incorporated the batch effect (experiment) and looked for the interaction between the ATc treatment and empty vs sgRNA-containing plasmid (~ 1 + experiment + plasmid + treatment + treatment:plasmid). Log₂ fold changes depending on contrasting conditions (experiment, treatment, plasmid) were generated and assigned significance values (adj. p-values corrected for multiple testing with Benjamini-Hochsberg method (Benjamini & Hochberg, 1995). Gene set enrichment was performed using GSEA, gseKEGG and gseMKEGG from the

clusterProfiler package (G. Yu et al., 2012) with the ranked log₂ fold-changes from the relevant DESeq2 contrast results.

5.4.7 RNA structure and binding prediction

RNA secondary structure prediction of as-phoR was performed with RNAfold (Gruber et al., 2007). Prediction of RNA interactions with antisense and differentially-expressed mRNA targets was performed with IntaRNA (Mann et al., 2017) with minimum seed sequence set to 6. TargetRNA3 (Tjaden, 2023) was used with the as-phoR sequence to scan for potential mRNA targets in the genome.

5.4.8 Analysis of publicly available RNA-seq datasets

Publicly available RNA-seq datasets were downloaded from SRA and processed as in Materials and Methods, Chapter 2. Mapped reads were visualised using Artemis (Carver et al., 2012) and IGV (Robinson et al., 2011) genome browsers. For *M. bovis* or *M. smegmatis* datasets, fastq reads were mapped to the appropriate genome: *M. bovis*, AF2122/97 (Accession: LT708304.1), *M. smegmatis* MC2 155 (Accession: CP000480.1) using *bwa-mem* with paired-end reads and default parameters (Li 2013). Differential expression analysis of as-phoR in ΔsigE (Baruzzo et al., 2023) was performed with DESeq2 using a model design that looked for the interaction between low and high phosphate conditions and ΔsigE mutant versus wild-type (~ 1 + condition + genotype + condition:genotype). Log₂ fold changes depending on contrasting conditions were generated and assigned significance values (adj. p-values corrected for multiple testing with Benjamini-Hochsberg method (Benjamini & Hochberg, 1995)).

All scripts for bioinformatics analysis are available at https://github.com/jenjane118/thesis_work/tree/main/Chapter 5. Data analysis and wrangling was performed with *dplyr* (Wickham et al., 2022) in R. Plots made with ggplot2 (Wickham, 2016), unless otherwise indicated.

5.5 RESULTS

5.5.1 An antisense transcript opposite *phoR* gene is predicted from *M. tuberculosis* RNA-seq data

Computational predictions of non-coding RNA from RNA-seq data (Chapter 2) identified an antisense transcript (ncRv0757c, putative_sRNA:m852286_852683, 'as-phoR') transcribed opposite the 5' end of the sensor kinase component member of the PhoPR two-component system (Figure 5.7). WGCNA analysis (Chapter 2) indicated as-phoR is a well-connected hub in a co-expression module enriched for sRNAs and PE/PPE genes ('darkturquoise'), while the *phoR* and *phoP* genes are clustered in different modules. PE/PPE genes are virulence factors thought to be involved in host-pathogen interactions in mycobacteria and to regulate transport across the outer membrane (Babu Sait et al., 2022; Boradia et al., 2022; Damen et al., 2022; Dechow et al., 2021). The module had a weakly negative correlation with the low iron condition (bicor=-0.37, $p_{\text{adj}} = 0.03$) and was slightly better correlated with stationary growth (bicor=0.43, $p_{\text{adj}}=0.007$) (Chapter 2, Figure 2.3).

The antisense overlaps what is predicted to be a signal peptide and extra-cellular sensing domain of the PhoR protein (Figure 5.7), as well as the 44 nucleotide intergenic region between *phoP* and *phoR*. The intergenic region contains a short open reading frame (sORF) and ribosome profiling studies have shown it to be translated (Sawyer et al., 2021; Smith et al., 2022) (Figure 5.7). Within the sORF there are located putative RNaseE cleavage sites (852193, 852199) predicted from RNA-seq data using a differential ligation strategy which maps all 5' transcript ends in treated libraries with removed 5' triphosphates versus untreated libraries (Zhou et al., 2023).

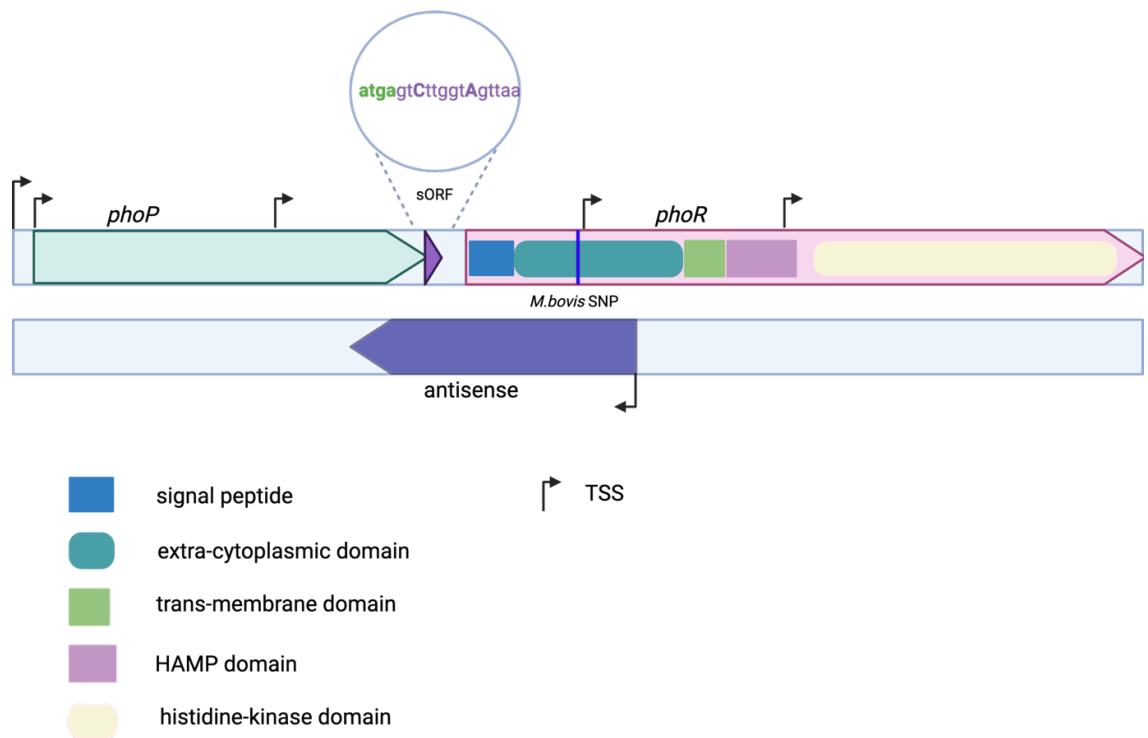


Figure 5.7. Diagram showing the orientation and length of as-phoR transcript relative to phoPR polycistronic operon. The antisense overlaps the region coding for the predicted extra-cytoplasmic and signal peptide domains of phoR and intergenic region between phoP and phoR, as well as the *M. bovis* phoR G/I SNP at codon 71. Close-up of the intergenic region shows a translated sORF of 18 nucleotides (852348-852365) that is in the same reading frame as phoR and overlaps the stop codon of phoP ('atga' in green). Putative RNaseE cleavage sites are predicted within the sORF at coordinates 852193 and 852199 (in bold) (Zhou et al., 2023). TSS coordinates, left to right: 851548, 851607, 852061, 852612, 852683, 852936 (Cortes et al., 2013; Ju et al., 2024; Shell et al., 2015). Figure created with BioRender.com.

The antisense transcript has a reported TSS within the first 10 nucleotides of its predicted start (Cortes et al., 2013; Ju et al., 2024; Shell et al., 2015) and overlaps the known SNP found in *M. bovis* (and other animal-adapted lineages) (Figure 5.8). This SNP is located approximately 4 nucleotides upstream from an alternative TSS in *phoR* (Cortes et al., 2013; Ju et al., 2024), detected in exponential growth conditions, which is 71 nucleotides downstream from the start of as-phoR. A motif similar to the SigE promoter motif (GGAAC-T/C-N17-18-GTT) appears within the upstream 35 nucleotides of as-phoR (Newton-Foot & Gey van Pittius, 2013; Song et al., 2008). SigE has been found to interact with PhoP and maintains redox and pH homeostasis as part of a stress-response system in *M. tuberculosis* (Bansal et al., 2017; Baruzzo et al., 2023; Goar et al., 2022). However, re-analysis of *M. tuberculosis* H37Rv Δ sigE strain versus wild-type RNA-seq data (Baruzzo et al., 2023), including non-coding RNA annotations (from Chapter 2), did not show differential expression of as-phoR, *phoR* or *phoP* transcripts in the conditions tested.

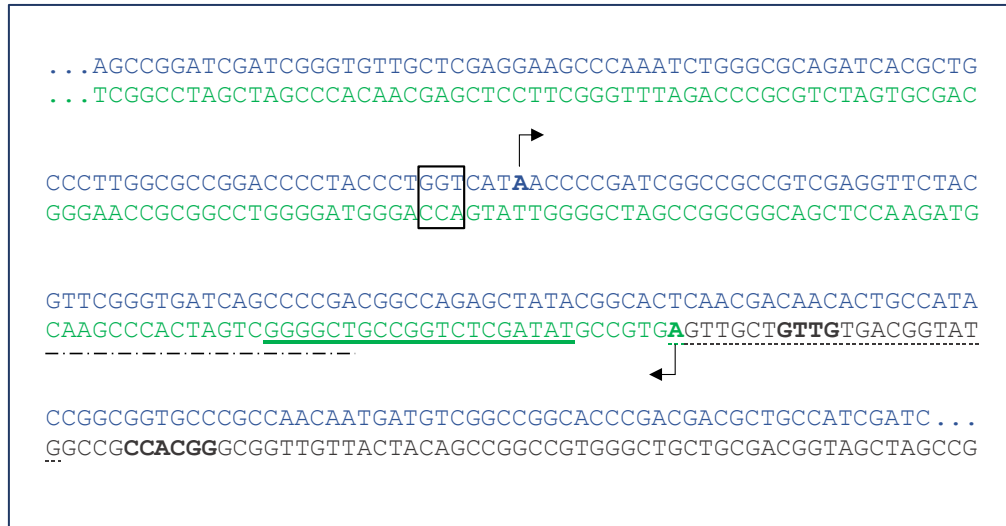


Figure 5.8. Sequence of phoR (blue) and as-phoR (green) in region 852525-852758 (*M. tuberculosis* H37Rv, AL123456.3). Transcriptional start sites marked with arrows (as-phoR TSS: 852683, phoR alt-TSS: 852617)(Cortes et al., 2013; Ju et al., 2024; Shell et al., 2015). The *M. bovis* G/I SNP (852606-852608) is outlined in black box. Positions of sgRNAs are underlined: sgRNA1 = dotted, sgRNA2 = solid green, sgRNA3 = dashed. Potential SigE promoter motif indicated in bold.

5.5.2 Antisense-phoR is expressed in multiple RNA-seq datasets, including from *M. bovis*

The antisense-phoR transcript was expressed in all conditions in the *M. tuberculosis* RNA-seq datasets analysed in Chapter 2, with increased expression in certain stress conditions (Figure 5.9). This transcript was also observed in *M. bovis* (AF2122/97) RNA-seq datasets (PRJNA390669, PRJNA774648) in stationary and rolling cultures but was not observed to be expressed in RNA-seq data (PRJNA820116) from the non-pathogenic, fast-growing strain, *M. smegmatis*, MC2-155.

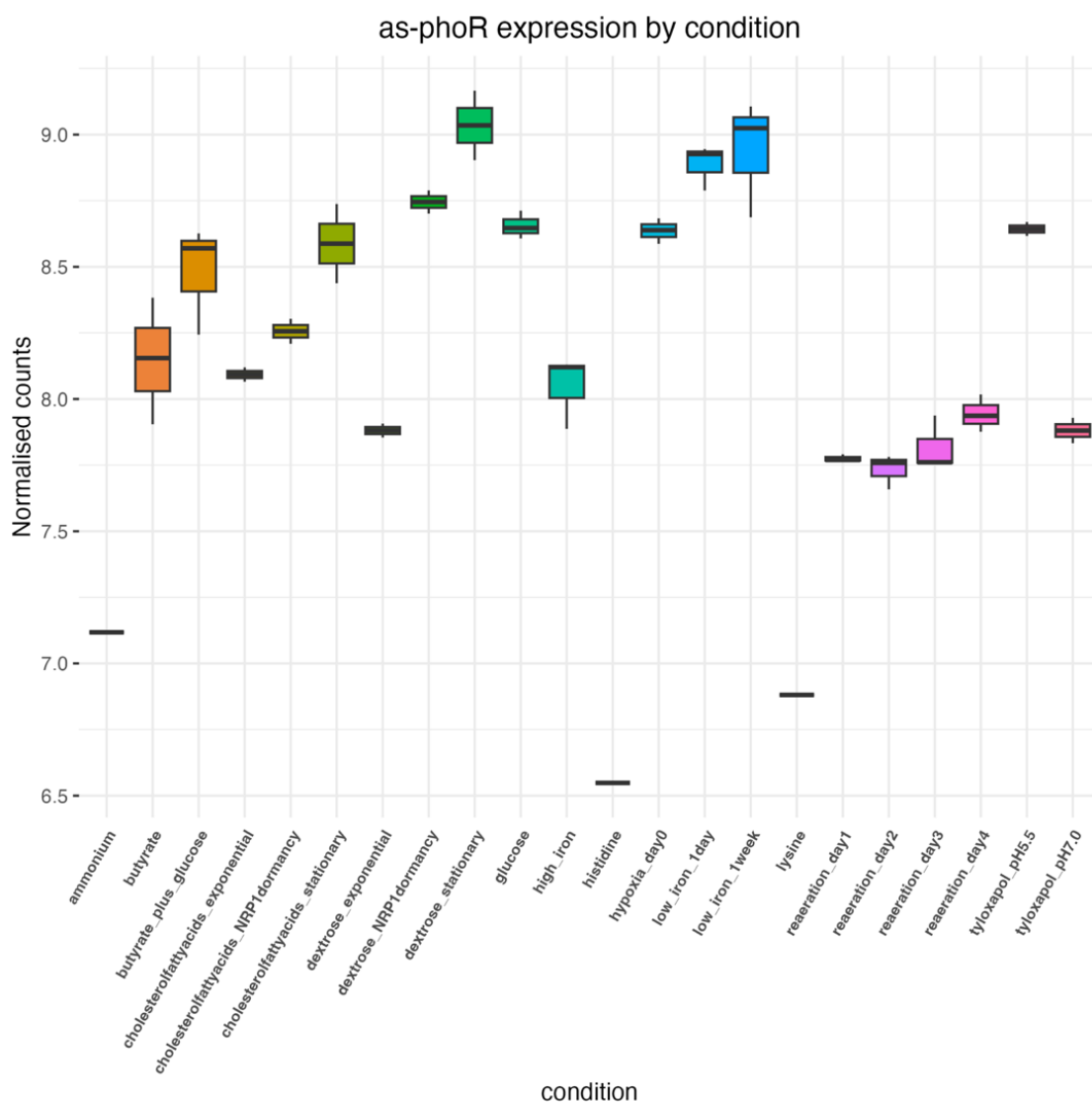


Figure 5.9. Boxplots showing relative expression levels of *as-phoR* across multiple conditions using normalised counts from publicly available RNA-seq data (SRA datasets: PRJEB65014_3, PRJNA278760, PRJNA327080, PRJNA390669, see Materials and Methods, Chapter 2). Solid line represents median expression across 1 to 3 replicates depending on condition.

5.5.3 Transcripts in region of antisense-*phoR* are detected at similar levels to housekeeping gene, *sigA*, in exponential growth conditions

In order to validate the prediction and to measure the levels of expression, RT-qPCR was used (Materials and Methods) to compare levels of the *as-phoR* transcript with *sigA*, a common housekeeping gene in total RNA from exponentially-growing *M. tuberculosis*. *As-phoR* was expressed at similar abundance to *sigA* (Figure 5.10). However, as random primers were used to generate the reverse-transcribed cDNA library used in the PCR reactions, the detected transcripts could originate from either strand and could include both sense and antisense expression.

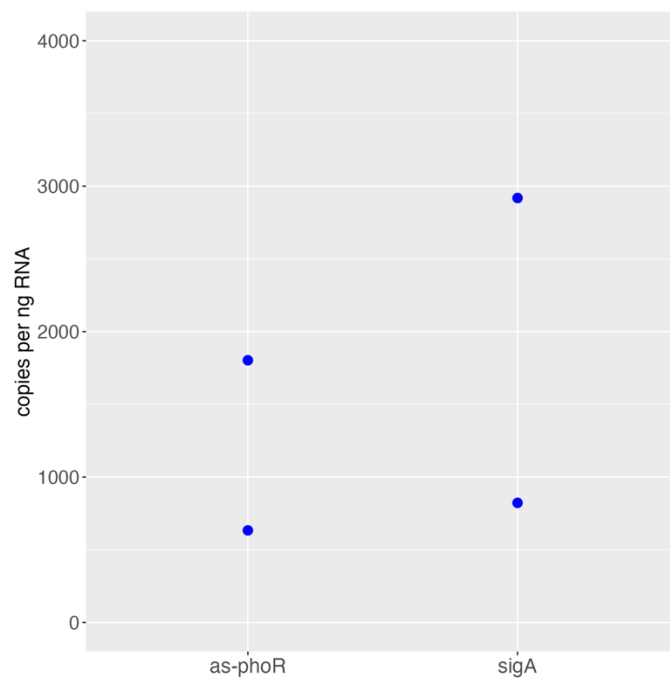


Figure 5.10. RT-qPCR with primers specific to *as-phoR* and *sigA* (Table 5.1) indicates expression in the region of the 5' end of *as-phoR* is comparable to *sigA* in *M. tuberculosis* total RNA. Each point represents the mean of two technical replicates of qPCR from independent RNA samples (biological replicate). Copy number calculated using absolute quantification against gDNA standard curves (Materials and Methods).

5.5.4 Antisense-phoR is silenced using CRISPRi

In order to determine how *as-phoR* might impact the *M. tuberculosis* transcriptome, and the PhoPR regulon specifically, we utilised a CRISPR inhibition system to knockdown its expression in exponentially-growing cultures of *M. tuberculosis* (Materials and Methods). Before RNA extraction, colony PCR was performed to confirm that each transformed strain (Table 5.2) included both the integrated and autonomous plasmids (Figure 5.11).

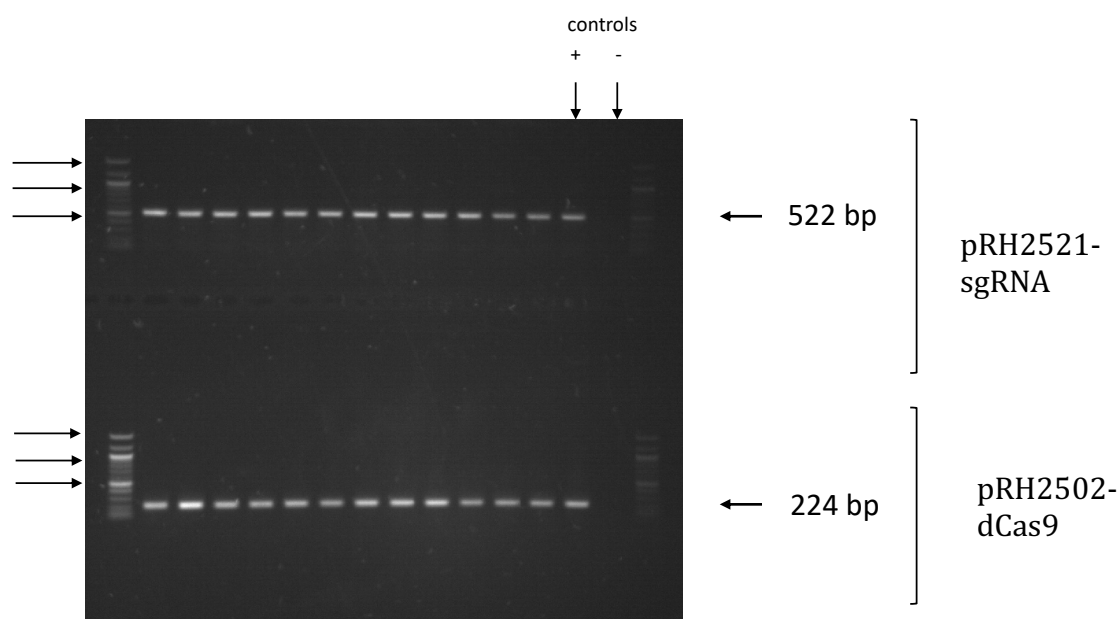


Figure 5.11. Colony PCR (Materials and Methods, 4.5.4) was used to confirm retention of episomal pRH2521 and integrated pRH2502 plasmids from plated colonies of transformed strains (Table 5.2). Top row shows 522bp fragment expected with primers pRH2521-sgRNA and bottom row, the 224 bp fragment from pRH2502-dCas9 primers (Table 5.1). Left-hand side arrows show DNA ladder bands at 1517, 1000, and 500 bp. Positive control with plasmid DNA and negative control with no DNA.

Initial PCA of the samples indicated that the samples were clustering by biological replicate with PC1 highly correlated to replicate experiment (Figure 5.12A). Surrogate variable analysis identified a variable that correlated well to the replicate number; and after batch correction, PC1 was most strongly correlated to treatment with ATc (Figure 5.12B). The surrogate variable was controlled by incorporating an additional factor ('experiment') into the design of the DESeq2 model (Materials and Methods).

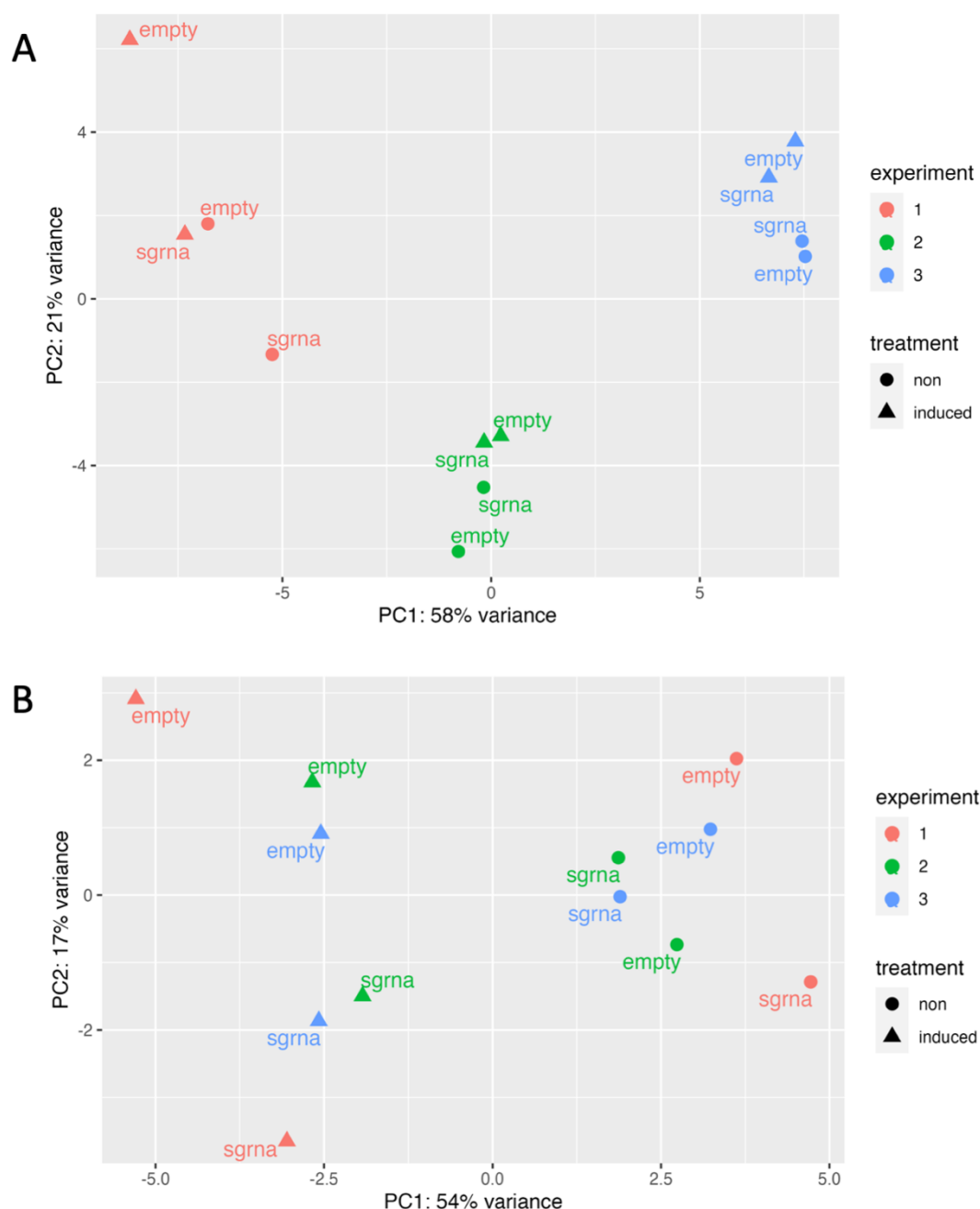


Figure 5.12. PCA plots before and after batch correction with Limma (A) PCA shows samples clustering by replicate ('experiment'). (B) Samples after batch correction no longer cluster by experimental replicate ('experiment'). PC1 shows clustering by ATc treatment ('treatment'). Samples labelled by plasmid: sgrRNA or empty vector control. Plots made with ggplot2 (Wickham, 2016).

Differential expression analysis using DESeq2 (Love et al., 2014) was applied across all RNA-seq samples (Materials and Methods)(A5.1 Supplemental Tables: Ch5_Supp_Table_1). *DCas9* expression was shown to be induced by ATc (log₂ fold-change = 4.8 in Mtb_{dCas9_ctrl} and 4.9 in Mtb_{dCas9_sgrRNA2}, $p_{adj} < 0.0001$). Using an interaction term in the linear model to reflect the differential effect of ATc treatment

on the Mtb_{dCas9_sgRNA2} versus Mtb_{dCas9_ctrl} strains, there were 19 differentially expressed transcripts (Figure 5.13): 5 downregulated transcripts (including as-phoR) and 14 upregulated transcripts. As-phoR expression was inhibited by 95% (\log_2 fold-change = -4.214).

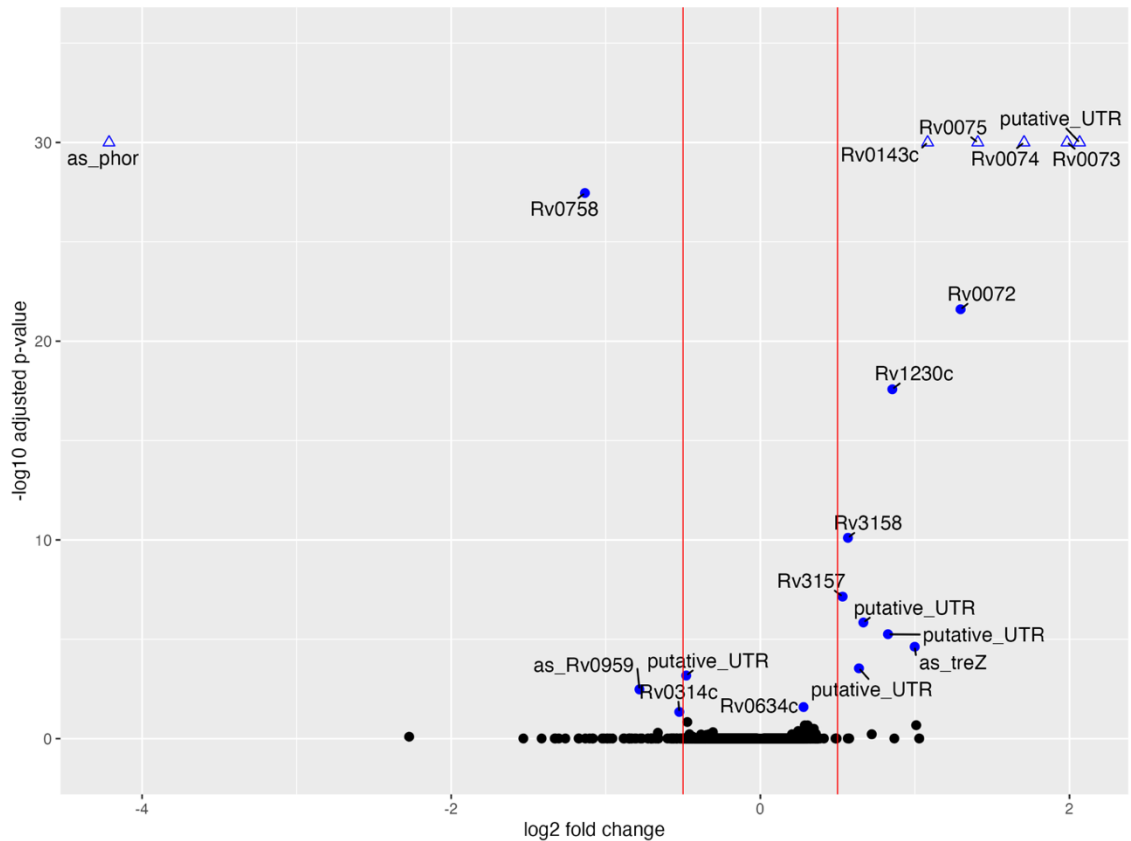


Figure 5.13. Volcano plot showing differentially expressed transcripts with knock-down of as-phoR expression. Blue points represent statistically significant \log_2 fold changes ($p_{adj} < 0.05$); red lines indicate \log_2 fold changes greater or less than 0.5; triangles indicate points with $p_{adj} < 10^{-30}$.

5.5.5 Antisense silencing impacts *phoR* expression

Unexpectedly, expression of *phoR* (Rv0758) was downregulated with antisense inhibition by approximately 50% (Figure 5.13, Figure 5.14). To better visualise the relative effects of antisense silencing on gene expression across the relevant genomic region, the ratio of mean per-base-pair read coverage in ATc-treated Mtb_{dCas9_sgRNA2}, versus untreated, was calculated for each transcript and plotted against genomic coordinates (Figure 5.15). The plot shows that read coverage of as-phoR is lower in the ATc-treated strain from the very start of the transcript, which can be explained by a transcriptional 'pause' by the RNA polymerase as it collides with the *dCas9* (Qi et al., 2013). Coverage gradually returns to the level of untreated with increasing distance from the target site. At the 5' end of *phoR*, coverage very

closely mirrors the decreased coverage of *as-phoR* at the overlapping coordinates, and the non-overlapping second-half of the transcript has less coverage than ATc-untreated levels. The decrease in *phoR* coverage occurs from the very start of *phoR* transcription, 260 bp upstream from where the sgRNA:*dcas9* complex is bound (Figure 5.14, Figure 5.15, Figure 5.16), however, it is possible that the decrease in *phoR* expression is due to the large sgRNA:*dcas9* complex blocking transcription elongation from both directions.

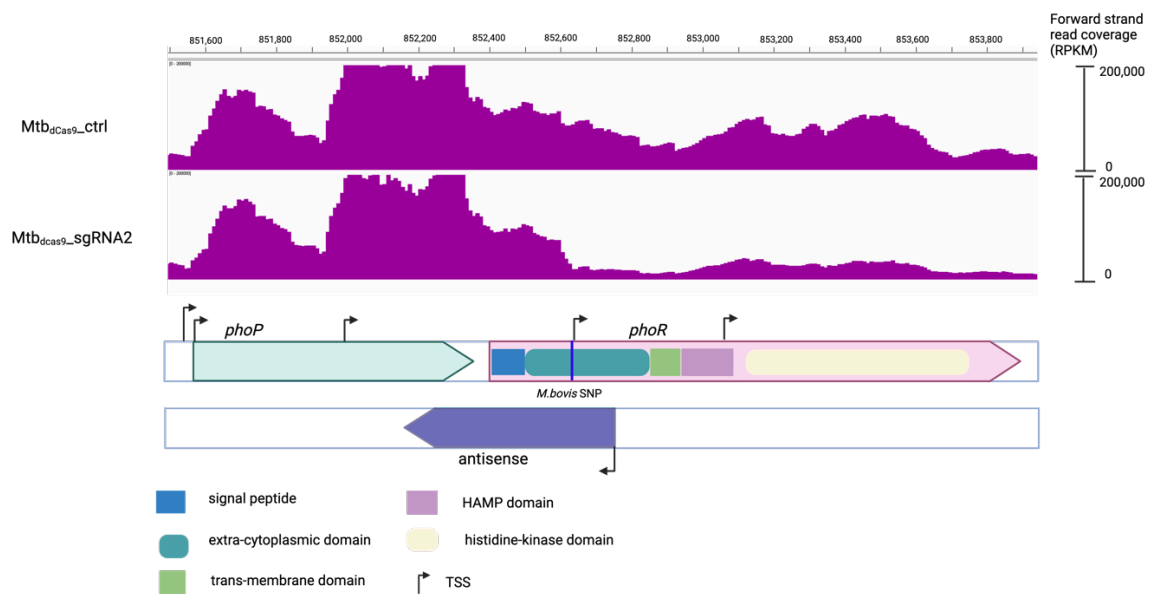


Figure 5.14. Read coverage from forward strand mapped to *phoR* annotations (representative sample). In silenced strain (*Mtb_{dCas9}_sgRNA2*), coverage drops-off at approximately 852,600, close to location of *M. bovis* SNP and first alternative TSS. Read coverage normalised by RPKM (Materials and Methods). Coverage visualised with IGV (Robinson et al., 2011). Figure made with BioRender.com.

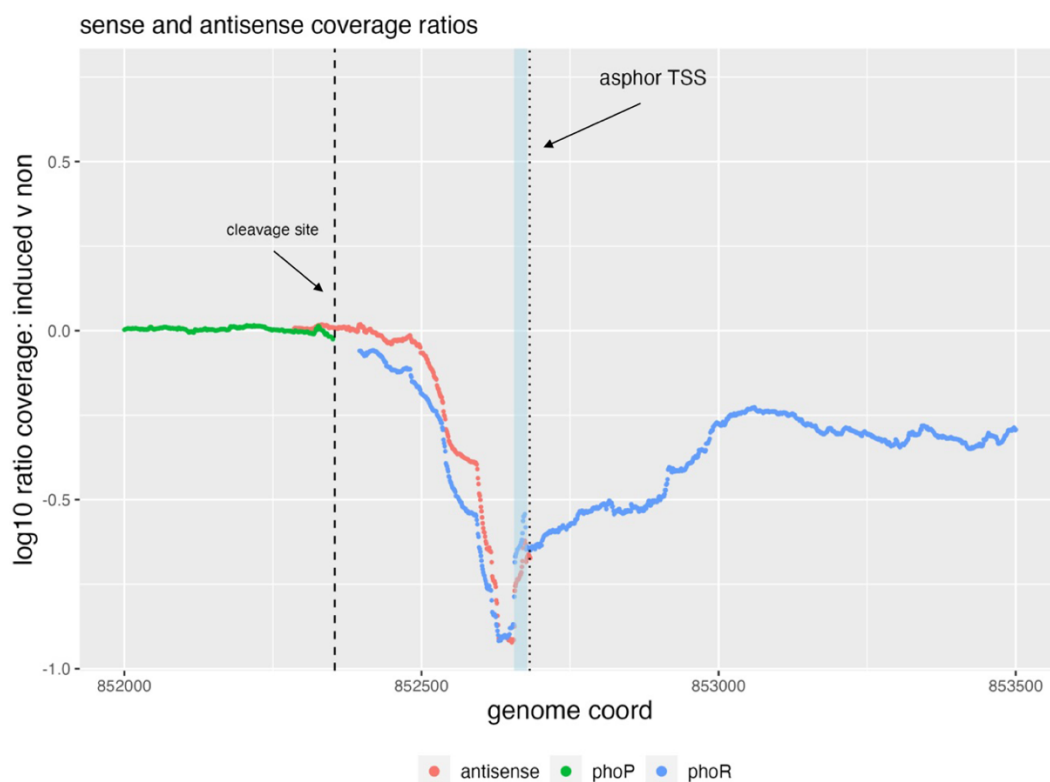


Figure 5.15. The \log_{10} ratio of mean base pair read coverage of ATc-treated (induced) vs. untreated *Mtb_{dCas9}_sgRNA2* is plotted for each transcript (phoP=green, as-phoR=coral, phoR=blue) in region of 852000-853500 (see Materials and Methods). Dashed line is location of predicted RNaseE cleavage site (Zhou et al., 2023). Dotted line is TSS for as-phoR. Light blue bar is location of 20nt sgRNA2 target sequence.

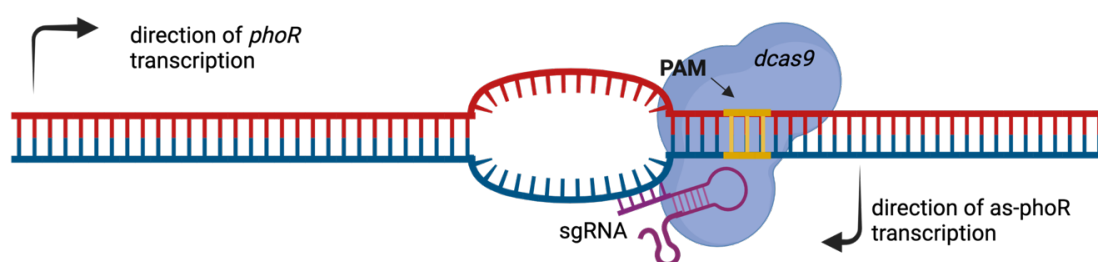


Figure 5.16. The orientation of the sgRNA:dcas9 complex relative to direction of transcription. In this experiment, the sgRNA is directed to the start of as-phoR by a sequence complementary to the non-template strand of the antisense transcript (i.e., antisense to phoR). This places dcas9 upstream of the sgRNA base-pairing sequence in relation to as-phoR. The transcription elongation complex is thought to collide with the dcas9 complex, thus inhibiting elongation of as-phoR. Transcription of phoR is initiated from the opposite strand (sense), and the sgRNA:dcas9 complex is bound 260 bp downstream from the start of the phoR transcript. In this case, sgRNA binding to target sequence may be disrupted by the RNAP helicase before collision (Qi et al., 2013). Created with BioRender.com and based on (Qi et al., 2013; Rock et al., 2017).

5.5.6 Protein-coding genes associated with the cell membrane were differentially expressed with antisense-silencing

Eight protein-coding genes were upregulated (\log_2 fold-change > 0.5) with antisense silencing (Table 5.3). These include an ATP-binding cassette glutamine transporter, and the related genes in the operon, (Rv0072-Rv0075) and both subunits of the membrane-associated Respiratory Complex I: NADH oxidoreductase, nuoNM (Rv3157-58). Genes Rv0073, Rv0074 and Rv0075, nuoN and nuoM are hubs in the same co-expression network module ('pink') which is enriched for the functional category, 'information pathways' (out of 148 total hub genes in the module, A2.1 Supplemental Tables: Ch2_Supp_Table_4). Two further transcripts had statistically-significant \log_2 fold changes of > 0.5 (Rv1230c and Rv0143c) and are both membrane-associated proteins. Rv1230c, UniProtKB ID: O86313_MYCTU, contains a transglycosylase SLT-2 domain and may be involved in peptidoglycan catabolism. Rv0143c, UniProtKB ID: P96820_MYCTU, is a probable voltage-gated Cl⁻ channel. (Paysan-Lafosse et al., 2023). All of these upregulated protein-coding transcripts showed downregulation in the ATc-treated Mtb_{dcas9_ctrl}, versus levels equal to untreated samples in ATc-treated Mtb_{dcas9_sgRNA2}. (Figure 5.17). All 6 differentially expressed UTRs were adjacent to the differentially expressed coding gene and were similarly upregulated (Table 5.3). Rv0314c, a gene for a possible membrane-associated protein with a domain of unknown function was the only protein-coding gene, other than *phoR*, downregulated with antisense silencing. None of the genes of the PhoP regulon were differentially expressed, as defined by (Cimino et al., 2012; Solans et al., 2014)(A5.1 Supplemental Tables: Ch5_Supp_Table_1).

It is possible that the antisense transcript is acting as an independent sRNA and binding mRNA of these differentially expressed genes directly. To explore this possibility, firstly, the secondary structure of the predicted as-*phoR* transcript (using the most highly expressed region, bases 1-144) was predicted with the RNAfold web server (Gruber et al., 2008). A long stem-loop is predicted by base-pairing probabilities, but there is no obvious seed region for target binding (Figure 5.18). The as-*phoR* sequence was then used with targetRNA3 (Tjaden, 2023) to find potential target mRNA binding partners in the *M. tuberculosis* transcriptome. TargetRNA3 is a machine-learning tool trained on known sRNA-mRNA interactions, of which there are very few in *M. tuberculosis*. The only hits were *phoR*, and

Rv2521/*bcp*, which was not one of the differentially expressed genes and likely a false positive. A more focused approach was tried with the *as-phoR* sequence and sequences of the differentially expressed protein-coding genes as input for IntaRNA to predict RNA-RNA interactions based on folding and hybridization energies (Materials and Methods 5.4.7) (Mann et al., 2017). Several plausible interactions with negative free energies were predicted with the differentially expressed genes, however, no clear seed region of *as-phoR* emerged (A5.1 Supplemental Tables: Ch5_Supp_Table_2).

Table 5.3. Differentially-expressed genes resulting from the interaction between ATc treatment and *Mtb*_{dCas9_ctrl} versus *Mtb*_{dCas9_sgRNA} with $p_{adj} < 0.05$. See Materials and Methods, 5.4.6. WGCNA module refers to co-expression network module to which the transcript was assigned in Chapter 2; * indicates a hub transcript with a high module connectivity score.

LOCUS	NAME	log2 FOLD-CHANGE	ADJ P-VALUE	FUNCTIONAL CATEGORY	WGCNA MODULE	PRODUCT
Rv0072		1.296	2.49E-22	cell wall and cell processes	tan*	glutamine transporter
Rv0073		1.985	2.82E-76	cell wall and cell processes	pink*	glutamine transporter
putative_UTR:p82669_82747	UTR Rv0073-Rv0074	2.066	1.31E-40	non-coding RNA	pink*	between Rv0073-74 conserved hypothetical prob
Rv0074		1.706	2.19E-76	conserved hypotheticals	pink*	aminotransferase
Rv0075		1.407	4.09E-61	intermediary metabolism and respiration	pink*	overlaps Rv0076c (head-to-head)
putative_UTR:p85169_85508	3' UTR Rv0075	0.667	1.42E-06	non-coding RNA	skyblue*	overlaps Rv0142
putative_UTR:m168648_168703	3' UTR Rv0143c	0.826	5.56E-06	non-coding RNA	brown	membrane protein (possible Cl- channel)
Rv0143c		1.083	2.19E-34	cell wall and cell processes	brown	
putative_UTR:m382621_382878	3' UTR Rv0314c	-0.479	6.66E-04	non-coding RNA	midnightblue	
Rv0314c		-0.524	4.50E-02	cell wall and cell processes	darkgreen	membrane protein
Rv0634c		0.280	2.57E-02	virulence, detoxification, adaptation	black	Possible glyoxalase II (hydroxyacylglutathi
putative_sRNA:m852286_852683	<i>as-phoR</i>	-4.214	4.95E-144	non-coding RNA	larkturquoise	antisense to <i>phoR</i>
Rv0758	<i>phoR</i>	-1.135	3.50E-28	regulatory proteins	lightgreen*	sensor kinase
putative_sRNA:m1073103_107330	antisense Rv0959	-0.782	3.31E-03	non-coding RNA	yellow*	
Rv1230c		0.854	2.62E-18	cell wall and cell processes	darkgreen	prob membrane protein
putative_sRNA:p1766474_1766851	antisense Rv1562 (<i>treZ</i>)	0.999	2.38E-05	non-coding RNA	red	
Rv3157	<i>nuoM</i>	0.532	7.10E-08	intermediary metabolism and respiration	pink*	NADH-ubiquinone oxidoreductase
Rv3158	<i>nuoN</i>	0.567	7.88E-11	intermediary metabolism and respiration	pink*	NADH-ubiquinone oxidoreductase
putative_UTR:p3527386_3527468	3'UTR Rv3158	0.638	2.88E-04	non-coding RNA	pink*	

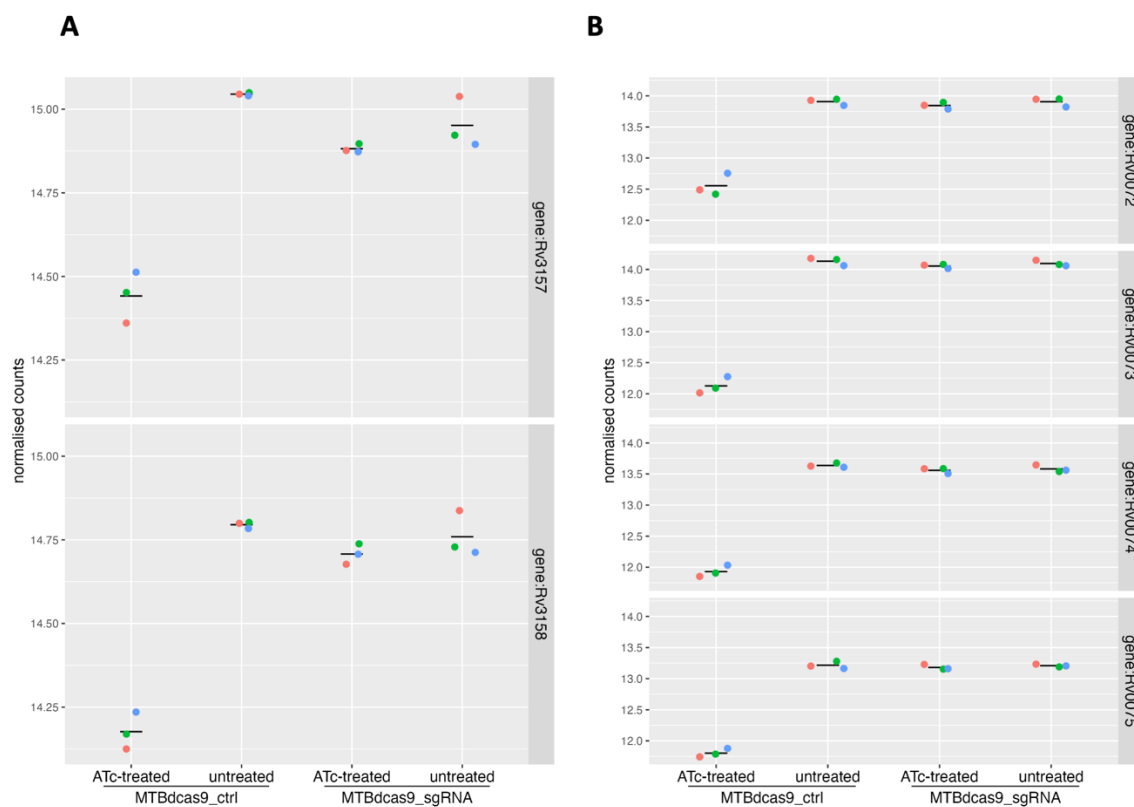


Figure 5.17. Plot of normalised counts versus strain for A) *nuoMN* (Rv3157 and Rv3158) and B) glutamine transport operon, Rv0073-Rv0075. Bar represents median of normalised counts from three independent biological replicates.

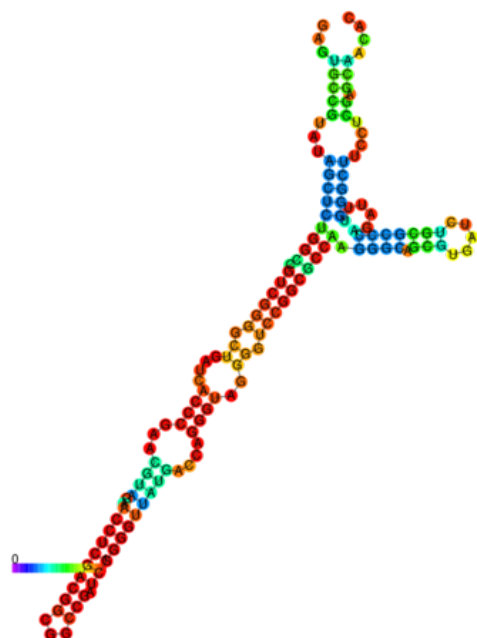


Figure 5.18. Secondary structure of *as-phoR* as predicted by RNAFold (Gruber et al., 2007). Base-pairing probabilities scaled 0-1, with red representing the highest confidence interactions

5.5.7 Two antisense transcripts are differentially expressed with as-phoR silencing

An antisense transcript (putative_sRNA:p1766474_1766851) coding opposite *treZ* (Rv1562) was also upregulated, however, unlike the upregulated protein-coding genes, this transcript was not downregulated in the ATc-treated Mtb_{dCas9_ctrl} strain (Figure 5.19A). TreZ is involved in trehalose synthesis--a disaccharide used for carbon storage, incorporated into phospholipids and involved in maintaining the mycolic acid cell wall in *M. tuberculosis*. (Kalscheuer & Koliwer-Brandl, 2014). Antisense transcription begins opposite a region in the second half of the gene with a TSS recorded at 1766472 (Figure 5.20A). An antisense transcript (putative_sRNA:m1073103_1073304) complementary to the 5' end of Rv0959 was downregulated with as-phoR silencing (Figure 5.19B). Transcription of this antisense initiates at a predicted TSS (1073304) from within the 54 bp intergenic region between Rv0959 and Rv0959A (*vapB9*) from what could be a bi-directional promoter (Figure 5.20B). Rv0959 is an uncharacterised protein thought to be associated with the cell membrane (Mawuenyega et al., 2005) and homologous to the transcriptional repressor of a nitrate reductase operon in *Corynebacterium glutanicum*, ArnR (Huang et al., 2015). VapB9 is the antitoxin member of a proposed toxin-antitoxin system (Ramage et al., 2009).

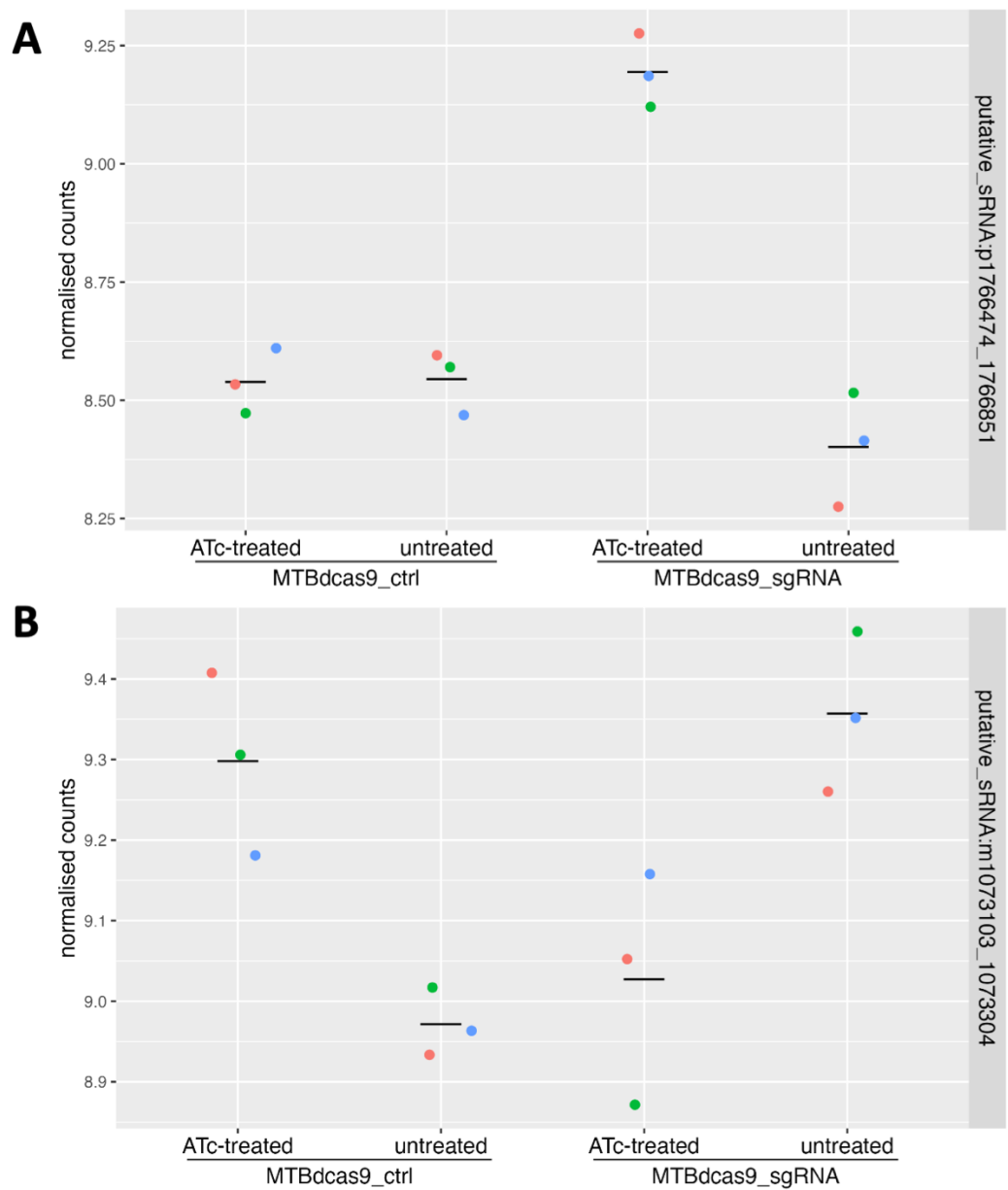


Figure 5.19. Plots of normalised counts versus strain for A) antisense-treZ (putative_sRNA:p1766474_1766851) and B) antisense-Rv0959 (putative_sRNA:m1073103_1073304). Bar represents median counts for three independent biological replicates.

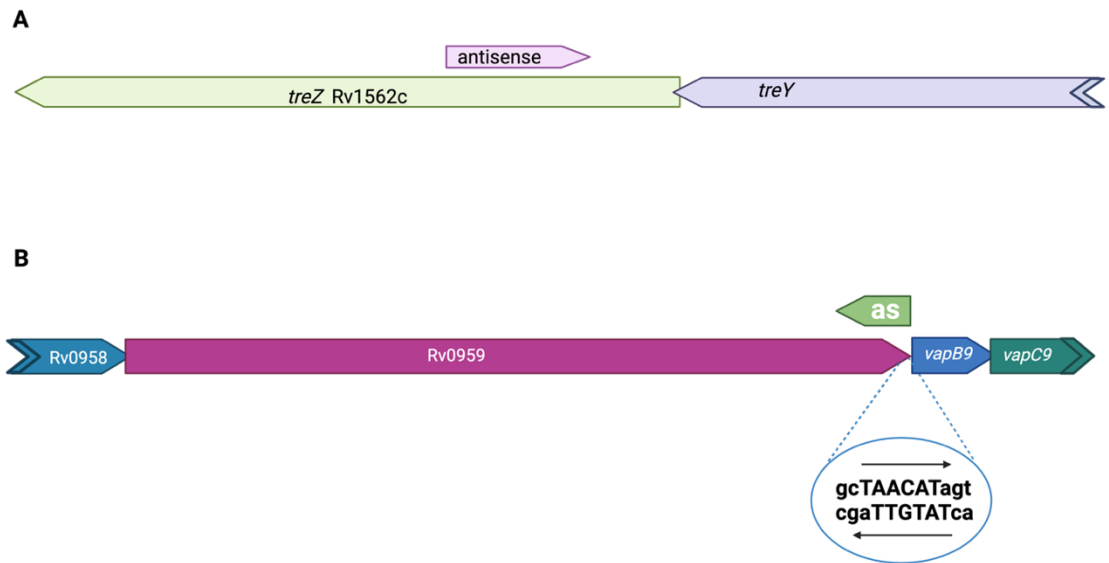


Figure 5.20. Schematic of the relative position and lengths of differentially expressed antisense transcripts. A) putative_sRNA:p1766474_176685 is coded opposite *treZ* with a TSS at 1766472. B) putative_sRNA:m1073103_1073304 ('as') is initiated in intergenic region between 3' end of *Rv0959* and 5' end of *vapB9* (*Rv0959A*) (TSS at 1073304). There are overlapping -10 promoter sequences (TAYgAT) for the antisense and *vapB9* (Newton-Foot & Gey van Pittius, 2013). Figures made with BioRender.com

5.5.8 ATc treatment results in differentially expressed genes in both control and sgRNA expressing strains

420 statistically significant differences ($p_{\text{adj}} < 0.05$) were observed between ATc-treated and untreated *Mtb_{dCas9_ctrl}* and 343 significant changes between the ATc-treated and untreated *Mtb_{dCas9_sgRNA2}* (Figure 5.21, A5.1 Supplemental Tables: Ch5_Supp_Table_3). This indicates between 5-6% of the transcripts evaluated (in total, 7059 transcripts, including predicted UTR regions and antisense RNAs) are differentially expressed in response to ATc treatment alone, in either plasmid context. The protein coding genes changed were enriched for KEGG pathways: "Biosynthesis of secondary metabolites", "Microbial metabolism in diverse environments", and "Biosynthesis of cofactors" (Kanehisa et al., 2022). Gene Set Enrichment Analysis (Subramanian et al., 2005) of ranked \log_2 fold-changes with ATc treatment in *Mtb_{dCas9_sgRNA2}* and *Mtb_{dCas9_ctrl}* showed higher ranked changes were enriched for genes from pathways: "ABC transporters", "Oxidative phosphorylation" and "Homologous recombination". Contrasting *Mtb_{dCas9_sgRNA2}* versus *Mtb_{dCas9_ctrl}* in the ATc-treated condition, there were 29 differentially expressed genes (A5.1 Supplemental Tables: Ch5_Supp_Table_3). There were no

statistically significant differentially expressed genes between untreated *Mtb_{dCas9_sgRNA2}* and *Mtb_{dCas9_ctrl}* samples.

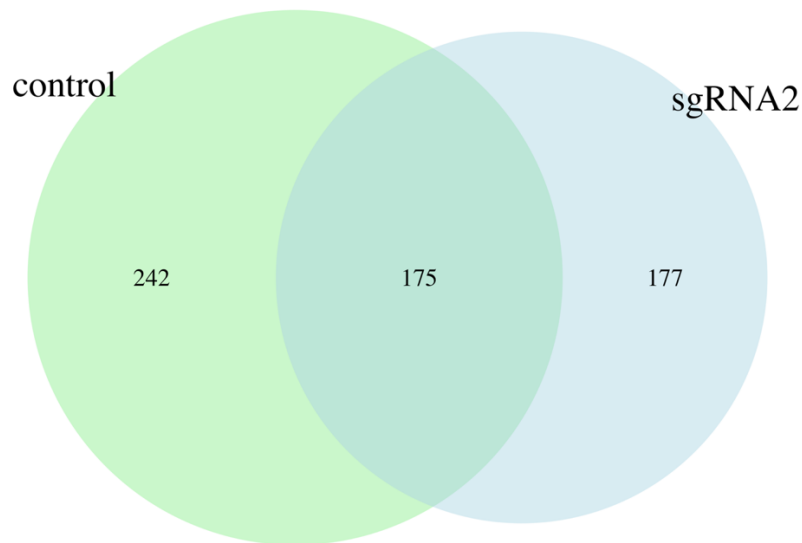


Figure 5.21. Number of differentially expressed genes with ATc treatment in *Mtb_{dCas9_sgRNA2}* and *Mtb_{dCas9_ctrl}* strains. $p_{adj} < 0.05$. 175 genes were differentially expressed with ATc treatment in both strains.

5.6 DISCUSSION

The aims of this chapter were to verify expression of a predicted antisense transcript that overlaps the 5' coding region of *phoR*, to silence expression of this transcript using CRISPRi and to evaluate the differentially expressed genes using RNA-seq. Expression of as-*phoR* in exponentially growing *M. tuberculosis* was successfully silenced using a CRISPRi system. RNA-seq analysis revealed that, with silencing of as-*phoR*, 10 protein-coding genes and two additional antisense transcripts were differentially expressed ($|\log_2 \text{fold-change}| > 0.5$, $p_{adj} < 0.05$), including the downregulation of *phoR* expression. Surprisingly, considering *phoR* expression was downregulated by 50%, none of the known genes of the PhoP regulon were differentially expressed.

5.6.1 As-*phoR* silencing reduces expression of *phoR*

Upon as-*phoR* silencing, there was a decrease in *phoR* expression that correlated well with as-*phoR* expression across the same region of the genome. It is important to consider that measurement of RNA transcript abundance by RNA-seq is a

'snapshot' of the relative abundances of various transcripts in the cells at a particular moment in time and therefore, related to both the transcription level of a particular gene and to the stability of the mRNA transcript. The pattern observed here could be interpreted to be a result of as-phoR regulation of *phoR* mRNA stability: in absence of the antisense, there is a decrease in stable *phoR* transcripts and/or more rapid degradation of the mRNA.

Binding of antisense transcripts to a mRNA can influence its vulnerability to cleavage by endoribonucleases. RNaseIII typically targets double-stranded RNA molecules, and widespread antisense transcription has been proposed as a mechanism for fine-tuning mRNA transcript levels by binding to sense transcripts and creating double-stranded templates (Dawson et al., 2022; Lasa et al., 2011; Lybecker et al., 2014; Ruiz de los Mozos et al., 2013). In a recent study in *Mycobacteria*, an antisense RNA was found to decrease transcript stability and protein expression, presumably by blocking access to the ribosome binding site (Li et al., 2022). Antisense expression was also responsible for differential expression of genes in a novel toxin-antitoxin bicistron in *M. tuberculosis*, where antisense binding creates a double-stranded RNA molecule which is specifically targeted by RNaseIII (with decreased antisense expression leading to inversely proportional increase in sense abundance) (Dawson et al., 2022).

In the work presented here, a decrease in as-phoR expression appears to lead to *decreased* mRNA stability, rather than the increase one would expect if *phoR* mRNA half-life was being regulated by RNaseIII. If this interpretation of the results is correct, it would be a novel mechanism of antisense regulation in *M. tuberculosis*. However, in other prokaryotes there are several documented examples of antisense stabilisation of mRNA. For example, in *E. coli*, overexpression of an antisense RNA, ArrS, was found to stabilise processed isoforms of *gadE* mRNA (Aiso et al., 2014). ArrS was strongly induced in stationary and acid pH conditions, similarly to as-phoR. In *Listeria monocytogenes*, UTRs from two different mRNAs bind and create a more stable, double-stranded chimera that resists 5'-3' exoribonuclease digestion by RNase J1 (Ignatov et al., 2020). In cyanobacteria, an antisense was found that stabilised the sense mRNA by occluding a RNaseE cleavage site (Sakurai et al., 2012).

There remains the possibility that the sgRNA:*dcas9* complex is interfering with transcription of both the targeted antisense transcript and *phoR* on the sense strand. Strand-specificity of the inhibition of transcriptional elongation is well-documented both in *E. coli* (Qi et al., 2013) and mycobacteria, where sgRNAs targeting the template strand of a gene target were ineffective at silencing the target transcript (Choudhary et al., 2015; Li et al., 2022; Singh et al., 2016). However, Howe et al. found that, at certain eukaryotic gene loci, CRISPRi was not strand-specific, causing transcriptional changes in both the sense and antisense directions due to chromatin interactions (Howe et al., 2017). As a decrease in *phoR* read coverage was observed beginning 260 bp upstream of where the sgRNA:*dCas9* complex was bound, steric hindrance of *phoR* transcription is not supported. To completely rule out this possibility of steric interference, or more profound chromatin changes, control sgRNAs targeting the sense strand opposite the antisense sgRNA (i.e. targeting the template strand of as-*phoR*) could be tested.

5.6.2 The intergenic region between *phoP* and *phoR* may be involved in translational regulation of *phoR*

The transcription and translation of *phoR* may involve several post-transcriptional regulatory elements that may allow for fine-tuning and more judicious protein production. Recent studies have implicated the translation of sORFs found in 5' leaders of *M. tuberculosis* coding transcripts in regulation of translation of the downstream gene (Kipkorir et al., 2024). A careful examination of the *phoPR* transcriptional unit and the 44 bp intergenic region found between the two genes revealed a short open reading frame (sORF) of 18 codons confirmed to be translated with ribosomal profiling (Figure 5.22) (Sawyer et al., 2021; Smith et al., 2022). This sORF is initiated at an overlapping stop/start codon at the end of *phoP* with an 'AUGA' pattern which is characteristic of the termination/re-initiation mechanism ('TeRe') described previously in other bacteria and in mycobacterial riboswitches (D'Halluin et al., 2023; Huber et al., 2019; Kipkorir et al., 2024) (Figure 5.23). The sORF is in the same reading frame as downstream *phoR*. Possibly the translation of the sORF initiates at the overlapping stop/start codon and facilitates handover of the ribosomes to the TIS of *phoR* such as proposed by (D'Halluin et al., 2023).



Figure 5.22. There is a sORF predicted between *phoP* and *phoR* that is actively translated. Screenshot from <https://mtb.wadsworth.org> which presents ribosomal profiling data (Ribo-seq and Ribo-RET) and sORF predictions from (Smith et al., 2022).

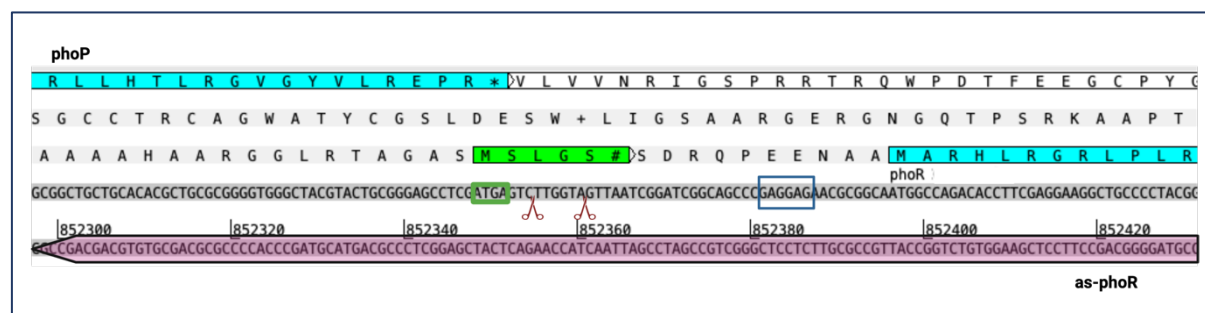


Figure 5.23. Close up of intergenic region between *phoP* and *phoR*. Coding sequences in cyan, antisense in rose, sORF is in green with 'AUGA' stop/start codon in green box. Possible purine-rich Shine-Dalgarno in blue box. Predicted cleavage sites indicated by scissors (Zhou et al., 2023). Sequences visualised with Artemis genome browser.

Within the sORF are reported potential RNaseE cleavage sites (Figure 5.23) (Zhou et al., 2023). Unlike RNaseIII which targets double-stranded RNA, RNaseE preferentially cleaves at single stranded regions of RNA, often between the genes of polycistronic transcripts, and is crucial for mRNA degradation in mycobacteria (Zhou et al., 2023), as well as in *E. coli* and *Bacillus subtilis* (DeLoughery et al., 2018; Trinquier et al., 2020). These cleavage sites can be masked by ribosomes or by non-coding RNA binding to nascent transcripts (Durand et al., 2015; Zhou et al., 2023). Cleavage in the single-stranded intergenic region between the two transcripts could impact mRNA stability of either transcript and may be a mechanism to regulate the stoichiometry of transcript expression (DeLoughery et al., 2018; Trinquier et al., 2020).

Active translation appears to promote transcriptional elongation in *M. tuberculosis*. In *E. coli*, translated mRNAs are recognised to be more stable as ribosomes can block or mask RNase cleavage sites, or interfere with RNase scanning (Iost & Dreyfus, 1995; Richards & Belasco, 2019); and the different mRNA secondary structure conformations can change the availability of translational initiation sites and RNase cleavage sites (Richards & Belasco, 2019; Trinquier et al., 2020). In *M. tuberculosis*, Ju, et al. (Ju et al., 2024), demonstrated that the RNA Polymerase complex is relatively inefficient compared to *E. coli*, and is prone to stalling 200-500 bp from the TSS, unless the transcript is being translated. Antisense transcripts, which are not typically translated, were observed to have even steeper drop-off of transcription than the coding transcripts. However, these incomplete transcripts were determined to be a result of the inefficient sigma factor in the Mycobacterial RNA polymerase complex, rather than RNase degradation (Ju et al., 2024).

Potentially, the translation efficiency of a transcript can be influenced by the folding of the nascent mRNA transcript, which may block ribosome binding to translation initiation sites (TIS). As-phoR binding could alter the secondary structure of the nascent *phoR* mRNA (including the sORF) and allow access to the TIS, resulting in increased mRNA stability. Further work to investigate the RNA folding of the polycistronic transcript could be done using in-line probing, which evaluates the relative rate of spontaneous cleavage of the RNA backbone. This rate is highest in single-stranded regions of a RNA molecule versus structured, more constrained, regions which are resistant to cleavage (Regulski & Breaker, 2008), and could be leveraged to predict secondary structure of the intergenic region between *phoP* and *phoR*. The region could be probed with increasing concentrations of as-phoR, using a similar strategy as probing riboswitch structure with ligands (Kipkorir et al., 2024; Regulski & Breaker, 2008).

5.6.3 Decrease in *phoR* expression does not impact genes of the PhoP regulon in exponential growth

In spite of the decrease in *phoR* expression, silencing of the as-phoR transcription had no apparent effect on expression of *phoP* or on the other genes of the PhoP regulon (Cimino et al., 2012; Gonzalo-Asensio et al., 2008). Other genes were differentially expressed in the silenced strain that have not previously been linked

to the PhoP regulon, and it is unclear whether these effects are due to the downregulation of *phoR* or by the abolition of direct interactions with the antisense. Off-target interactions of the sgRNA were ruled out by following up on all BLAST hits of the target sequence plus two nucleotides of the PAM sequence that were greater or equal to 9 bp, to confirm that none of the potential off-target hits were differentially expressed (Cui et al., 2018; Larson et al., 2013).

This study showed rather modest log₂ fold-changes with inhibition of the antisense. This was not unexpected in exponential growth conditions where neither the antisense or the *phoPR* operon are strongly expressed, and we have included a relatively low (+/- 0.5) log₂ fold-change cut-off to evaluate subtle changes in the transcriptome. Genes related to a proposed glutamine transport operon were downregulated upon ATc treatment in the control strain (Mtb_{dcas9_ctrl}) and unaffected by ATc treatment in the antisense-silenced strain (Mtb_{dcas9_sgRNA2}). Glutamine is an essential component of the cell wall of pathogenic mycobacteria and necessary for nitrogen assimilation and adaptation to low pH (Harth et al., 1994; Parveen et al., 2023; Tripathi et al., 2013). The respiratory Complex I NADH dehydrogenase system (nuoNM), which is regulated to maintain redox homeostasis in response to stress (Liang et al., 2023), and two other membrane-associated genes were similarly affected by ATc treatment and antisense silencing.

It is possible that these genes are downregulated in response to stress associated with the undirected *Spy dCas9* in Mtb_{dcas9_ctrl}, which may have more toxic effects than the sgRNA-bound complex in Mtb_{dcas9_sgRNA2} (Rock et al., 2017). Alternatively, ATc itself could potentially cause membrane stress as the molecules diffuse through and interact with the cell membrane (Ehrt et al., 2005; Oliva et al., 1992). However, the strains tested here did not differ in viability after 24 hours, and toxic effects on growth of CRISPRi strains, with and without the sgRNA insert, have not previously been observed in this lab with ATc concentrations of 50-200 ng/mL (Faulkner, 2021), in agreement with other published results (Choudhary et al., 2015; Singh et al., 2016). The low constitutive expression of *phoPR* in exponential growth and the subtle effects of silencing the target, may mean these changes are more obvious than in systems where silencing causes a more impressive transcriptomic response. It is also possible that these changes have been previously

undetected with the more routinely used RT-qPCR assessment with specific genes of interest. If the experiment is repeated in a condition where the *phoPR* regulon is induced, such as low pH, hypoxia or oxidative stress, there may be a more profound effect on the PhoP regulon; and if these are background effects rather than significant interactions, they might become less obvious.

5.6.4 Further Work

There is much to do to further characterise the role of as-phoR in *M. tuberculosis*. Firstly, it would be prudent to confirm the strand-specific CRISPR-mediated inhibition of transcription in *M. tuberculosis* by creating strains with sgRNAs targeting the strand opposite the sgRNA used for targeting as-phoR. These would target the *phoR* sense strand about 260 bp downstream from its start. sgRNAs targeting regions far from the TSS of a gene have been shown to be inefficient inhibitors of transcription, so no effect on sense transcription, or antisense transcription, would be expected.

Repeating the experiment in conditions, such as low pH or hypoxia, which induce the *phoPR* system and increased level of as-phoR expression, may give a more biologically relevant context in which to test the inhibition of as-phoR. Mapping the 3' end of the transcript using 3'RACE (Arnvig & Young, 2009) would be a prerequisite first step to identifying a functional transcript, and should be repeated in relevant conditions to identify any conditionally-processed transcripts. Complementing the CRISPRi strategy described here with an experiment overexpressing the antisense with an exogenous plasmid could be insightful (Li et al., 2022). Another orthogonal approach to confirm as-phoR's direct effect on *phoR* expression might be to overexpress a synthetic sequence complementary to as-phoR with a tet-responsive promoter which could act as a sponge and reduce the ability of as-phoR to interact with *phoR* (Sakurai et al., 2012) and measure *phoR* transcript abundance with RT-qPCR with primers outside the as-phoR-complementary region. Future RT-qPCR experiments should use custom primers to generate strand-specific cDNA and accurately measure sense versus antisense expression. Proteomics assays, such as MS-SWATH, with the CRISPRi and overexpressed strains could be used to determine any differences in the expression of protein products with antisense expression. Half-lives of *phoR* mRNA and as-phoR

could be directly measured by treating with rifampicin to block *de novo* RNA synthesis (Moores et al., 2017). As mentioned earlier, it would be useful to understand the secondary structure of the *phoR* intergenic leader by analysing the cleavage pattern with and without as-phoR RNA (Kipkorir et al., 2024).

Finally, antisense transcription is pervasive in *M. tuberculosis*, and in Chapter 2, RNA-seq-based prediction methods identified over 1100 antisense transcripts. Some of these transcripts overlap 5' UTRs or inter-cistronic UTRs that may contain sORFs as regulatory features. For example, a short transcript that may be part of the 5' UTR of Rv0756c, is antisense to the 5' end of *phoP* and has a TSS at 851736 (Shell et al., 2015) (Figure 5.24). Within the 5' UTR of *phoP* is a 51bp predicted sORF (Sawyer et al., 2021; Smith et al., 2022). Translation of this sORF may regulate *phoP* expression in certain conditions by inhibiting translation, as in other cases in *M. tuberculosis* (Kipkorir et al., 2024), though it lacks an overlapping stop/start codon with the downstream *phoP* ORF. It would be interesting to do a genome-wide investigation of antisense transcripts that overlap predicted sORFs in 5' and inter-cistronic UTRs to look for functional enrichments or other ways of characterising these transcripts and the genes they may regulate.



Figure 5.24. Translation of a sORF found in 5' leader of *phoP* may be regulated by antisense transcript. Figure made with BioRender.com.

5.7 CONCLUSIONS

In this chapter, the expression of a non-coding antisense transcript, antisense-phoR, was verified opposite the membrane-bound kinase member of an important two-component system in *M. tuberculosis*, PhoPR. Expression was successfully inhibited

in exponentially-growing culture using a CRISPR-inhibition system. Differential expression analysis of the CRISPRi strains with untreated controls revealed that many stress response genes were dysregulated in the ATc-treated strains, but several genes were returned to untreated levels of expression upon antisense silencing. These included genes involved in nitrogen assimilation and redox homeostasis, but no known members of the PhoP regulon.

Unexpectedly, expression of the *phoR* transcript was downregulated by 50%. This could be caused by steric inhibition or chromatin interference by the sgRNA:dCas9 complex or have a biologically relevant cause, such as stabilisation of the *phoR* mRNA by the antisense. As translation and mRNA stability are intrinsically linked in mycobacteria, antisense-phoR could potentially impact the translation, and therefore stability, of *phoR* by binding and stabilising mRNA secondary structure and enhancing translation of a sORF within the 5' leader. Translation of the sORF may impact access to a cleavage site and/or by facilitating *phoR* translation by improving the availability of the *phoR* TIS. Dissecting the role of antisense RNA in this well-studied system will be useful for understanding the role of pervasive antisense transcription in the MTBC.

Chapter 6: Conclusion

The host-adapted species of the *Mycobacterium tuberculosis* complex have nearly identical genomes but have adapted to specific host niches. Phylogenomic research has identified large gene losses and the acquisition of virulence gene groups such as toxin/antitoxin systems and genes involved in regulation of lipid metabolism as important steps in the evolution from free-living bacteria to a pathogenic lifestyle (Sapriel et al., 2019). More limited gene losses have led to divergence within the MTBC as the pathogens exploited particular host niches (Brites et al., 2018; Gagneux, 2018). However, the limited genetic differences within the complex may provide enough diversity to support a highly flexible system of post-transcriptional gene regulation that has allowed mycobacteria to adapt and respond to highly variable host extracellular and intracellular environments. In this thesis, this possibility is probed using whole genome assays including transcriptomic analysis and global phenotypic assays.

6.1 Exploring beyond the protein-coding genome

In Chapter 2, a collection of *M. tuberculosis* transcriptomic data in 15 different *in vitro* culture conditions are searched for expressed transcripts outside the known protein coding annotation that could represent non-coding regulators or unannotated short peptides. The abundance and conditional expression of these 'extra-annotated' transcripts, including nearly 1200 mostly ignored antisense transcripts and over 1700 untranslated regions within mRNAs, indicates huge potential for regulatory control that extends far beyond the limited number of short intergenic non-coding RNA and riboswitches that have been explored to date. Ribosome profiling studies have indicated hundreds of unannotated short ORFs in the UTRs of coding genes that are uncharacterised but may be involved in regulating the transcription or translation of downstream genes. It is hard to see how a complete understanding of gene regulation in the MTBC can ignore the potential role of these transcripts. To hint at the functional pathways involving these transcripts, the principle of 'guilt by association' was applied through the creation of a network that clustered both annotated and unannotated expressed transcripts by their co-expression across the range of tested culture conditions. Within the

resulting modules, non-coding transcripts were among the best-connected nodes, indicating their transcription is likely to be co-regulated with the highly connected protein coding genes. Proof of concept was supported by individual modules that include a large proportion of genes from known regulons, such as KstR and DosR. The network is a resource for mycobacterial researchers to find potential ncRNA actors in gene pathways. To facilitate the exploration of modules, a web-app was created which allows users to find ncRNA associated with particular modules, transcripts or genomic coordinates.

6.2 Back to the basics: discovering what is essential in the MTBC

Determining the basic gene requirements for a cell to survive in a particular environmental context began with efforts to determine the 'minimal' genome and the development of global transposon mutagenesis (Hutchison et al., 1999). Tn-seq experiments in *M. tuberculosis* and other bacteria demonstrated that the 'essentiality' of genes was subject to the conditions in which the cell was exposed to. Differences in nutrient availability, extracellular pH, oxygen levels, metal concentrations and drug treatments each produce a different set of 'essential' genes. The particular ecological niche created by the host immune system and inhabited by the members of the MTBC will be different for the human versus animal-adapted species, and consequently, the set of essential genes may also be different. In Chapter 3, parallel tn-seq libraries from human-adapted *M. tuberculosis* and animal-adapted *M. bovis* were analysed to identify differences in the gene requirements for *in vitro* exponential growth between the species. Tn-seq assays are always subject to stochastic factors related to the limited number of insertion sites in any gene region, the number and representation of unique clones sequenced and technical bottlenecks. There are several statistical methods to identify truly essential genes versus genes that may appear essential because the mutants are not represented in the sample sequenced (either because they have a mild growth defect or because of chance) but comparing the results of different tn-seq libraries created in different bacterial species is not straightforward.

The analysis in this chapter presented two different approaches to identify genes that are more required in either *M. tuberculosis* or *M. bovis*. The TRANSIT HMM

method was used for each library independently to determine the essentiality probability for each site depending on the essentiality of the adjacent sites. The results were then compared for orthologous genes in the two species. A quantitative statistical approach, TRANSIT resampling, was also used to determine if there was any statistically significant difference in the mean number of reads within the specified gene region. The results of both methods indicated a number of orthologous genes that code for identical proteins are differently required between the two species, especially those related to intermediary metabolism, lipid metabolism and cell wall processes. One intriguing candidate for further characterisation is a potential transcriptional regulator, Mb1859/Rv1828, which is identical in both genomes, however, only appears to be essential for growth in *M. tuberculosis* and may be involved in nitrogen metabolism. In addition, 15 non-coding RNA transcripts were found to have different essentiality calls with the HMM method.

In Chapter 4, the requirements for *M. bovis* survival were extended to investigate genes essential to survive conditions of oxidative stress—a condition the pathogen will encounter within the phagosomes of mammalian macrophages. Parallel tn-seq libraries grown with and without menadione were compared using TRANSIT resampling. 18 genes were identified that were conditionally-essential in the menadione treated library. These included genes involved in the electron transport chain, lipid metabolism and those associated with membrane integrity. Two genes that were only conditionally essential for *M. bovis* in menadione were *required* for normal *in vitro* growth in *M. tuberculosis* (Chapter 3). *FadD30*, a potential fatty acid ligase (Mb0411/Rv0404) was found to cause a growth defect for *M. tuberculosis in vitro* growth and tolerated fewer insertions than the ortholog in *M. bovis*. Iron transport regulator, *irtA*, showed a difference in essentiality between the human and animal-adapted strains, as well, with *M. bovis* more tolerant of insertions than *M. tuberculosis*. Regulating iron levels is crucial for maintaining redox homeostasis and differences in the regulation of iron transport may be related to known variations in utilisation of heme between the strains.

6.3 Making 'antisense' of the PhoPR two-component system

Chapter 5 uses a more focussed approach, targeting an antisense transcript, identified in Chapter 2, that is transcribed on the opposite strand of an important two-component gene regulatory system acting at the host-pathogen interface. Using tn-seq studies in mammalian hosts, the PhoPR system has been identified as a requirement for virulence in both *M. tuberculosis* and in *M. bovis* (Gibson et al., 2022; Smith, C.M. et al., 2022), despite a SNP in *M. bovis* that causes lower virulence when transferred to *M. tuberculosis* (Gonzalo-Asensio et al., 2014). Using a CRISPRi approach, this transcript, as-phoR, was silenced in *M. tuberculosis* and RNA-seq was used to identify differentially expressed genes. The abundance of *phoR* mRNA was downregulated in the silenced strains, indicating either that the antisense impacts *phoR* abundance, or that the CRISPRi system interferes with transcription on both strands. More experiments are required to rule out the latter scenario, but a model where an antisense RNA stabilises the *phoR* mRNA should be investigated further. The presence of a sORF and regulatory sequences in the UTR between the *phoP* and *phoR* gene that may influence the translation of *phoR* is intriguing. If antisense regulation is essential to this system in *M. tuberculosis*, there may be similar examples elsewhere in the genome. This kind of post-transcriptional regulation could be useful for rapid detection and 'fine-tuning' in the face of changing intracellular conditions.

6.4 Non-coding RNA in host-adapted gene systems

Members of the MTBC have evolved to infect a broad range of hosts, despite the narrow differences in the various host-adapted genomes. Pathogenic mycobacteria have evolved from free-living species to intracellular pathogens and then further adapted to survive and spread among specific mammalian hosts (Gagneux, 2018; Sapriel et al., 2019). Phylogenomics can identify gene deletions, acquisitions and mutations that indicate host leaps among the phyla, but when it comes to comparing the highly similar members of the MTBC, 'the devil is in the details'—such as the little more than 2000 SNPs that differentiate the reference strains of *M. tuberculosis* and *M. bovis* (Bigi et al., 2016; Garnier et al., 2003). A limited number of comparative transcriptomic studies between different strains and species within the MTBC indicate differences in gene expression and regulation, including the use of different

transcriptional start sites and antisense transcription resulting from SNPs (Chiner-Oms et al., 2019; Dinan et al., 2014; Golby et al., 2007, 2013); but the function of these differently expressed non-coding transcripts remains unclear.

The tn-seq work presented in this thesis has also demonstrated a difference in the required genes for unrestricted growth between human-adapted *M. tuberculosis* and the animal-adapted *M. bovis*. Tn-seq studies in culture conditions that replicate intracellular challenges, such as presented here with *M. bovis* oxidative stress, indicate that the essential set of genes is also flexible depending on the external conditions. Constitutive gene expression means that the correlation between essentiality and transcription is not robust but differently required genes may be regulated post-transcriptionally. Non-coding RNA may be involved in adapting protein expression to respond to the unique microenvironment created by the specific host immune system. Transcriptomic and proteomic studies have highlighted differences that indicate the PhoPR system may respond to different signals and regulate different genes, including non-coding RNA (García et al., 2021; Malone et al., 2018; Solans et al., 2014). Therefore, this important virulence system, differing in genomic sequence, expression, and phenotype between the human and animal-adapted lineages, is an attractive system to examine the function of antisense regulation and its possible implications for host-specific regulation.

APPENDIX

Appendix 1. Published works

Stiens, J., Tan, Y. Y., Joyce, R., Arnvig, K. B., Kendall, S. L., & Nobeli, I. (2023). Using a whole genome co-expression network to inform the functional characterisation of predicted genomic elements from *Mycobacterium tuberculosis* transcriptomic data. *Molecular Microbiology*, 119(4), 381-400.

<https://doi.org/https://doi.org/10.1111/mmi.15055>

*Gibson, A. J., *Stiens, J., Passmore, I. J., Faulkner, V., Miculob, J., Willcocks, S., Coad, M., Berg, S., Werling, D., Wren, B. W., Nobeli, I., Villarreal-Ramos, B., & Kendall, S. L. (2022). Defining the Genes Required for Survival of *Mycobacterium bovis* in the Bovine Host Offers Novel Insights into the Genetic Basis of Survival of Pathogenic Mycobacteria. *MBio*, 13(4). <https://doi.org/10.1128/mbio.00672-22> (*co-first authors)

Stiens, J., Arnvig, K. B., Kendall, S. L., & Nobeli, I. (2022). Challenges in defining the functional, non-coding, expressed genome of members of the *Mycobacterium tuberculosis* complex. *Molecular Microbiology*, 117(1), 20–31.

<https://doi.org/10.1111/mmi.14862>

Gibson, A. J., Passmore, I. J., Faulkner, V., Xia, D., Nobeli, I., Stiens, J., Willcocks, S., Clark, T. G., Sobkowiak, B., Werling, D., Villarreal-Ramos, B., Wren, B. W., & Kendall, S. L. (2021). Probing Differences in Gene Essentiality Between the Human and Animal Adapted Lineages of the *Mycobacterium tuberculosis* Complex Using TnSeq. *Frontiers in Veterinary Science*, 8(December), 1–12.

<https://doi.org/10.3389/fvets.2021.760717>

Appendix 2. Chapter 2

A2.1 Supplemental Tables

All supplemental data files are available for download at

<https://zenodo.org/records/13820446>.

A2.2 Supplemental Figures

Figure S1. Dispersion of count data and variance of the mean for non-normalized expression data by sample

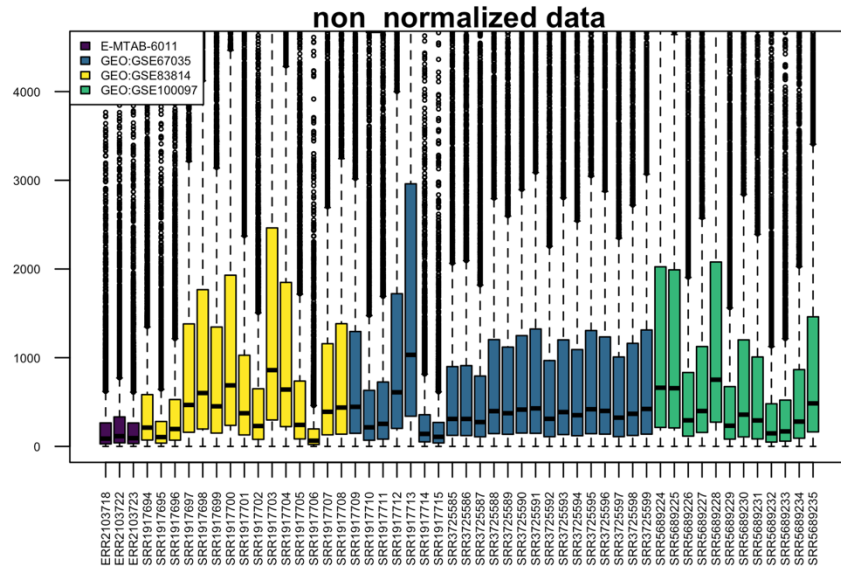


Figure S2. Dispersion of count data and variance around the mean for rlog transformed expression data by sample

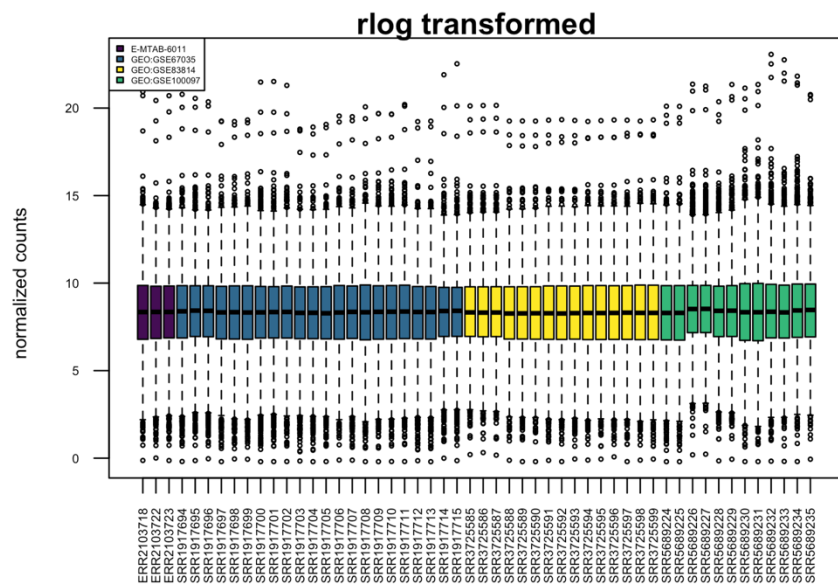


Figure S3. Hierarchical clustering dendrogram of rlog transformed data before limma batch effect correction

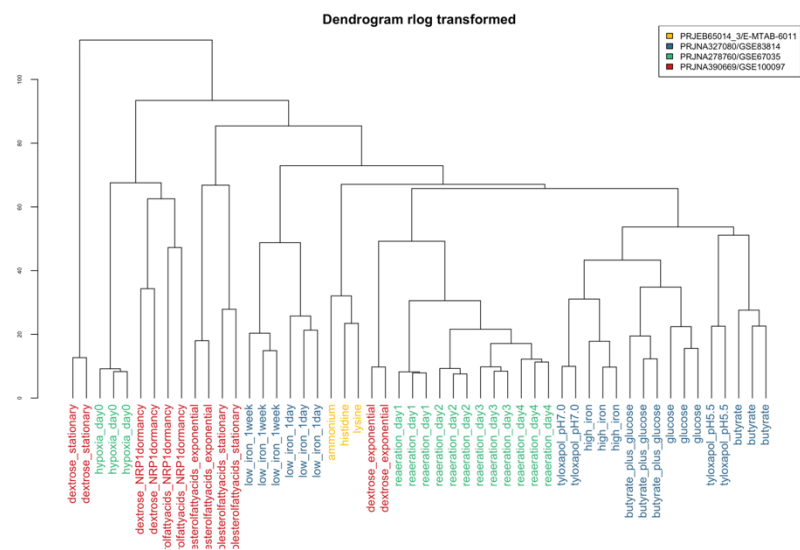


Figure S4. Choice of soft-thresholding power based on scale-free topology model. A) Scale-free topology fit index as a function of the soft-thresholding power and B) mean connectivity as a function of soft-thresholding power.

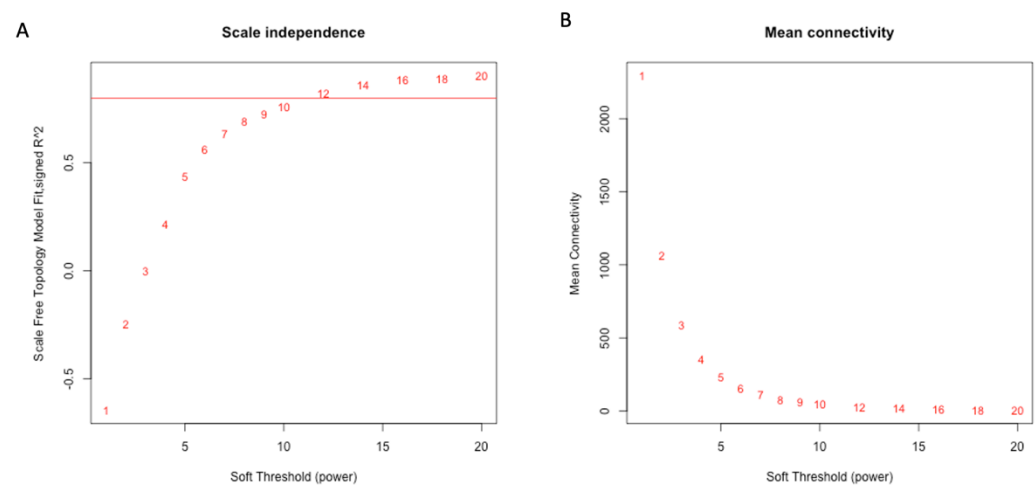


Figure S5. Dispersion of count data and variance around the mean for rlog transformed expression data by sample, filtered for coding regions only

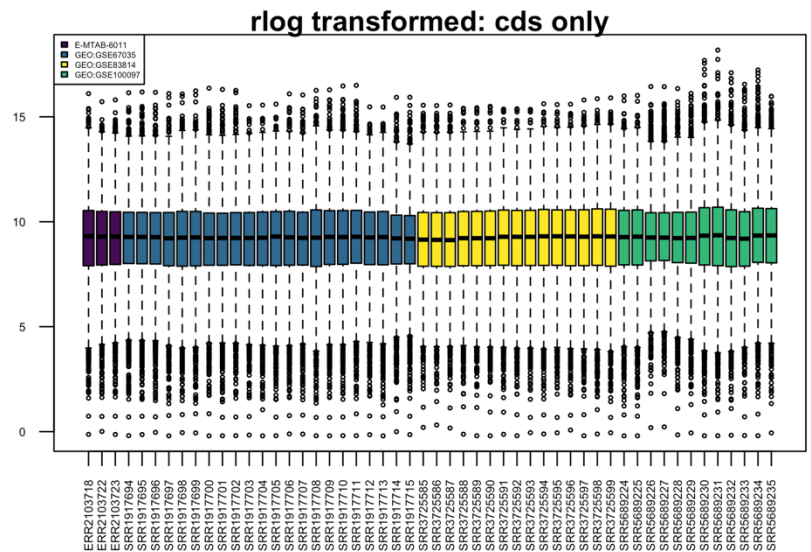


Figure S6. Dispersion of count data and variance around the mean for rlog transformed expression data by sample, filtered for putative sRNAs

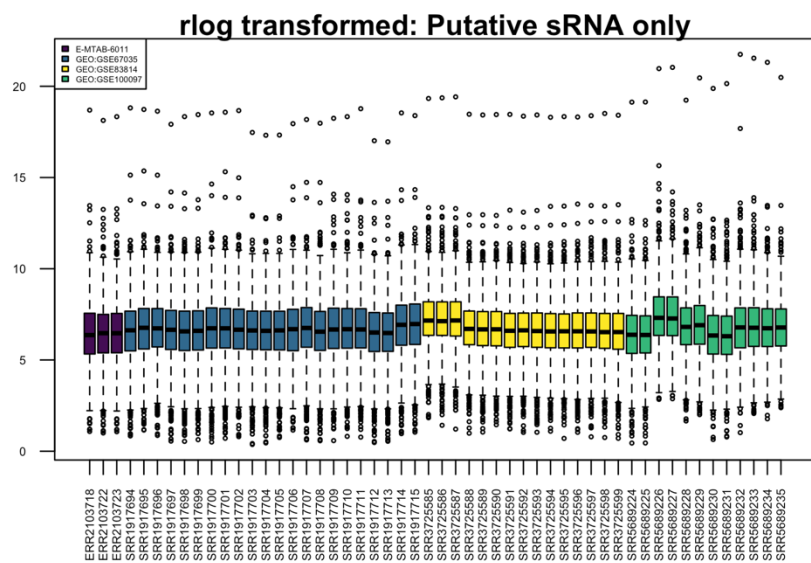


Figure S7. Dispersion of count data and variance around the mean for rlog transformed expression data by sample, filtered for putative UTRs

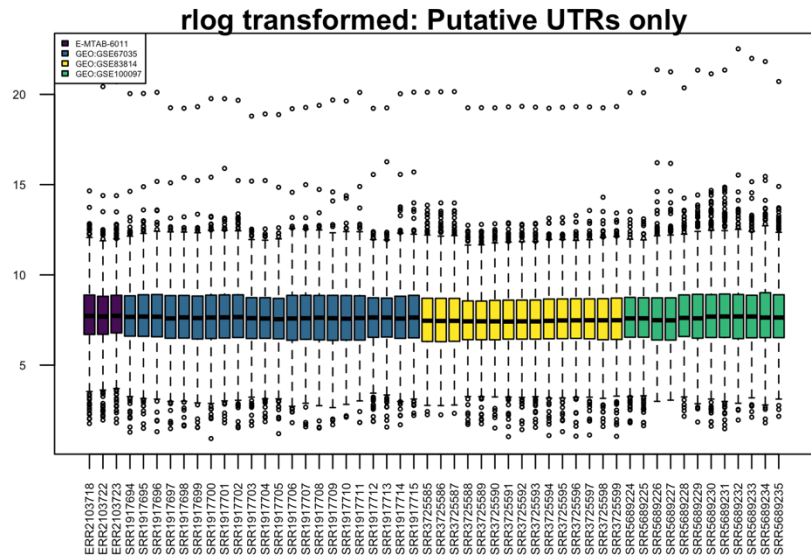


Figure S8. Cluster dendrogram of modules in WGCNA analysis. The network was comprised of 53 modules here indicated by colour. Genes at the tips of the branches are the least connected to the module, and most highly-connected genes form the nodes, or branch-points of the module.

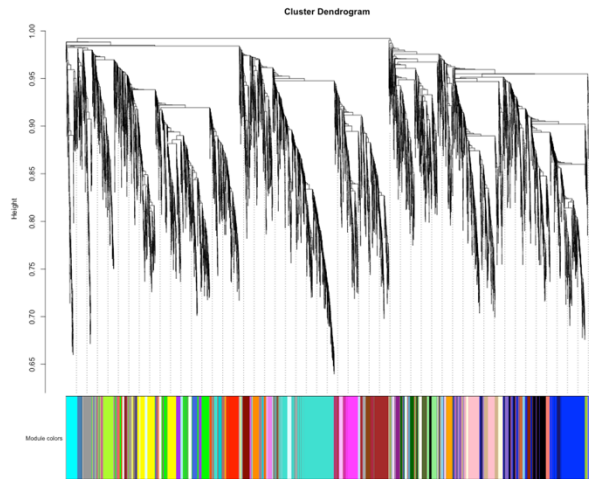


Figure S9. Network of modules based on eigengene adjacency. A) Correlation between module eigengenes represented in a heatmap. Boxes of red along the diagonal indicate clusters of more related modules. B) Cluster diagram demonstrating hierarchical relationship between module eigengenes.

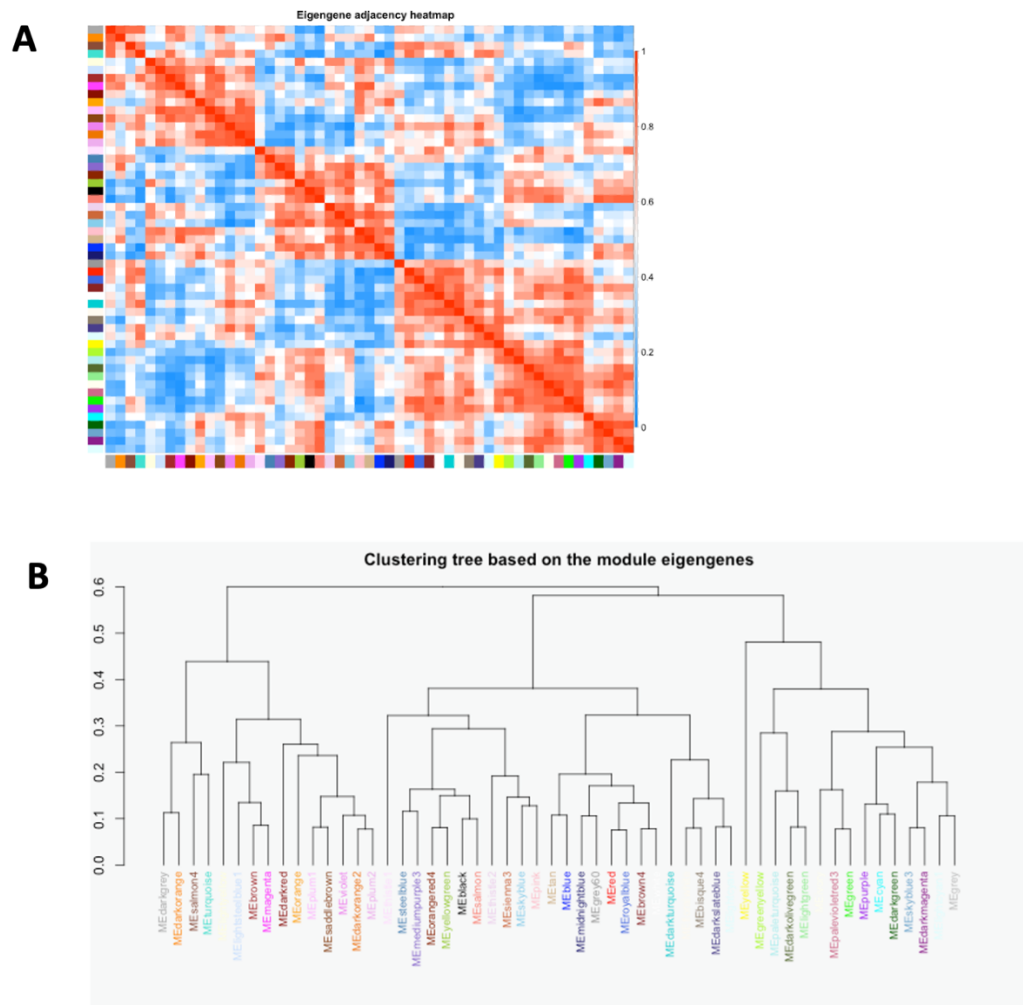
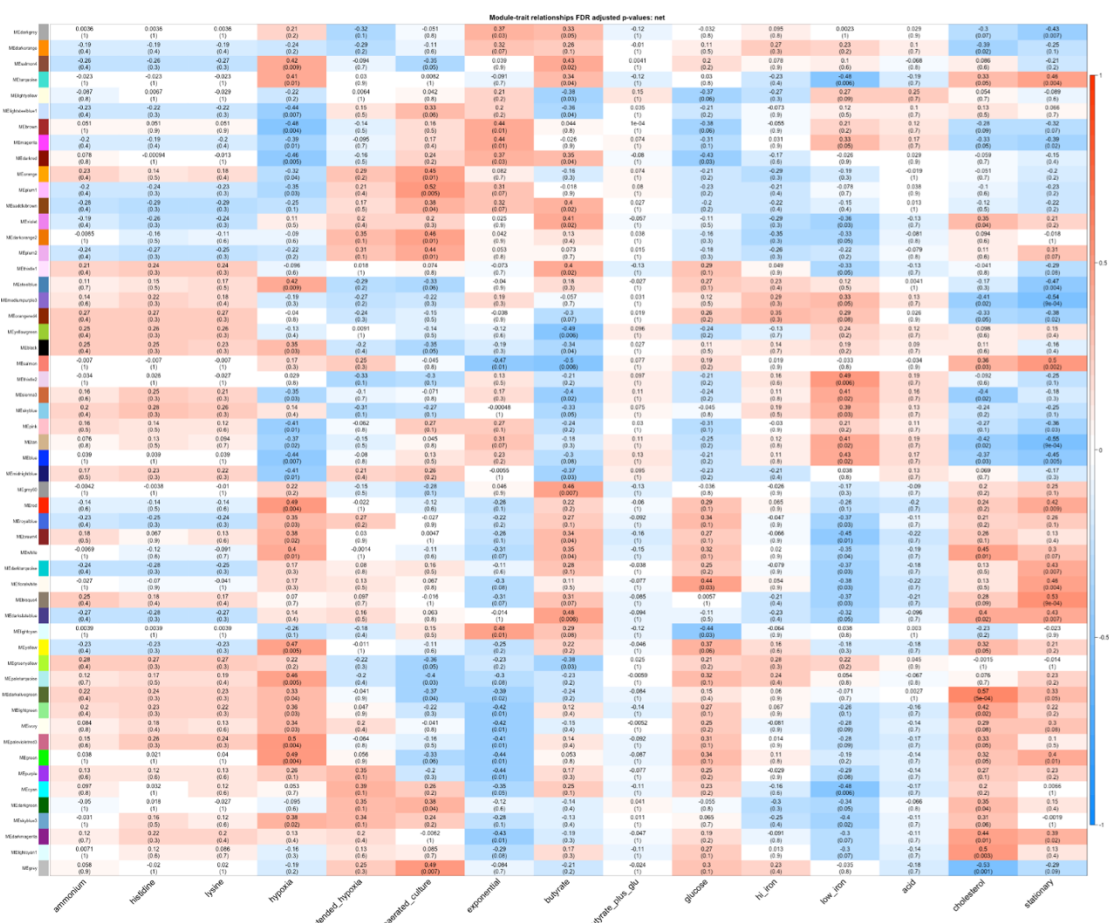


Figure S10. Heat map of correlation of module eigengene (ME) of each module with all experimental conditions. Correlation was calculated using biweight midcorrelation (bicor) and p-values were adjusted for multiple testing (fdr). Positive correlation is red, negative correlation is blue.



Appendix 3. Chapter 3

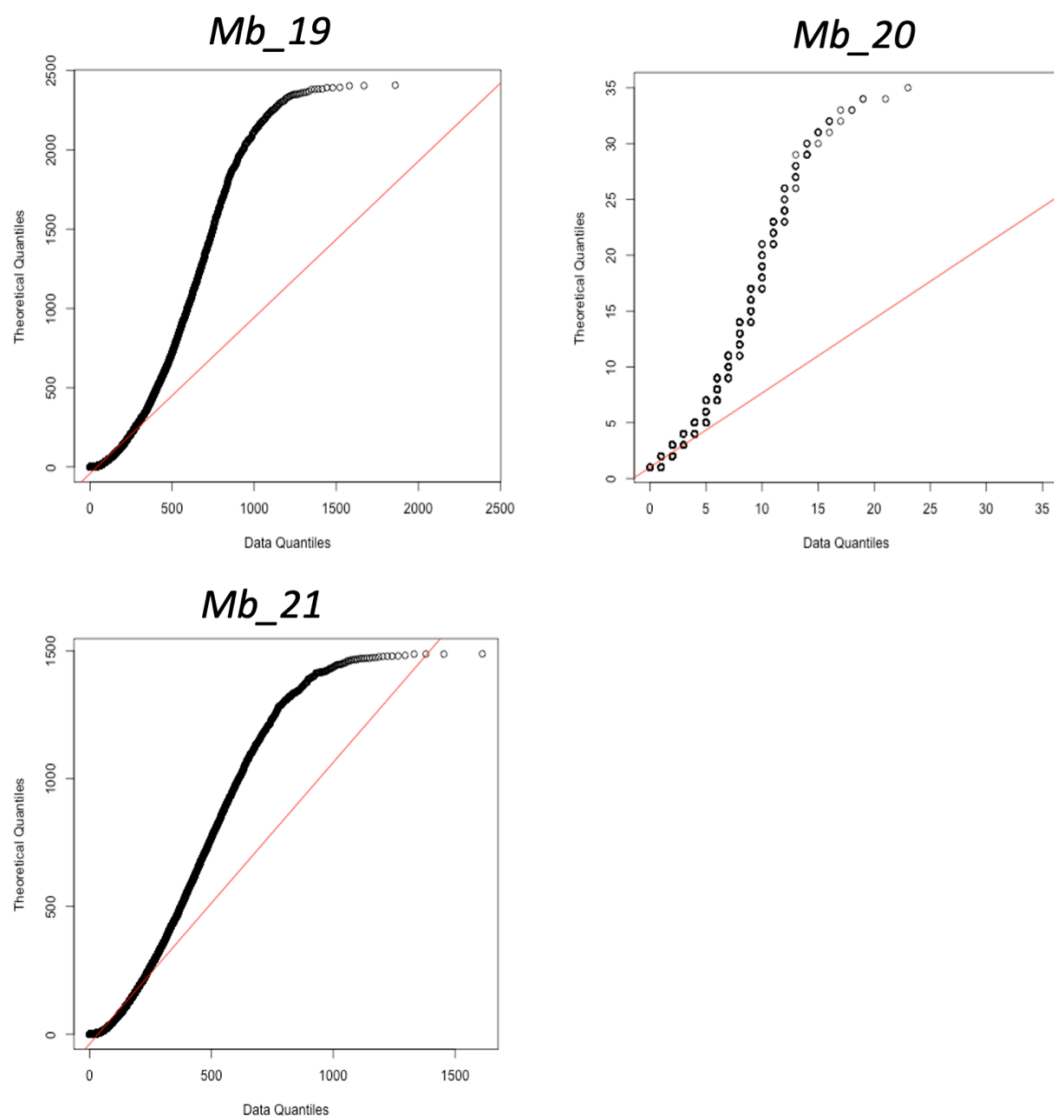
A3.1 Supplemental Tables

All supplemental data files are available for download at:

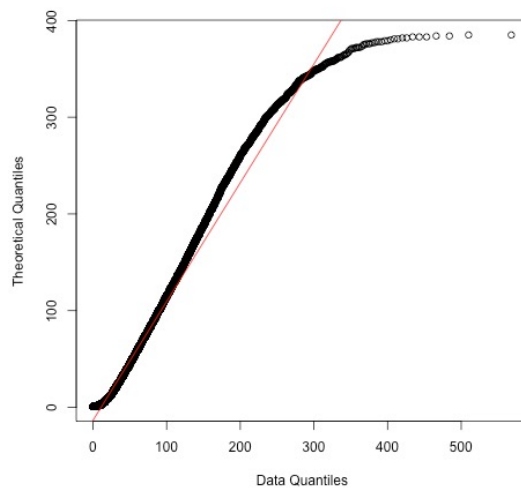
<https://zenodo.org/records/13820446>.

A3.2 Supplemental Figures

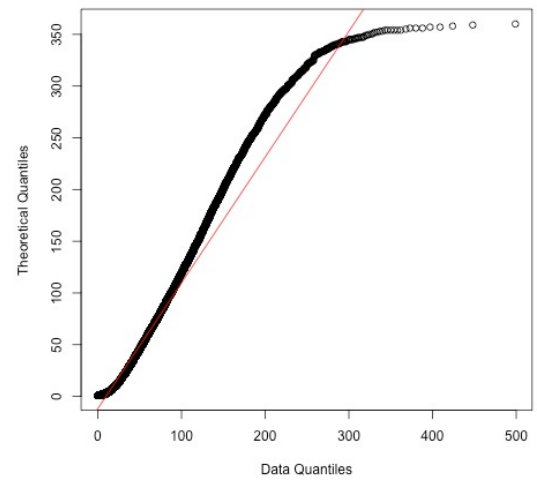
S1 Quartile-quartile plots of geometric distribution against *M. bovis* subsample distributions



S2 Quartile-quartile plots of geometric distribution against *M. tb* replicate distributions



Mtb_22



Mtb_23

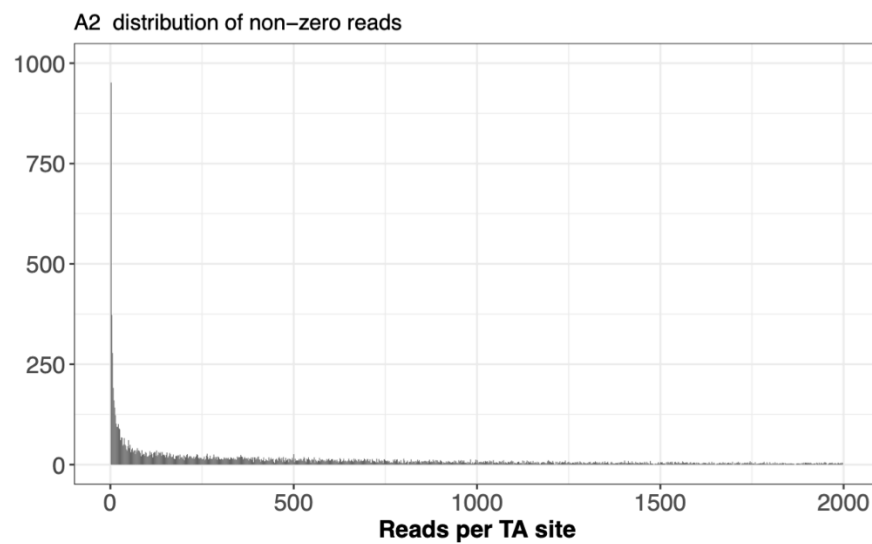
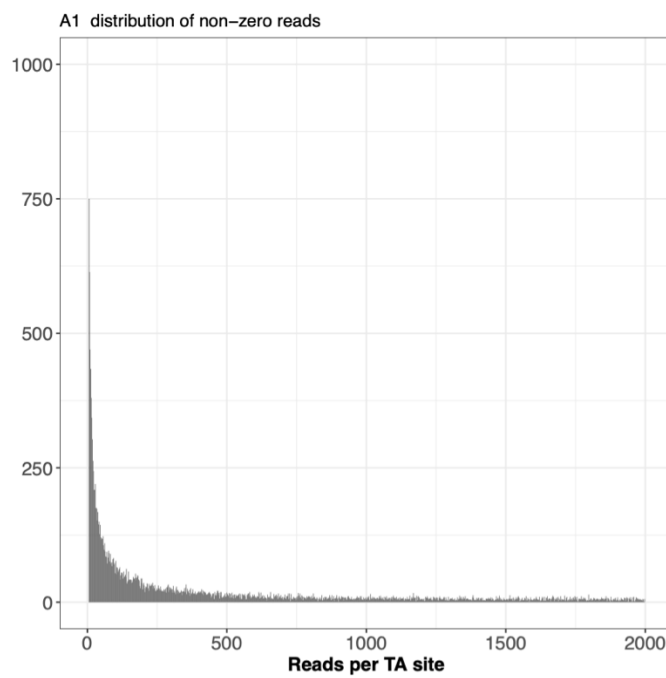
Appendix 4. Chapter 4

A4.1 Supplemental Tables

All supplemental data files are available for download at <https://zenodo.org/records/13820446>.

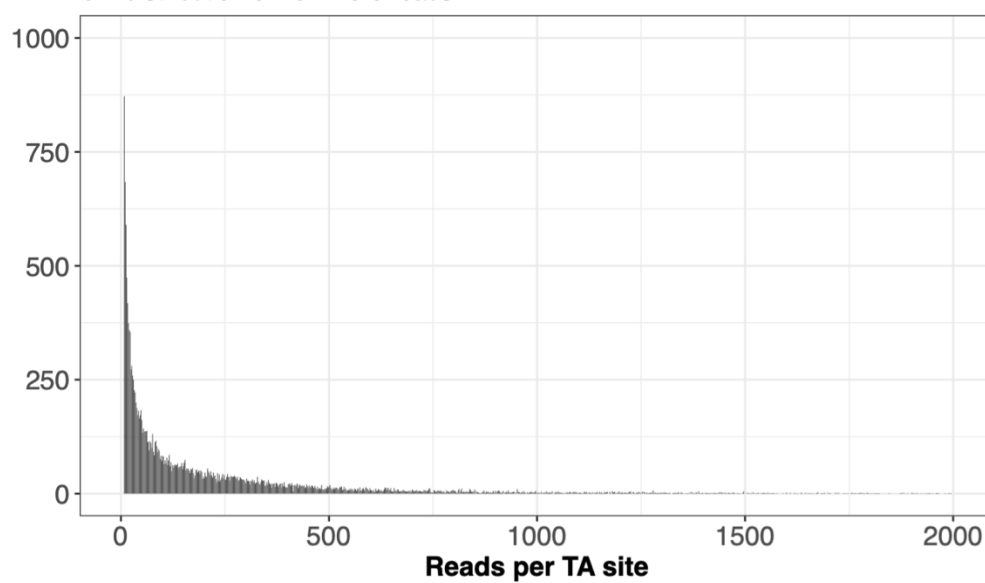
A4.2 Supplemental Figures

Histograms of non-zero read frequency

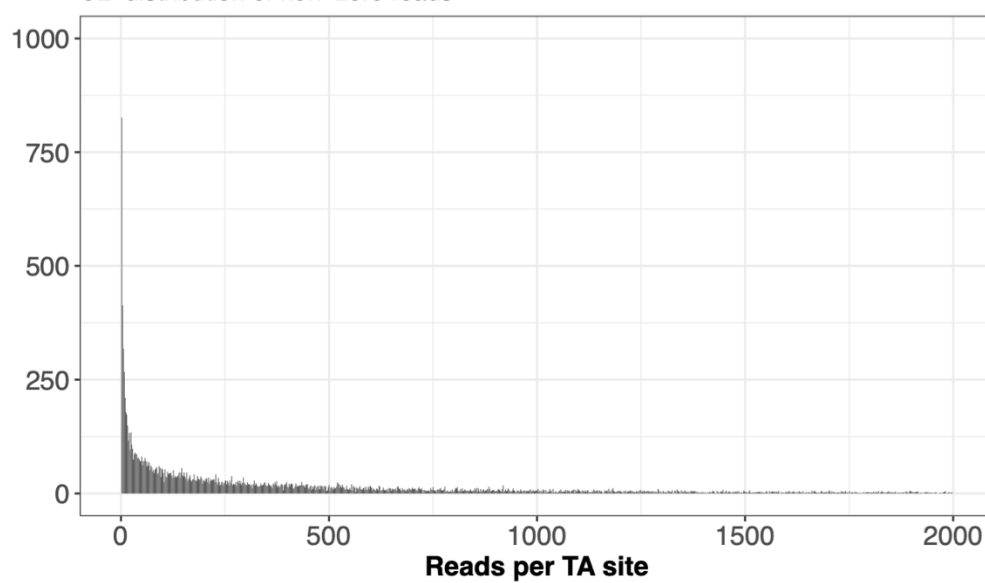


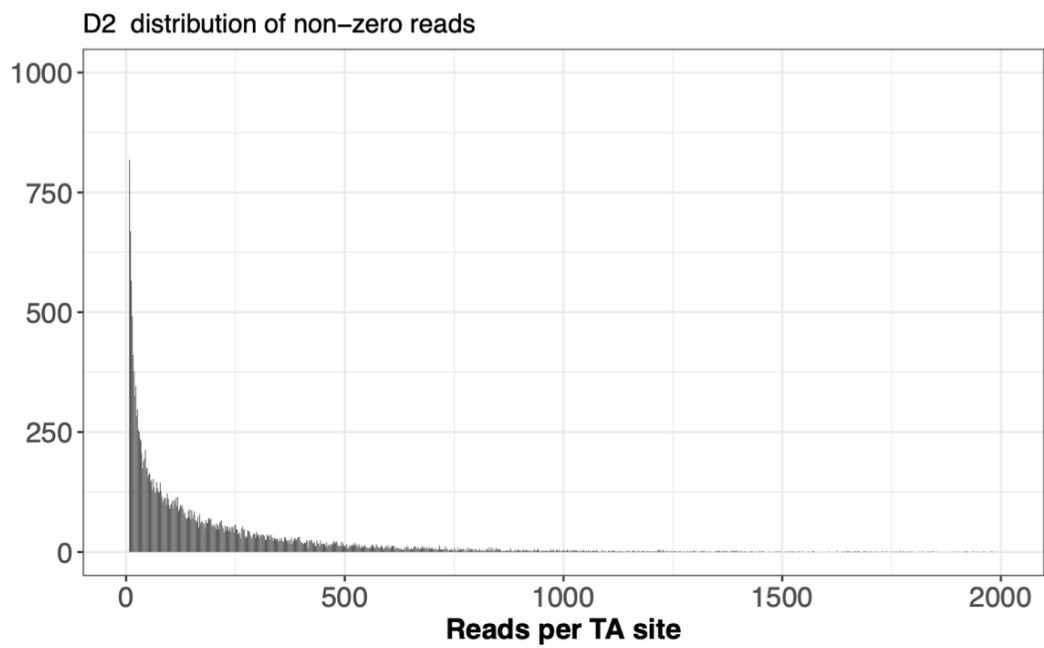


C1 distribution of non-zero reads



C2 distribution of non-zero reads





Appendix 5. Chapter 5

A5.1 Supplemental Tables

All supplemental data files are available for download at

<https://zenodo.org/records/13820446>.

REFERENCES

- Adams, P. P., Baniulyte, G., Esnault, C., Chegiredy, K., Singh, N., Monge, M., Dale, R. K., Storz, G., & Wade, J. T. (2020). Regulatory roles of 5' UTR and ORF-internal RNAs detected by 3' end mapping. *eLife*, 2020.07.18.207399. <https://doi.org/10.1101/2020.07.18.207399>
- Agrawal, R., Pandey, A., Rajankar, M. P., Dixit, N. M., & Saini, D. K. (2015). The two-component signalling networks of *Mycobacterium tuberculosis* display extensive cross-talk in vitro. *Biochemical Journal*, 469(1), 121–134. <https://doi.org/10.1042/BJ20150268>
- Aguilar-Ayala, D. A., Tilleman, L., Van Nieuwerburgh, F., Deforce, D., Palomino, J. C., Vandamme, P., Gonzalez-Y-Merchand, J. A., & Martin, A. (2017). The transcriptome of *Mycobacterium tuberculosis* in a lipid-rich dormancy model through RNAseq analysis. *Scientific Reports*, 7(1), 17665–17665. PubMed. <https://doi.org/10.1038/s41598-017-17751-x>
- Aiso, T., Kamiya, S., Yonezawa, H., & Gamou, S. (2014). Overexpression of an antisense RNA, ArrS, increases the acid resistance of *Escherichia coli*. *Microbiology (United Kingdom)*, 160(PART 5), 954–961. <https://doi.org/10.1099/mic.0.075994-0>
- Alkam, D., Wongsurawat, T., Nookaew, I., Richardson, A. R., Ussery, D., Smeltzer, M. S., & Jenjaroenpun, P. (2021). Is amplification bias consequential in transposon sequencing (Tnseq) assays? A case study with a *staphylococcus aureus* tnseq library subjected to pcr-based and amplification-free enrichment methods. *Microbial Genomics*, 7(10). <https://doi.org/10.1099/mgen.0.000655>
- 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(167), 1–12.
- Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data* [Computer software]. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Arnvig, K. B., Cortes, T., & Young, D. B. (2014). Noncoding RNA in *Mycobacteria*. *Microbiology Spectrum*, 2(2), 1–16. <https://doi.org/10.1128/microbiolspec>

- Arnvig, K. B., & Young, D. B. (2009). Identification of small RNAs in *Mycobacterium tuberculosis*. *Molecular Microbiology*, 73(3), 397–408.
<https://doi.org/10.1111/j.1365-2958.2009.06777.x>
- 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*. *PLOS Pathogens*, 7(11), e1002342. <https://doi.org/10.1371/journal.ppat.1002342>
- Arnvig, K., & Young, D. (2012). Non-coding RNA and its potential role in *Mycobacterium tuberculosis* pathogenesis. *RNA Biology*, 9(4), 427–436.
<https://doi.org/10.4161/rna.20105>
- 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. *Nature Genetics*, 25(1), 25–29. <https://doi.org/10.1038/75556>
- Babu Sait, M. R., Koliwer-Brandl, H., Stewart, J. A., Swarts, B. M., Jacobsen, M., Ioerger, T. R., & Kalscheuer, R. (2022). PPE51 mediates uptake of trehalose across the mycomembrane of *Mycobacterium tuberculosis*. *Scientific Reports*, 12(1), 2097. <https://doi.org/10.1038/s41598-022-06109-7>
- Bailey, M., Christoforidou, Z., & Lewis, M. C. (2013). The evolutionary basis for differences between the immune systems of man, mouse, pig and ruminants. ECMIS 2011 - *Escherichia Coli* and the Mucosal Immune System : Interaction, Modulation and Vaccination, 152(1), 13–19.
<https://doi.org/10.1016/j.vetimm.2012.09.022>
- Baker, J. J., Dechow, S. J., & Abramovitch, R. B. (2019). Acid Fasting: Modulation of *Mycobacterium tuberculosis* Metabolism at Acidic pH. *Trends in Microbiology*, 27(11), 942–953. <https://doi.org/10.1016/j.tim.2019.06.005>
- Baker, J. J., Johnson, B. K., & Abramovitch, R. B. (2014). Slow growth of *Mycobacterium tuberculosis* at acidic pH is regulated by phoPR and host-associated carbon sources. *Molecular Microbiology*, 94(1), 56–69.
<https://doi.org/10.1111/mmi.12688>
- Bansal, R., Anil Kumar, V., Sevalkar, R. R., Singh, P. R., & Sarkar, D. (2017). *Mycobacterium tuberculosis* virulence-regulator PhoP interacts with

- alternative sigma factor SigE during acid-stress response. *Molecular Microbiology*, 104(3), 400–411. <https://doi.org/10.1111/mmi.13635>
- Barquist, L., Mayho, M., Cummins, C., Cain, A. K., Boinett, C. J., Page, A. J., Langridge, G. C., Quail, M. A., Keane, J. A., & Parkhill, J. (2016). The TraDIS toolkit: Sequencing and analysis for dense transposon mutant libraries. *Bioinformatics*, 32(7), 1109–1111. <https://doi.org/10.1093/bioinformatics/btw022>
- Bartens, M., Willcocks, S., Werling, D., & Gibson, A. J. (2024). Respiratory bioenergetics is enhanced in human , but not bovine macrophages after exposure to *M. bovis* PPD : Exploratory insights into overall similar Cellular Metabolic Profiles. *Innate Immunity*. 2024;30(6-8):136-149. <https://doi.org/10.1177/17534259241296630>
- Basaraba, R. J., & Hunter, R. L. (2017). Pathology of Tuberculosis: How the Pathology of Human Tuberculosis Informs and Directs Animal Models. *Microbiology Spectrum*, 5(3), 10.1128/microbiolspec.tbtb2-0029–2016. <https://doi.org/10.1128/microbiolspec.tbtb2-0029-2016>
- Baruzzo, G., Serafini, A., Finotello, F., Sanavia, T., Cioetto-Mazzabò, L., Boldrin, F., Lavezzo, E., Barzon, L., Toppo, S., Provvedi, R., Manganeli, R., & Di Camillo, B. (2023). Role of the Extracytoplasmic Function Sigma Factor SigE in the Stringent Response of *Mycobacterium tuberculosis*. *Microbiology Spectrum*, 11(2). <https://doi.org/10.1128/spectrum.02944-22>
- 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. <https://doi.org/10.1093/molbev/msm111>
- Bellerose, M. M., Proulx, M. K., Smith, C. M., Baker, R. E., Ioerger, T. R., & Sassetti, C. M. (2020). Distinct Bacterial Pathways Influence the Efficacy of Antibiotics against *Mycobacterium tuberculosis*. *mSystems*, 5(4), e00396-20. <https://doi.org/10.1128/mSystems.00396-20>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>

- Bhusal, R. P., Bashiri, G., Kwai, B. X. C., Sperry, J., & Leung, I. K. H. (2017). Targeting isocitrate lyase for the treatment of latent tuberculosis. *Drug Discovery Today*, 22(7), 1008–1016. <https://doi.org/10.1016/j.drudis.2017.04.012>
- Bidnenko, E., & Bidnenko, V. (2018). Transcription termination factor Rho and microbial phenotypic heterogeneity. *Current Genetics*, 64(3), 541–546. <https://doi.org/10.1007/s00294-017-0775-7>
- Bigi, M. M., Blanco, F. C., Araújo, F. R., Thacker, T. C., Zumárraga, M. J., Cataldi, A. A., Soria, M. A., & Bigi, F. (2016). Polymorphisms of 20 regulatory proteins between *Mycobacterium tuberculosis* and *Mycobacterium bovis*. *Microbiology and Immunology*, 60(8), 552–560. <https://doi.org/10.1111/1348-0421.12402>
- Birhanu, A. G., Yimer, S. A., Kalayou, S., Riaz, T., Zegeye, E. D., Holm-Hansen, C., Norheim, G., Aseffa, A., Abebe, M., & Tønjum, T. (2019). Ample glycosylation in membrane and cell envelope proteins may explain the phenotypic diversity and virulence in the *Mycobacterium tuberculosis* complex. *Scientific Reports*, 9(1), 2927. <https://doi.org/10.1038/s41598-019-39654-9>
- Black, K. A., Duan, L., Mandyoli, L., Selbach, B. P., Xu, W., Ehrt, S., Sacchettini, J. C., & Rhee, K. Y. (2021). Metabolic bifunctionality of Rv0812 couples folate and peptidoglycan biosynthesis in *Mycobacterium tuberculosis*. *Journal of Experimental Medicine*, 218(7), e20191957. <https://doi.org/10.1084/jem.20191957>
- Blighe K, Lun A. (2024). *PCAtools: PCAtools: Everything Principal Components Analysis* (Version 2.16.0) [R]. <https://github.com/kevinblighe/PCAtools>
- Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., Marchler-Bauer, A., ... Finn, R. D. (2020). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, 49(D1), D344–D354. <https://doi.org/10.1093/nar/gkaa977>
- Boradia, V., Frando, A., & Grundner, C. (2022). The *Mycobacterium tuberculosis* PE15/PPE20 complex transports calcium across the outer membrane. *PLOS Biology*, 20(11), e3001906. <https://doi.org/10.1371/journal.pbio.3001906>

- Brenner, E. P., & Sreevatsan, S. (2023). Global-scale GWAS associates a subset of SNPs with animal-adapted variants in *M. tuberculosis* complex. *BMC Medical Genomics*, 16(1), 260. <https://doi.org/10.1186/s12920-023-01695-5>
- Brites, D., Loiseau, C., Menardo, F., Borrell, S., Boniotti, M. B., Warren, R., Dippenaar, A., Parsons, S. D. C., Beisel, C., Behr, M. A., Fyfe, J. A., Coscolla, M., & Gagneux, S. (2018). A new phylogenetic framework for the animal-adapted *mycobacterium tuberculosis* complex. *Frontiers in Microbiology*, 9(NOV), 1–14. <https://doi.org/10.3389/fmicb.2018.02820>
- Budell, W. C., Germain, G. A., Janisch, N., McKie-Krisberg, Z., Jayaprakash, A. D., Resnick, A. E., & Quadri, L. E. N. (2020). Transposon mutagenesis in *Mycobacterium kansasii* links a small RNA gene to colony morphology and biofilm formation and identifies 9,885 intragenic insertions that do not compromise colony outgrowth. *MicrobiologyOpen*, 9(4), 1–22. <https://doi.org/10.1002/mbo3.988>
- Butler, R. E., Smith, A. A., Mendum, T. A., Chandran, A., Wu, H., Lefrancois, L., Chambers, M., Soldati, T., & Stewart, G. R. (2020). *Mycobacterium bovis* uses the ESX-1 Type VII secretion system to escape predation by the soil-dwelling amoeba *Dictyostelium discoideum*. *The ISME Journal*, 14(4), 19–930. <https://doi.org/10.1038/s41396-019-0572-z>
- Cain, A. K., Barquist, L., Goodman, A. L., Paulsen, I. T., Parkhill, J., & van Opijnen, T. (2020). A decade of advances in transposon-insertion sequencing. *Nature Reviews Genetics* 21(9), 526–540. <https://doi.org/10.1038/s41576-020-0244-x>
- Carey, A. F., Rock, J. M., Krieger, I. V., Chase, M. R., Fernandez-Suarez, M., Gagneux, S., Sacchettini, J. C., Ioerger, T. R., & Fortune, S. M. (2018). TnSeq of *Mycobacterium tuberculosis* clinical isolates reveals strain-specific antibiotic liabilities. *PLoS Pathogens*, 14(3), e1006939–e1006939. <https://doi.org/10.1371/journal.ppat.1006939>
- Carver, T., Harris, S. R., Berriman, M., Parkhill, J., & McQuillan, J. A. (2012). Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, 28(4), 464–469. <https://doi.org/10.1093/bioinformatics/btr703>
- Castro, F. A. V., Mariani, D., Panek, A. D., Eleutherio, E. C. A., & Pereira, M. D. (2008). Cytotoxicity mechanism of two naphthoquinones (menadione and

- plumbagin) in *Saccharomyces cerevisiae*. *PLoS ONE*, 3(12).
<https://doi.org/10.1371/journal.pone.0003999>
- Chakravarty, S., & Massé, E. (2019). RNA-Dependent Regulation of Virulence in Pathogenic Bacteria. *Frontiers in Cellular and Infection Microbiology* 9.
<https://doi.org/10.3389/fcimb.2019.00337>
- Chao, M. C., Abel, S., Davis, B. M., & Waldor, M. K. (2016). The Design and Analysis of Transposon-Insertion Sequencing Experiments. *Nat Rev Microbiol*, 14(2), 119–128. <https://doi.org/10.1038/nrmicro.2015.7>
- Chao, Y., Papenfort, K., Reinhardt, R., Sharma, C. M., & Vogel, J. (2012). An atlas of Hfq-bound transcripts reveals 3' UTRs as a genomic reservoir of regulatory small RNAs. *The EMBO Journal*, 31(20), 4005–4019.
<https://doi.org/10.1038/emboj.2012.229>
- Chao, Y., & Vogel, J. (2016). A 3' UTR-Derived Small RNA Provides the Regulatory Noncoding Arm of the Inner Membrane Stress Response. *Molecular Cell*, 61(3), 352–363. <https://doi.org/10.1016/j.molcel.2015.12.023>
- Chen, J., Fruhauf, A., Fan, C., Ponce, J., Ueberheide, B., Bhabha, G., & Ekiert, D. C. (2023). Structure of an endogenous mycobacterial MCE lipid transporter. *Nature*, 620(7973), 445–452. <https://doi.org/10.1038/s41586-023-06366-0>
- Chen, J., & Xie, J. (2011). Role and regulation of bacterial LuxR-like regulators. *Journal of Cellular Biochemistry*, 112(10), 2694–2702.
<https://doi.org/10.1002/jcb.23219>
- Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884–i890.
<https://doi.org/10.1093/bioinformatics/bty560>
- Chen, X., Chen, J., Yan, B., Zhang, W., Guddat, L. W., Liu, X., & Rao, Z. (2020). Structural basis for the broad substrate specificity of two acyl-CoA dehydrogenases FadE5 from mycobacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 117(28), 16324–16332.
<https://doi.org/10.1073/pnas.2002835117>
- Chetal, K., & Janga, S. C. (2015). OperomeDB: A Database of Condition-Specific Transcription Units in Prokaryotic Genomes. *BioMed Research International*, 2015, 318217–318217. <https://doi.org/10.1155/2015/318217>

- Chiner-Oms, Á., Berney, M., Boinett, C., González-Candelas, F., Young, D. B., Gagneux, S., Jacobs, W. R., Parkhill, J., Cortes, T., & Comas, I. (2019). Genome-wide mutational biases fuel transcriptional diversity in the *Mycobacterium tuberculosis* complex. *Nature Communications*, *10*(1), 1–11. <https://doi.org/10.1038/s41467-019-11948-6>
- Choudhary, E., Thakur, P., Pareek, M., & Agarwal, N. (2015). Gene silencing by CRISPR interference in mycobacteria. *Nature Communications*, *6*. <https://doi.org/10.1038/ncomms7267>
- Cimino, M., Thomas, C., Namouchi, A., Dubrac, S., Gicquel, B., & Gopaul, D. N. (2012). Identification of DNA Binding Motifs of the *Mycobacterium tuberculosis* PhoP/PhoR Two-Component Signal Transduction System. *PLoS ONE*, *7*(8), e42876. <https://doi.org/10.1371/journal.pone.0042876>
- 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., ... Barrell, B. G. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, *393*(6685), 537–544. <https://doi.org/10.1038/31159>
- Cortes, T., Schubert, O. T., Rose, G., Arnvig, K. B., Comas, I., Aebersold, R., & Young, D. B. (2013). Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*. *Cell Reports*, *5*(4), 1121–1131. <https://doi.org/10.1016/j.celrep.2013.10.031>
- Cui, L., Vigouroux, A., Rousset, F., Varet, H., Khanna, V., & Bikard, D. (2018). A CRISPRi screen in *E. coli* reveals sequence-specific toxicity of dCas9. *Nature Communications*, *9*(1), 1912. <https://doi.org/10.1038/s41467-018-04209-5>
- Cumming, B. M., Lamprecht, D. A., Wells, R. M., Saini, V., Mazorodze, J. H., & Steyn, A. J. C. (2014). The Physiology and Genetics of Oxidative Stress in Mycobacteria. *Microbiology Spectrum*, *2*(3). <https://doi.org/10.1128/microbiolspec.mgm2-0019-2013>
- Damen Merel P. M., Meijers Aniek S., Keizer Esther M., Piersma Sander R., Jiménez Connie R., Kuijl Coenraad P., Bitter Wilbert, & Houben Edith N. G. (2022). The ESX-1 Substrate PPE68 Has a Key Function in ESX-1-Mediated

- Secretion in *Mycobacterium marinum*. *mBio*, 13(6), e02819-22.
<https://doi.org/10.1128/mbio.02819-22>
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), giab008.
<https://doi.org/10.1093/gigascience/giab008>
- Dar, D., Shamir, M., Mellin, J. R., Koutero, M., Stern-Ginossar, N., Cossart, P., & Sorek, R. (2016). Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science*, 352(6282), aad9822.
<https://doi.org/10.1126/science.aad9822>
- Dar, D., & Sorek, R. (2018). Bacterial noncoding RNAs excised from within protein-coding transcripts. *mBio*, 9(5), 1730–18.
<https://doi.org/10.1128/mBio.01730-18>
- Darling, A. E., Mau, B., & Perna, N. T. (2010). Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, 5(6).
<https://doi.org/10.1371/journal.pone.0011147>
- Dawson, C. C., Cummings, J. E., Starkey, J. M., & Slayden, R. A. (2022). Discovery of a novel type RelBE toxin-antitoxin system in *Mycobacterium tuberculosis* defined by co-regulation with an antisense RNA. *Molecular Microbiology*, 117(6), 1419–1433. <https://doi.org/10.1111/mmi.14917>
- De Maio, F., Berisio, R., Manganelli, R., & Delogu, G. (2020). PE_PGRS proteins of *Mycobacterium tuberculosis*: A specialized molecular task force at the forefront of host–pathogen interaction. *Virulence*, 11(1), 898–915.
<https://doi.org/10.1080/21505594.2020.1785815>
- Dechow, S. J., Baker, J. J., Murto, M. R., & Abramovitch, R. B. (2021). Ppe51 variants promote non-replicating *Mycobacterium tuberculosis* to grow at acidic pH by selectively promoting glycerol uptake. *Journal of Bacteriology*, 2021.05.19.444820. <https://doi.org/10.1128/jb.00212-22>
- DeJesus, M. A., Ambadipudi, C., Baker, R., Sassetti, C., & Ioerger, T. R. (2015). TRANSIT - A Software Tool for Himar1 TnSeq Analysis. *PLoS Computational Biology*, 11(10), 1–17. <https://doi.org/10.1371/journal.pcbi.1004401>
- 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. <https://doi.org/10.1128/mBio.02133-16>
- DeJesus, M. A., & Ioerger, T. R. (2013). A Hidden Markov Model for identifying essential and growth-defect regions in bacterial genomes from transposon insertion sequencing data. *BMC Bioinformatics*, 14(1). <https://doi.org/10.1186/1471-2105-14-303>
- DeJesus, M. A., & Ioerger, T. R. (2015). Reducing type I errors in Tn-Seq experiments by correcting the skew in read count distributions. *Proceedings of the 7th International Conference on Bioinformatics and Computational Biology, BICOB 2015*, 45–50.
- DeJesus, M. A., Zhang, Y. J., Sassetti, C. M., Rubin, E. J., Sacchettini, J. C., & Ioerger, T. R. (2013). Bayesian analysis of gene essentiality based on sequencing of transposon insertion libraries. *Bioinformatics*, 29(6), 695–703. <https://doi.org/10.1093/bioinformatics/btt043>
- Del Portillo, P., García-Morales, L., Menéndez, M. C., Anzola, J. M., Rodríguez, J. G., Helguera-Repetto, A. C., Ares, M. A., Prados-Rosales, R., Gonzalez-y-Merchand, J. A., & García, M. J. (2019). Hypoxia Is Not a Main Stress When *Mycobacterium tuberculosis* Is in a Dormancy-Like Long-Chain Fatty Acid Environment. *Frontiers in Cellular and Infection Microbiology*, 8, 449–449.
- DeLoughery, A., Lalanne, J.-B., Losick, R., & Li, G.-W. (2018). Maturation of polycistronic mRNAs by the endoribonuclease RNase Y and its associated Y-complex in *Bacillus subtilis*. *Proceedings of the National Academy of Sciences*, 115(24), E5585 LP-E5594. <https://doi.org/10.1073/pnas.1803283115>
- Deng, J., Bi, L., Zhou, L., Guo, S., Fleming, J., Jiang, H., Zhou, Y., Gu, J., Zhong, Q., Wang, Z., Liu, Z., Deng, R., Gao, J., Chen, T., Li, W., Wang, J., Wang, X., Li, H., Ge, F., ... Zhang, X. E. (2014). *Mycobacterium Tuberculosis* Proteome Microarray for Global Studies of Protein Function and Immunogenicity. *Cell Reports*, 9(6), 2317–2329. <https://doi.org/10.1016/j.celrep.2014.11.023>
- Denman RB. (1993). Using RNAFOLD to predict the activity of small catalytic RNAs. *Biotechniques*, 15(6), 1090–1095.
- Desgranges, E., Barrientos, L., & Caldelari, I. (2021). The 3'UTR-derived sRNA RsaG coordinates redox homeostasis and metabolism adaptation in response to glucose-6-phosphate uptake in *Staphylococcus aureus*. *Molecular Microbiology*. 117(1), 193-214. <https://doi.org/10.1111/MMI.14845>

- D'Halluin, A., Polgar, P., Kipkorir, T., Patel, Z., Cortes, T., & Arnvig, K. B. (2023). Premature termination of transcription is shaped by Rho and translated uORFS in *Mycobacterium tuberculosis*. *iScience*, 26(4).
<https://doi.org/10.1016/j.isci.2023.106465>
- 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.
<https://doi.org/10.1093/nar/gkq101>
- Dinan, A. M., Tong, Pin, Lohan, Amanda J., Conlon, Kevin M., Miranda-CasoLuengo Aleksandra A., Malone, Kerri M., Gordon, Stephen V., & Loftus, Brendan J. (2014). Relaxed Selection Drives a Noisy Noncoding Transcriptome in Members of the *Mycobacterium tuberculosis* Complex. *mBio*, 5(4), e01169-14. <https://doi.org/10.1128/mBio.01169-14>
- Dragset, M. S., Iøerger, T. R., Zhang, Y. J., Mærk, M., Ginbot, Z., Sacchettini, J. C., Flo, T. H., Rubin, E. J., & Steigedal, M. (2019). Genome-wide Phenotypic Profiling Identifies and Categorizes Genes Required for Mycobacterial Low Iron Fitness. *Scientific Reports*, 9(1), 1–11. <https://doi.org/10.1038/s41598-019-47905-y>
- Du, P., Sohaskey, C. D., & Shi, L. (2016). Transcriptional and physiological changes during *Mycobacterium tuberculosis* reactivation from non-replicating persistence. *Frontiers in Microbiology*, 7(AUG).
<https://doi.org/10.3389/fmicb.2016.01346>
- Durand, S., Tomasini, A., Braun, F., Condon, C., & Romby, P. (2015). sRNA and mRNA turnover in Gram-positive bacteria. *FEMS Microbiology Reviews*, 39(3), 316–330. <https://doi.org/10.1093/femsre/fuv007>
- Dutta, D. (2018). Advance in Research on *Mycobacterium tuberculosis* FabG4 and Its Inhibitor. *Frontiers in Microbiology*, 9(June).
<https://www.frontiersin.org/article/10.3389/fmicb.2018.01184>
- Ehrt, S., Guo, X. V., Hickey, C. M., Ryou, M., Monteleone, M., Riley, L. W., & Schnappinger, D. (2005). Controlling gene expression in mycobacteria with anhydrotetracycline and Tet repressor. *Nucleic Acids Research*, 33(2), 1–11. <https://doi.org/10.1093/nar/gni013>

- Ellis, M. J., & Haniford, D. B. (2016). Riboregulation of bacterial and archaeal transposition. *WIREs RNA*, 7(3), 382–398.
<https://doi.org/10.1002/wrna.1341>
- Eoh, H., Wang, Z., Layre, E., Rath, P., Morris, R., Branch Moody, D., & Rhee, K. Y. (2017). Metabolic anticipation in *Mycobacterium tuberculosis*. *Nature Microbiology*, 2(8), 17084. <https://doi.org/10.1038/nmicrobiol.2017.84>
- Faulkner, V. (2021). *The identification of novel therapeutic targets for Mycobacteria using CRISPR/dCas9 genome interference technology*. [Doctoral dissertation, Royal Veterinary College]
- Fay, M. P. (2009). Confidence intervals that match Fisher's exact or Blaker's exact tests. *Biostatistics*, 11(2), 373–374.
<https://doi.org/10.1093/biostatistics/kxp050>
- Feng, L., Chen, S., & Hu, Y. (2018). PhoPR positively regulates *whiB3* expression in response to low pH in pathogenic mycobacteria. *Journal of Bacteriology*, 200(8). <https://doi.org/10.1128/JB.00766-17>
- Fishbein, S., van Wyk, N., Warren, R. M., & Sampson, S. L. (2015). Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Molecular Microbiology*, 96(5), 901–916.
<https://doi.org/10.1111/mmi.12981>
- Forrellad, M. A., McNeil, M., Santangelo, M. D. L. P., Blanco, F. C., García, E., Klepp, L. I., Huff, J., Niederweis, M., Jackson, M., & Bigi, F. (2014). Role of the Mce1 transporter in the lipid homeostasis of *Mycobacterium tuberculosis*. *Tuberculosis*, 94(2), 170–177. <https://doi.org/10.1016/j.tube.2013.12.005>
- Gagneux, S. (2018). Ecology and evolution of *Mycobacterium tuberculosis*. *Nature Reviews Microbiology*, 16(4), 202–213.
<https://doi.org/10.1038/nrmicro.2018.8>
- Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., & Koonin, E. V. (2021). COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research*, 49(D1), D274–D281. <https://doi.org/10.1093/nar/gkaa1018>
- García, E. A., Blanco, F. C., Bigi, M. M., Vazquez, C. L., Forrellad, M. A., Rocha, R. V., Golby, P., Soria, M. A., & Bigi, F. (2018). Characterization of the two component regulatory system PhoPR in *Mycobacterium bovis*. *Veterinary*

- Microbiology*, 222(May), 30–38.
<https://doi.org/10.1016/j.vetmic.2018.06.016>
- García, E. A., Blanco, F. C., Klepp, L. I., Pazos, A., McNeil, M. R., Jackson, M., & Bigi, F. (2021). Role of PhoPR in the response to stress of *Mycobacterium bovis*. *Comparative Immunology, Microbiology and Infectious Diseases*, 74(November 2020), 101593.
<https://doi.org/10.1016/j.cimid.2020.101593>
- Garnier, T., Eiglmeier, K., Camus, J.-C., Medina, N., Mansoor, H., Pryor, M., Duthoy, S., Grondin, S., Lacroix, C., Monsempe, C., Simon, S., Harris, B., Atkin, R., Doggett, J., Mayes, R., Keating, L., Wheeler, P. R., Parkhill, J., Barrell, B. G., ... Hewinson, R. G. (2003). The complete genome sequence of *Mycobacterium bovis*. *Proceedings of the National Academy of Sciences*, 100(13), 7877–7882.
<https://doi.org/10.1073/pnas.1130426100>
- Geneva: World Health Organisation. (2023). *Global Tuberculosis Report 2023*.
 Geneva: World Health Organisation.
<https://www.who.int/publications/i/item/9789240083851>
- Georg, J., & Hess, W. R. (2018). Widespread Antisense Transcription in Prokaryotes. *Microbiology Spectrum*, 6(4).
<https://doi.org/10.1128/microbiolspec.rwr-0029-2018>
- Gerrick, E. R. (2018). *Discovery of Small RNAs and Characterization of Their Regulatory Roles in Mycobacterium Tuberculosis* [PhD Thesis, Harvard University, Graduate School of Arts & Sciences].
<https://dash.harvard.edu/handle/1/41129159>
- Gerrick, E. R., Barbier, T., Chase, M. R., Xu, R., François, 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.
<https://doi.org/10.1073/pnas.1718003115>
- Getahun, H., Matteelli, A., Chaisson, R. E., & Ravigliione, M. (2015). Latent *Mycobacterium tuberculosis* Infection. *New England Journal of Medicine*, 372(22), 2127–2135. <https://doi.org/10.1056/NEJMra1405427>
- Gibson, A. J., Passmore, I. J., Faulkner, V., Xia, D., Nobeli, I., Stiens, J., Willcocks, S., Clark, T. G., Sobkowiak, B., Werling, D., Villarreal-Ramos, B., Wren, B. W., &

- Kendall, S. L. (2021). Probing Differences in Gene Essentiality Between the Human and Animal Adapted Lineages of the *Mycobacterium tuberculosis* Complex Using TnSeq. *Frontiers in Veterinary Science*, 8(December), 1–12. <https://doi.org/10.3389/fvets.2021.760717>
- Gibson, A. J., Stiens, J., Passmore, I. J., Faulkner, V., Miculob, J., Willcocks, S., Coad, M., Berg, S., Werling, D., Wren, B. W., Nobeli, I., Villarreal-Ramos, B., & Kendall, S. L. (2022). Defining the Genes Required for Survival of *Mycobacterium bovis* in the Bovine Host Offers Novel Insights into the Genetic Basis of Survival of Pathogenic Mycobacteria. *mBio*, 13(4). <https://doi.org/10.1128/mbio.00672-22>
- Gibson, Amanda J., Stiens, Jennifer, Passmore Ian J., Faulkner Valwynne, Miculob Josephous, Willcocks Sam, Coad Michael, Berg Stefan, Werling Dirk, Wren Brendan W., Nobeli Irene, Villarreal-Ramos Bernardo, & Kendall Sharon L. (2022). Defining the Genes Required for Survival of *Mycobacterium bovis* in the Bovine Host Offers Novel Insights into the Genetic Basis of Survival of Pathogenic Mycobacteria. *mBio*, 13(4), e00672-22. <https://doi.org/10.1128/mbio.00672-22>
- Girardin, R. C., & McDonough, K. A. (2020). Small RNA Mcr11 requires the transcription factor AbmR for stable expression and regulates genes involved in the central metabolism of *Mycobacterium tuberculosis*. *Molecular Microbiology*, 113(2), 504–520. <https://doi.org/10.1111/mmi.14436>
- Goar, H., Paul, P., Khan, H., & Sarkar, D. (2022). Molecular Connectivity between Extracytoplasmic Sigma Factors and PhoP Accounts for Coupled Mycobacterial Stress Response. *Journal of Bacteriology*, 204(6). <https://doi.org/10.1128/jb.00110-22>
- Golby, P., Hatch, K. A., Bacon, J., Cooney, R., Riley, P., Allnutt, J., Hinds, J., Nunez, J., Marsh, P. D., Hewinson, R. G., & Gordon, S. V. (2007). Comparative transcriptomics reveals key gene expression differences between the human and bovine pathogens of the *Mycobacterium tuberculosis* complex. *Microbiology*, 153(10), 3323–3336. <https://doi.org/10.1099/mic.0.2007/009894-0>
- Golby, P., Nunez, J., Witney, A., Hinds, J., Quail, M. A., Bentley, S., Harris, S., Smith, N., Hewinson, R. G., & Gordon, S. V. (2013). Genome-level analyses of

- Mycobacterium bovis* lineages reveal the role of SNPs and antisense transcription in differential gene expression. *BMC Genomics*, 14(1).
<https://doi.org/10.1186/1471-2164-14-710>
- Gómez-Lozano, M., Marvig, R., Molin, S., & Long, K. (2014). Identification of Bacterial Small RNAs by RNA Sequencing. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 1149, pp. 433–456). https://doi.org/10.1007/978-1-4939-0473-0_34
- Gonzalo-Asensio, J., Malaga, W., Pawlik, A., Astarie-Dequeker, C., Passemar, C., Moreau, F., Laval, F., Daffé, M., Martin, C., Brosch, R., & Guilhot, C. (2014). Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proceedings of the National Academy of Sciences of the United States of America*, 111(31), 11491–11496.
<https://doi.org/10.1073/pnas.1406693111>
- Gonzalo-Asensio, J., Mostowy, S., Harders-Westerveen, J., Huygen, K., Hernández-Pando, R., Thole, J., Behr, M., Gicquel, B., & Martín, C. (2008). PhoP: a missing piece in the intricate puzzle of *Mycobacterium tuberculosis* virulence. *PloS One*, 3(10), e3496–e3496. <https://doi.org/10.1371/journal.pone.0003496>
- Grange, J. M. (2001). *Mycobacterium bovis* infection in human beings. *Tuberculosis*, 81(1-2), 71–77. <https://doi.org/10.1054/tube.2000.0263>
- Griffin, J. E., Gawronski, J. D., DeJesus, M. A., Ioerger, T. R., Akerley, B. J., & Sassetti, C. M. (2011). High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathogens*, 7(9), 1–9. <https://doi.org/10.1371/journal.ppat.1002251>
- Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R., & Hofacker, I. L. (2008). The Vienna RNA Websuite. *Nucleic Acids Research*, 36(Web Server), W70–W74. <https://doi.org/10.1093/nar/gkn188>
- Gruber, A. R., Neuböck, 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–W338. <https://doi.org/10.1093/nar/gkm222>
- Grzegorzewicz, A. E., Ma, Y., Jones, V., Crick, D., Liav, A., & McNeil, M. R. (2008). Development of a microtitre plate-based assay for lipid-linked glycosyltransferase products using the mycobacterial cell wall

- rhamnosyltransferase WbbL. *Microbiology*, 154(12), 3724–3730.
<https://doi.org/10.1099/mic.0.2008/023366-0>
- Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). Circlize implements and enhances circular visualization in R. *Bioinformatics*, 30(19), 2811–2812.
<https://doi.org/10.1093/bioinformatics/btu393>
- Harold, L. K., Antoney, J., Ahmed, F. H., Hards, K., Carr, P. D., Rapson, T., Greening, C., Jackson, C. J., & Cook, G. M. (2019). FAD-sequestering proteins protect mycobacteria against hypoxic and oxidative stress. *Journal of Biological Chemistry*, 294(8), 2903–5814. <https://doi.org/10.1074/jbc.RA118.006237>
- Harth, G., Clemens, D. L., & Horwitz, M. A. (1994). Glutamine synthetase of *Mycobacterium tuberculosis*: Extracellular release and characterization of its enzymatic activity. *Proceedings of the National Academy of Sciences*, 91(20), 9342–9346. <https://doi.org/10.1073/pnas.91.20.9342>
- He, X., & Wang, S. (2014). DNA Consensus Sequence Motif for Binding Response Regulator PhoP, a Virulence Regulator of *Mycobacterium tuberculosis*. *Biochemistry*, 53(51), 8008–8020. <https://doi.org/10.1021/bi501019u>
- Holder, T., Srinivasan, S., McGoldrick, A., Williams, G. A., Palmer, S., Clarke, J., O'Brien, A., Conlan, A. J. K., Juleff, N., Vordermeier, H. M., Jones, G. J., & Kapur, V. (2024). Temporal dynamics of the early immune response following *Mycobacterium bovis* infection of cattle. *Scientific Reports*, 14(1), 2600.
<https://doi.org/10.1038/s41598-024-52314-x>
- Houghton, Joanna, Rodgers, Angela, Rose, Graham, D'Halluin, Alexandre, Kipkorir, Terry, Barker, Declan, Waddell, Simon J., Arnvig, Kristine B., & Oglesby, Amanda G. (2021). The *Mycobacterium tuberculosis* sRNA F6 Modifies Expression of Essential Chaperonins, GroEL2 and GroES. *Microbiology Spectrum*, 9(2), e01095-21. <https://doi.org/10.1128/Spectrum.01095-21>
- Howe, F. S., Russell, A., Lamstaes, A. R., El-Sagheer, A., Nair, A., Brown, T., & Mellor, J. (2017). CRISPRi is not strand-specific at all loci and redefines the transcriptional landscape. *eLife*, 6, e29878.
<https://doi.org/10.7554/eLife.29878>
- Hu, Y., Wang, Z., Feng, L., Chen, Z., Mao, C., Zhu, Y., & Chen, S. (2016). σ E-dependent activation of RbpA controls transcription of the *furA-katG* operon in response to oxidative stress in mycobacteria. *Molecular Microbiology*, 102(1), 107–120. <https://doi.org/10.1111/mmi.13449>

- Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009a). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1), 1–13.
<https://doi.org/10.1093/nar/gkn923>
- Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1), 44–57. <https://doi.org/10.1038/nprot.2008.211>
- Huang, Q., Abdalla, A. E., & Xie, J. (2015). Phylogenomics of Mycobacterium Nitrate Reductase Operon. *Current Microbiology*, 71(1), 121–128.
<https://doi.org/10.1007/s00284-015-0838-2>
- Huber, M., Faure, G., Laass, S., Kolbe, E., Seitz, K., Wehrheim, C., Wolf, Y. I., Koonin, E. V., & Soppa, J. (2019). Translational coupling via termination-reinitiation in archaea and bacteria. *Nature Communications*, 10(1), 4006.
<https://doi.org/10.1038/s41467-019-11999-9>
- Hutchison, C. A., Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., Smith, H. O., & Venter, J. C. (1999). Global transposon mutagenesis and a minimal mycoplasma genome. *Science*, 286(5447), 2165–2169.
<https://doi.org/10.1126/science.286.5447.2165>
- 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(1), 954.
<https://doi.org/10.1186/s12864-015-2197-6>
- Ignatov, D., Vaitkevicius, K., Durand, S., Cahoon, L., Sandberg, S. S., Liu, X., Kallipolitis, B. H., Rydén, P., Freitag, N., Condon, C., & Johansson, J. (2020). An mRNA-mRNA Interaction Couples Expression of a Virulence Factor and Its Chaperone in *Listeria monocytogenes*. *Cell Reports*, 30(12), 4027–4040.e7. <https://doi.org/10.1016/j.celrep.2020.03.006>
- Iost, I., & Dreyfus, M. (1995). The stability of *Escherichia coli lacZ* mRNA depends upon the simultaneity of its synthesis and translation. *The EMBO Journal*, 14(13), 3252–3261. <https://doi.org/10.1002/j.1460-2075.1995.tb07328.x>
- Jiang, J., Lin, C., Zhang, J., Wang, Y., Shen, L., Yang, K., Xiao, W., Li, Y., Zhang, L., & Liu, J. (2020). Transcriptome Changes of *Mycobacterium marinum* in the Process of Resuscitation From Hypoxia-Induced Dormancy. *Frontiers in Genetics*, 10(February), 1–13. <https://doi.org/10.3389/fgene.2019.01359>

- Jiang, J., Sun, X., Wu, W., Li, L., Wu, H., Zhang, L., Yu, G., & Li, Y. (2016). Construction and application of a co-expression network in *Mycobacterium tuberculosis*. *Scientific Reports*, 6(March 2015), 1–18.
<https://doi.org/10.1038/srep28422>
- Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., & Lempicki, R. A. (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, 28(13), 1805–1806.
<https://doi.org/10.1093/bioinformatics/bts251>
- Jinich, A., Zaveri, A., Dejesus, M. A., Flores-bautista, E., & Smith, C. M. (2021). *The Mycobacterium tuberculosis transposon sequencing database (MtbTnDB): A large-scale guide to genetic conditional essentiality. i*, 1–21.
- Ju, X., Li, D., & Liu, S. (2019). Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. *Nature Microbiology*, 4(11), 1907–1918. <https://doi.org/10.1038/s41564-019-0500-z>
- Ju, X., Li, S., Froom, R., Wang, L., Lilic, M., Campbell, E. A., Rock, J. M., & Liu, S. (2024). Incomplete transcripts dominate the Mycobacterium tuberculosis transcriptome. *Nature*, 2023.03.10.532058.
<https://doi.org/10.1038/s41586-024-07105-9>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589.
<https://doi.org/10.1038/s41586-021-03819-2>
- Kalscheuer, R., Palacios, A., Anso, I., Cifuentes, J., Anguita, J., Jacobs, W. R., Jr, Guerin, M. E., & Prados-Rosales, R. (2019). The *Mycobacterium tuberculosis* capsule: A cell structure with key implications in pathogenesis. *Biochemical Journal*, 476(14), 1995–2016. <https://doi.org/10.1042/BCJ20190324>
- Kalscheuer Rainer & Koliwer-Brandl Hendrik. (2014). Genetics of Mycobacterial Trehalose Metabolism. *Microbiology Spectrum*, 3, 10.1128/microbiolspec.mgm2-0002-2013.
<https://doi.org/10.1128/microbiolspec.mgm2-0002-2013>
- Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z.,

- Rivas, E., Eddy, S. R., Finn, R. D., Bateman, A., & Petrov, A. I. (2021). Rfam 14: Expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, 49(D1), D192–D200.
<https://doi.org/10.1093/nar/gkaa1047>
- Kanehisa, M., Sato, Y., & Kawashima, M. (2022). KEGG mapping tools for uncovering hidden features in biological data. *Protein Science*, 31(1), 47–53.
<https://doi.org/10.1002/pro.4172>
- Kang, C.-M., Nyayapathy, S., Lee, J.-Y., Suh, J.-W., & Husson, R. N. (2008). Wag31, a homologue of the cell division protein DivIVA, regulates growth, morphology and polar cell wall synthesis in mycobacteria. *Microbiology*, 154(3), 725–735. <https://doi.org/10.1099/mic.0.2007/014076-0>
- 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.
<https://doi.org/10.1016/j.tube.2010.09.006>
- Keating, L. A., Wheeler, P. R., Mansoor, H., Inwald, J. K., Dale, J., Hewinson, R. G., & Gordon, S. V. (2005). The pyruvate requirement of some members of the *Mycobacterium tuberculosis* complex is due to an inactive pyruvate kinase: Implications for in vivo growth. *Molecular Microbiology*, 56(1), 163–174.
<https://doi.org/10.1111/j.1365-2958.2005.04524.x>
- Kendall, S. L., Burgess, P., Balhana, R., Withers, M., Ten Bokum, A., Lott, J. S., Gao, C., Uhia-Castro, I., & Stoker, N. G. (2010). Cholesterol utilization in mycobacteria is controlled by two TetR-type transcriptional regulators: *kstR* and *kstR2*. *Microbiology*, 156(5), 1362–1371.
<https://doi.org/10.1099/mic.0.034538-0>
- Kendall, S. L., Withers, M., Soffair, C. N., Moreland, N. J., Gurcha, S., Sidders, B., Frita, R., Ten Bokum, A., Besra, G. S., Lott, J. S., & Stoker, N. G. (2007). A highly conserved transcriptional repressor controls a large regulon involved in lipid degradation in *Mycobacterium smegmatis* and *Mycobacterium tuberculosis*. *Molecular Microbiology*, 65(3), 684–699.
<https://doi.org/10.1111/j.1365-2958.2007.05827.x>
- Khan, H., Paul, P., Goar, H., Bamniya, B., Baid, N., & Sarkar, D. (2024). *Mycobacterium tuberculosis* PhoP integrates stress response to intracellular

survival by maintenance of cAMP level. *eLife*, 1–42.

<https://elifesciences.org/articles/92136>

Kieser, K. J., Baranowski, C., Chao, M. C., Long, J. E., Sassetti, C. M., Waldor, M. K., Sacchettini, J. C., Ioerger, T. R., & Rubin, E. J. (2015). Peptidoglycan synthesis in *Mycobacterium tuberculosis* is organized into networks with varying drug susceptibility. *Proceedings of the National Academy of Sciences of the United States of America*, 112(42), 13087–13092.

<https://doi.org/10.1073/pnas.1514135112>

Kipkorir, T., Polgar, P., Barker, D., D'Halluin, A. D., Patel, Z., & Arnvig, K. B. (2024). A novel regulatory interplay between atypical B 12 riboswitches and uORF translation in *Mycobacterium tuberculosis*. *Nucleic Acids Research*, v, 1–17.

[https://academic.oup.com/nar/advance-](https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae338/7665635?login=false)

[article/doi/10.1093/nar/gkae338/7665635?login=false](https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae338/7665635?login=false)

Kipkorir, Terry, Mashabela, Gabriel T., de Wet, Timothy J., Koch, Anastasia, Dawes Stephanie S., Wiesner, Lubbe, Mizrahi, Valerie, Warner, Digby F., & Henkin, Tina M. (2021). De Novo Cobalamin Biosynthesis, Transport, and Assimilation and Cobalamin-Mediated Regulation of Methionine Biosynthesis in *Mycobacterium smegmatis*. *Journal of Bacteriology*, 203(7), e00620-20. <https://doi.org/10.1128/JB.00620-20>

Klepp, L. I., Sabio y Garcia, J., & FabianaBigi. (2022). Mycobacterial MCE proteins as transporters that control lipid homeostasis of the cell wall. *Tuberculosis*, 132, 102162. <https://doi.org/10.1016/j.tube.2021.102162>

Kock, R., Michel, A. L., Yeboah-Manu, D., Azhar, E. I., Torrelles, J. B., Cadmus, S. I., Brunton, L., Chakaya, J. M., Marais, B., Mboera, L., Rahim, Z., Haider, N., & Zumla, A. (2021). Zoonotic Tuberculosis – The Changing Landscape. *Commemorating World Tuberculosis Day March 24th, 2021: “The Clock Is Ticking”*, 113, S68–S72. <https://doi.org/10.1016/j.ijid.2021.02.091>

Kumar, K., Chakraborty, A., & Chakrabarti, S. (2020). PresRAT: A server for identification of bacterial small-RNA sequences and their targets with probable binding region. *RNA Biology*, 2020.04.03.024935. <https://doi.org/10.1101/2020.04.03.024935>

Lamichhane, G., Arnvig, K. B., & McDonough, K. A. (2013). Definition and annotation of (myco)bacterial non-coding RNA. *Tuberculosis*, 93(1), 26–29. <https://doi.org/10.1016/j.tube.2012.11.010>

- Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, 9.
<https://doi.org/10.1186/1471-2105-9-559>
- Langridge, G. C., Phan, M.-D., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., Charles, I., Maskell, D. J., Peters, S. E., Dougan, G., Wain, J., Parkhill, J., & Turner, A. K. (2009). Simultaneous assay of every *Salmonella Typhi* gene using one million transposon mutants. *Genome Research*, 19(12), 2308–2316.
<https://doi.org/10.1101/gr.097097.109>
- Larson, M. H., Gilbert, L. A., Wang, X., Lim, W. A., Weissman, J. S., & Qi, L. S. (2013). CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature Protocols*, 8(11), 2180–2196.
<https://doi.org/10.1038/nprot.2013.132>
- Lasa, I., Toledo-Arana, A., Dobin, A., Villanueva, M., de los Mozos, I. R., Vergara-Irigaray, M., Segura, V., Fagegaltier, D., Penadés, J. R., Valle, J., Solano, C., & Gingeras, T. R. (2011). Genome-wide antisense transcription drives mRNA processing in bacteria. *Proceedings of the National Academy of Sciences*, 108(50), 20172–20177. <https://doi.org/10.1073/pnas.1113521108>
- Lee, J. J., Lim, J., Gao, S., Lawson, C. P., Odell, M., Raheem, S., Woo, J. I., Kang, S. H., Kang, S. S., Jeon, B. Y., & Eoh, H. (2018). Glutamate mediated metabolic neutralization mitigates propionate toxicity in intracellular *Mycobacterium tuberculosis*. *Scientific Reports*, 8(1), 1–13.
<https://doi.org/10.1038/s41598-018-26950-z>
- Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6), 882–883.
<https://doi.org/10.1093/bioinformatics/bts034>
- Lefrançois LH, Nitschke J, Wu H, Panis G, Prados J, Butler RE, Mendum TA, Hanna N, Stewart GR, Soldati T. 2024. Temporal genome-wide fitness analysis of *Mycobacterium marinum* during infection reveals the genetic requirement for virulence and survival in amoebae and microglial cells. *mSystems*, 9:e01326-23. <https://doi.org/10.1128/msystems.01326-23>
- Lei, L., Stipp, R. N., Chen, T., Wu, S. Z., Hu, T., & Duncan, M. J. (2018). Activity of *Streptococcus mutans* VicR Is Modulated by Antisense RNA. *Journal of*

- Dental Research*, 97(13), 1477–1484.
<https://doi.org/10.1177/0022034518781765>
- Lejars, M., Caillet, J., Solchaga-Flores, E., Guillier, M., Plumbridge, J., & Hajnsdorf, E. (2022). Regulatory Interplay between RNase III and Antisense RNAs in *E. coli*: The Case of AsflhD and FlhD, Component of the Master Regulator of Motility. *mBio*, 3:e00981-22. <https://doi.org/10.1128/mbio.00981-22>
- Lejars, M., Kobayashi, A., & Hajnsdorf, E. (2019). Physiological roles of antisense RNAs in prokaryotes. *Biochimie*, 164, 3–16.
<https://doi.org/10.1016/j.biochi.2019.04.015>
- Leonard, S., Meyer, S., Lacour, S., Nasser, W., Hommais, F., & Reverchon, S. (2019). APERO: a genome-wide approach for identifying bacterial small RNAs from RNA-Seq data. *Nucleic Acids Research*, 47(15), e88–e88.
<https://doi.org/10.1093/nar/gkz485>
- Lew, J. M., Kapopoulou, A., Jones, L. M., & Cole, S. T. (2011). TubercuList—10 years after. *Tuberculosis (Edinburgh, Scotland)*, 91(1), 1–7.
<https://doi.org/10.1016/j.tube.2010.09.008>
- Li, Heng. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. <https://doi.org/10.48550/arXiv.1303.3997>
- Li, L., Huang, D., Cheung, M. K., Nong, W., Huang, Q., & Kwan, H. S. (2013). BSRD: a repository for bacterial small regulatory RNA. *Nucleic Acids Research*, 41(Database issue), D233–D238. PubMed.
<https://doi.org/10.1093/nar/gks1264>
- Li, X., Chen, F., Liu, X., Xiao, J., Andongma, B. T., Tang, Q., Cao, X., Chou, S. H., Galperin, M. Y., & He, J. (2022). Clp protease and antisense RNA jointly regulate the global regulator CarD to mediate mycobacterial starvation response. *eLife*, 11, 1–22. <https://doi.org/10.7554/eLife.73347>
- Liang, Y., Plourde, A., Bueler, S. A., Liu, J., Brezezinski, P., Vahidi, S., & Rubinstein, J. L. (2023). Structure of mycobacterial respiratory complex I. *Proceedings of the National Academy of Sciences*, 120(13), 2023.
<https://doi.org/10.1073/pnas>
- Liu, C., He, Y., & Chang, Z. (2004). Truncated hemoglobin o of *Mycobacterium tuberculosis*: The oligomeric state change and the interaction with membrane components. *Biochemical and Biophysical Research*

- Communications*, 316(4), 1163–1172.
<https://doi.org/10.1016/j.bbrc.2004.02.170>
- Liu, W., Rochat, T., Toffano-Nioche, C., Le Lam, T. N., Boulloc, P., & Morvan, C. (2018). Assessment of Bona Fide sRNAs in *Staphylococcus aureus*. In *Frontiers in Microbiology* (Vol. 9).
<https://www.frontiersin.org/article/10.3389/fmicb.2018.00228>
- Livny, J., Teonadi, H., Livny, M., & Waldor, M. K. (2008). High-Throughput, Kingdom-Wide Prediction and Annotation of Bacterial Non-Coding RNAs. *PLOS ONE*, 3(9), e3197. <https://doi.org/10.1371/journal.pone.0003197>
- Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W.-H., Bork, P., Glass, J. I., Serrano, L., & Lluch-Senar, M. (2016). Bacterial antisense RNAs are mainly the product of transcriptional noise. *Science Advances*, 2(3), e1501363. <https://doi.org/10.1126/sciadv.1501363>
- Loiseau, C., Menardo, F., Aseffa, A., Hailu, E., Gumi, B., Ameni, G., Berg, S., Rigouts, L., Robbe-Austerman, S., Zinsstag, J., Gagneux, S., & Brites, D. (2020). An African origin for *Mycobacterium bovis*. *Evolution, Medicine, and Public Health*, 2020(1), 49–59. <https://doi.org/10.1093/emph/eoaa005>
- Long, J. E., Dejesus, M., Ward, D., Baker, R. E., Ioerger, T., & Sassetti, C. M. (2015). Global Phenotypic Profiling. In *Gene Essentiality: Methods and Protocols* (Vol. 1279, pp. 79–95). <https://doi.org/10.1007/978-1-4939-2398-4>
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 1–21. <https://doi.org/10.1186/s13059-014-0550-8>
- Lu, L., Wei, R., Bhakta, S., Waddell, S. J., & Boix, E. (2021). Weighted gene co-expression network analysis to identify key modules and hub genes associated with Mycobacterial Infection of Human Macrophages. *Antibiotics*, 10(97). <https://doi.org/10.3390/antibiotics10020097>
- Lunge, A., Gupta, R., Choudhary, E., & Agarwal, N. (2020). The unfoldase ClpC1 of *Mycobacterium tuberculosis* regulates the expression of a distinct subset of proteins having intrinsically disordered termini. *Journal of Biological Chemistry*, 295(28), 9455–9473.
<https://doi.org/10.1074/jbc.RA120.013456>

- Lybecker, M., Bilusic, I., & Raghavan, R. (2014). Pervasive transcription: Detecting functional RNAs in bacteria. *Transcription*, 5(4).
<https://doi.org/10.4161/21541272.2014.944039>
- Lybecker, M., Zimmermann, B., Bilusic, I., Tukhtubaeva, N., & Schroeder, R. (2014). The double-stranded transcriptome of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 111(8), 3134–3139.
<https://doi.org/10.1073/pnas.1315974111>
- Ma, R., Farrell, D., Gonzalez, G., Browne, J. A., Nakajima, C., Suzuki, Y., & Gordon, S. V. (2022). The TbD1 Locus Mediates a Hypoxia-Induced Copper Response in *Mycobacterium bovis*. *Frontiers in Microbiology*, 13(April), 1–14.
<https://doi.org/10.3389/fmicb.2022.817952>
- Maciąg, A., Dainese, E., Rodriguez, G., Milano, A., Provvedi, R., Pasca, M.R., Smith, I., Palù, G., Riccardi, G., & Manganelli, R. (2007). Global Analysis of the *Mycobacterium tuberculosis* Zur (FurB) Regulon. *Journal of Bacteriology*, 189(3), 730–740. <https://doi.org/10.1128/JB.01190-06>
- Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res*, 47(W1), W636–W641. PubMed. <https://doi.org/10.1093/nar/gkz268>
- Magee, D. A., Conlon, K. M., Nalpas, N. C., Browne, J. A., Pirson, C., Healy, C., McLoughlin, K. E., Chen, J., Vordermeier, H. M., Gormley, E., MacHugh, D. E., & Gordon, S. V. (2014). Innate cytokine profiling of bovine alveolar macrophages reveals commonalities and divergence in the response to *Mycobacterium bovis* and *Mycobacterium tuberculosis* infection. *Tuberculosis*, 94(4), 441–450. <https://doi.org/10.1016/j.tube.2014.04.004>
- Mahapatra, A., Mativandlela, S. P. N., Binneman, B., Fourie, P. B., Hamilton, C. J., Meyer, J. J. M., van der Kooy, F., Houghton, P., & Lall, N. (2007). Activity of 7-methyljuglone derivatives against *Mycobacterium tuberculosis* and as subversive substrates for mycothiol disulfide reductase. *Bioorganic and Medicinal Chemistry*, 15(24), 7638–7646.
<https://doi.org/10.1016/j.bmc.2007.08.064>
- Mahmutovic, A., Abel zur Wiesch, P., & Abel, S. (2020). Selection or drift: The population biology underlying transposon insertion sequencing

- experiments. *Computational and Structural Biotechnology Journal*, 18, 791–804. <https://doi.org/10.1016/j.csbj.2020.03.021>
- 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. <https://doi.org/10.1093/nar/gkz149>
- Majumdar, G., Mbau, R., Singh, V., Warner, D. F., Dragset, M. S., & Mukherjee, R. (2017). Genome-wide transposon mutagenesis in *Mycobacterium tuberculosis* and *Mycobacterium smegmatis*. *Methods in Molecular Biology*, 1498, 321–335. https://doi.org/10.1007/978-1-4939-6472-7_21
- Malone, K. M., & Gordon, S. (2017). Strain Variation in the *Mycobacterium tuberculosis* Complex: Its Role in Biology, Epidemiology and Control. In Gagneaux, S. (Ed.), *Mycobacterium tuberculosis Complex Members Adapted to Wild and Domestic Animals* (135–153). Springer. <https://doi.org/10.1007/978-3-319-64371-7>
- Malone, K. M., Rue-Albrecht, K., Magee, D. A., Conlon, K., Schubert, O. T., Nalpas, N. C., Browne, J. A., Smyth, A., Gormley, E., Aebersold, R., MacHugh, D. E., & Gordon, S. V. (2018). Comparative 'omics analyses differentiate *Mycobacterium tuberculosis* and *Mycobacterium bovis* and reveal distinct macrophage responses to infection with the human and bovine tubercle bacilli. *Microbial Genomics*, 4(3). <https://doi.org/10.1099/mgen.0.000163>
- Malone, K. M., Farrell, D., Stuber T. P., Schubert, O. T., Aebersold, R., Robbe-Austerman, S., & Gordon, S. (2017). Updated Reference Genome Sequence and Annotation of *Mycobacterium bovis* AF2122/97. *Genome Announcements*, 5(14), 10.1128/genomea.00157-17. <https://doi.org/10.1128/genomea.00157-17>
- Mann, M., Wright, P. R., & Backofen, R. (2017). IntaRNA 2.0: Enhanced and customizable prediction of RNA-RNA interactions. *NAR*, 45(W1), W435–W439. <https://doi.org/10.1093/nar/gkx279>
- Mao, X., Ma, Q., Liu, B., Chen, X., Zhang, H., & Xu, Y. (2015). Revisiting operons: An analysis of the landscape of transcriptional units in *E. coli*. *BMC Bioinformatics*, 16(1), 356. <https://doi.org/10.1186/s12859-015-0805-8>
- Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLOS*

- Computational Biology*, 14(1), e1005944.
<https://doi.org/10.1371/journal.pcbi.1005944>
- Martini, M. C., Zhou, Y., Sun, H., & Shell, S. S. (2019). Defining the Transcriptional and Post-transcriptional Landscapes of *Mycobacterium smegmatis* in Aerobic Growth and Hypoxia. *Frontiers in Microbiology*, 10(March).
<https://www.frontiersin.org/article/10.3389/fmicb.2019.00591>
- Mawuenyega, K. G., Forst, C. V., Dobos, K. M., Belisle, J. T., Chen, J., Bradbury, E. M., Bradbury, A. R. M., & Chen, X. (2005). *Mycobacterium tuberculosis* Functional Network Analysis by Global Subcellular Protein Profiling. *Molecular Biology of the Cell*, 16(1), 396–404.
<https://doi.org/10.1091/mbc.e04-04-0329>
- McClure, R., Balasubramanian, D., Sun, Y., Bobrovskyy, M., Sumby, P., Genco, C. A., Vanderpool, C. K., & Tjaden, B. (2013). Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Research*, 41(14), e140–e140.
<https://doi.org/10.1093/nar/gkt444>
- Mendum, T. A., Chandran, A., Williams, K., Vordermeier, H. M., Villarreal-Ramos, B., Wu, H., Singh, A., Smith, A. A., Butler, R. E., Prasad, A., Bharti, N., Banerjee, R., Kasibhatla, S. M., Bhatt, A., Stewart, G. R., & McFadden, J. (2019). Transposon libraries identify novel *Mycobacterium bovis* BCG genes involved in the dynamic interactions required for BCG to persist during in vivo passage in cattle. *BMC Genomics*, 20(1), 1–13. <https://doi.org/10.1186/s12864-019-5791-1>
- Menendez-Gil, P., Caballero, C., Catalan-Moreno, A., Irurzun, N., Barrio-Hernandez, I., Caldelari, I., & Toledo-Arana, A. (2020). Differential evolution in 3'UTRs leads to specific gene expression in Staphylococcus. *Nucleic Acids Research*, 48. <https://doi.org/10.1093/nar/gkaa047>
- Menendez-Gil, P., & Toledo-Arana, A. (2021). Bacterial 3'UTRs: A Useful Resource in Post-transcriptional Regulation. *Frontiers in Molecular Biosciences*, 7. <https://www.frontiersin.org/article/10.3389/fmolb.2020.617633>
- Meng, E. C., Goddard, T. D., Pettersen, E. F., Couch, G. S., Pearson, Z. J., Morris, J. H., & Ferrin, T. E. (2023). UCSF ChimeraX: Tools for structure building and analysis. *Protein Science*, 32(11), e4792. <https://doi.org/10.1002/pro.4792>
- Minato, Y., Gohl, D. M., Thiede, J. M., Chacón, 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. <https://doi.org/10.1128/msystems.00070-19>
- Miotto, P., Forti, F., Ambrosi, A., Pellin, D., Veiga, D. F., Balazsi, G., Gennaro, M. L., Di Serio, C., Ghisotti, D., & Cirillo, D. M. (2012). Genome-Wide Discovery of Small RNAs in *Mycobacterium tuberculosis*. *PLOS ONE*, 7(12), e51950. <https://doi.org/10.1371/journal.pone.0051950>
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1), D412–D419. <https://doi.org/10.1093/nar/gkaa913>
- Mitra, A., Speer, A., Lin, K., Ehrt, Sabine, Niederweis, Michael, & Stallings, Christina L. (2017). PPE Surface Proteins Are Required for Heme Utilization by *Mycobacterium tuberculosis*. *mBio*, 8(1), 1–14. <https://doi.org/10.1128/mBio.01720-16>
- Modlin, S. J., Afif, E., Deepika, G., Zlotnicki, A. M., Dillon, N. A., Dhillon, N., Kuo, N., Robinhold, C., Chan, C. K., Baughn, A. D., & Valafar, F. (2021). Structure-Aware *Mycobacterium tuberculosis* Functional Annotation Uncloaks Resistance, Metabolic, and Virulence Genes. *mSystems*, 0(0), e00673-21. <https://doi.org/10.1128/mSystems.00673-21>
- Mölder, F., Jablonski, K., Letcher, B., Hall, M., Tomkins-Tinch, C., Sochat, V., Forster, J., Lee, S., Twardziok, S., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake [version 2; peer review: 2 approved]. *F1000Research*, 10(33). <https://doi.org/10.12688/f1000research.29032.2>
- Moore, A., Riesco, A. B., Schwenk, S., & Arnvig, K. B. (2017). Expression, maturation and turnover of DrrS, an unusually stable, DosR regulated small RNA in *Mycobacterium tuberculosis*. *PLOS ONE*, 12(3), e0174079. <https://doi.org/10.1371/journal.pone.0174079>
- Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., ... Groop, L. C. (2003). PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately

- downregulated in human diabetes. *Nature Genetics*, 34(3), 267–273.
<https://doi.org/10.1038/ng1180>
- Morra, R., Pratama, F., Butterfield, T., Tomazetto, G., Young, K., Lopez, R., & Dixon, N. (2023). *arfA* antisense RNA regulates MscL excretory activity. *Life Science Alliance*, 6(6), e202301954. <https://doi.org/10.26508/lsa.202301954>
- NCBI Resource Coordinators. (2014). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 42(D1), D7–D17.
<https://doi.org/10.1093/nar/gkt1146>
- Negri, A., Javidnia, P., Mu, R., Zhang, X., Vendome, J., Gold, B., Roberts, J., Barman, D., Ioerger, T., Sacchettini, J. C., Jiang, X., Burns-Huang, K., Warriar, T., Ling, Y., Warren, J. D., Oren, D. A., Beuming, T., Wang, H., Wu, J., ... Somersan-Karakaya, S. (2018). Identification of a Mycothiol-Dependent Nitroreductase from *Mycobacterium tuberculosis*. *ACS Infectious Diseases*, 4(5), 771–787. <https://doi.org/10.1021/acsinfecdis.7b00111>
- Negri, L. B., Mannaa, Y., Korupolu, S., Farinelli, W. A., Anderson, R. R., & Gelfand, J. A. (2023). Vitamin K3 (Menadione) is a multifunctional microbicide acting as a photosensitizer and synergizing with blue light to kill drug-resistant bacteria in biofilms. *Journal of Photochemistry and Photobiology B: Biology*, 244(May), 112720. <https://doi.org/10.1016/j.jphotobiol.2023.112720>
- Nesbitt, N. M., Yang, X., Fontán, P., Kolesnikova, I., Smith, I., Sampson, N. S., & Dubnau, E. (2010). A Thiolase of *Mycobacterium tuberculosis* Is Required for Virulence and Production of Androstenedione and Androstadienedione from Cholesterol. *Infection and Immunity*, 78(1), 275 LP – 282.
<https://doi.org/10.1128/IAI.00893-09>
- Newton-Foot, M., & Gey van Pittius, N. C. (2013). The complex architecture of mycobacterial promoters. *Tuberculosis*, 93(1), 60–74.
<https://doi.org/10.1016/j.tube.2012.08.003>
- Neyrolles, O., Wolschendorf, F., Mitra, A., & Niederweis, M. (2015). Mycobacteria, metals, and the macrophage. *Immunological Reviews*, 264(1), 249–263.
<https://doi.org/10.1111/imr.12265>
- Olea-Popelka, F., Muwonge, A., Perera, A., Dean, A. S., Mumford, E., Erlacher-Vindel, E., Forcella, S., Silk, B. J., Ditiu, L., El Idrissi, A., Raviglione, M., Cosivi, O., LoBue, P., & Fujiwara, P. I. (2017). Zoonotic tuberculosis in human beings caused by *Mycobacterium bovis*—A call for action. *The Lancet Infectious*

Diseases, 17(1), e21–e25. [https://doi.org/10.1016/S1473-3099\(16\)30139-6](https://doi.org/10.1016/S1473-3099(16)30139-6)

- Oliva B, Gordon G, McNicholas P, Ellestad G, & Chopra I. (1992). Evidence that tetracycline analogs whose primary target is not the bacterial ribosome cause lysis of *Escherichia coli*. *Antimicrobial Agents and Chemotherapy*, 36(5), 913–919. <https://doi.org/10.1128/aac.36.5.913>
- Ozuna, A., Liberto, D., Joyce, R. M., Arnvig, K. B., & Nobeli, I. (2019). baerhunter: An R package for the discovery and analysis of expressed non-coding regions in bacterial RNA-seq data. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btz643>
- Pacl, H. T., Reddy, V. P., Saini, V., Chinta, K. C., & Steyn, A. J. C. (2018). Host-pathogen redox dynamics modulate *Mycobacterium tuberculosis* pathogenesis. *Pathogens and Disease*, 76(5), 1–14. <https://doi.org/10.1093/femspd/fty036>
- Parveen, S., Shen, J., Lun, S., Zhao, L., Alt, J., Koleske, B., Leone, R. D., Rais, R., Powell, J. D., Murphy, J. R., Slusher, B. S., & Bishai, W. R. (2023). Glutamine metabolism inhibition has dual immunomodulatory and antibacterial activities against *Mycobacterium tuberculosis*. *Nature Communications*, 14(1), 7427. <https://doi.org/10.1038/s41467-023-43304-0>
- Patil, S., Palande, A., Lodhiya, T., Pandit, A., & Mukherjee, R. (2021). Redefining genetic essentiality in *Mycobacterium tuberculosis*. *Gene*, 765, 145091. <https://doi.org/10.1016/j.gene.2020.145091>
- Pawaria, S., Rajamohan, G., Gambhir, V., Lama, A., Varshney, G. C., & Dikshit, K. L. (2007). Intracellular growth and survival of *Salmonella enterica* serovar Typhimurium carrying truncated hemoglobins of *Mycobacterium tuberculosis*. *Microbial Pathogenesis*, 42(4), 119–128. <https://doi.org/10.1016/j.micpath.2006.12.001>
- Pawaria Sudesh, Lama Amrita, Raje Manoj, & Dikshit Kanak L. (2008). Responses of *Mycobacterium tuberculosis* Hemoglobin Promoters to *In Vitro* and *In Vivo* Growth Conditions. *Applied and Environmental Microbiology*, 74(11), 3512–3522. <https://doi.org/10.1128/AEM.02663-07>
- Pawełczyk, J., Brzostek, A., Minias, A., Płociński, P., Rumijowska-Galewicz, A., Strapagiel, D., Zakrzewska-Czerwińska, J., & Dziadek, J. (2021). Cholesterol-dependent transcriptome remodeling reveals new insight into the

- contribution of cholesterol to *Mycobacterium tuberculosis* pathogenesis. *Scientific Reports*, 11(1), 12396. <https://doi.org/10.1038/s41598-021-91812-0>
- Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., Bileschi, M. L., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, D. A., Orengo, C. A., Pandurangan, A. P., Rivoire, C., ... Bateman, A. (2023). InterPro in 2022. *Nucleic Acids Res*, 51(D1), D418–D427. PubMed. <https://doi.org/10.1093/nar/gkac993>
- Pellin, D., Miotto, P., Ambrosi, A., Cirillo, D. M., & Di Serio, C. (2012). A Genome-Wide Identification Analysis of Small Regulatory RNAs in *Mycobacterium tuberculosis* by RNA-Seq and Conservation Analysis. *PLOS ONE*, 7(3), e32723. <https://doi.org/10.1371/journal.pone.0032723>
- 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. <https://doi.org/10.1016/j.gene.2012.03.044>
- Peterson, E. J. R., Reiss, D. J., Turkarslan, S., Minch, K. J., Rustad, T., Plaisier, C. L., Longabaugh, W. J. R., Sherman, D. R., & Baliga, N. S. (2014). A high-resolution network model for global gene regulation in *Mycobacterium tuberculosis*. *Nucleic Acids Research*, 42(18), 11291–11303. <https://doi.org/10.1093/nar/gku777>
- Ponath, F., Hör, J., & Vogel, J. (2022). An overview of gene regulation in bacteria by small RNAs derived from mRNA 3' ends. *FEMS Microbiology Reviews*, 46(5), fuac017. <https://doi.org/10.1093/femsre/fuac017>
- Puniya, B. L., Kulshreshtha, D., Verma, S. P., Kumar, S., & Ramachandran, S. (2013). Integrated gene co-expression network analysis in the growth phase of *Mycobacterium tuberculosis* reveals new potential drug targets. *Molecular BioSystems*, 9(11), 2798–2815. <https://doi.org/10.1039/c3mb70278b>
- Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., & Lim, W. A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, 152(5), 1173–1183. <https://doi.org/10.1016/j.cell.2013.02.022>
- Queval, C. J., Fearn, A., Botella, L., Smyth, A., Schnettger, L., Mitermite, M., Wooff, E., Villarreal-Ramos, B., Garcia-Jimenez, W., Heunis, T., Trost, M., Werling, D.,

- Salguero, F. J., Gordon, S. V., & Gutierrez, M. G. (2021). Macrophage-specific responses to human- and animal-adapted tubercle bacilli reveal pathogen and host factors driving multinucleated cell formation. *PLOS Pathogens*, 17(3), e1009410. <https://doi.org/10.1371/journal.ppat.1009410>
- Rajwani, R., Galata, C., Lee, A. W. T., So, P.-K., Leung, K. S. S., Tam, K. K. G., Shehzad, S., Ng, T. T. L., Zhu, L., Lao, H. Y., Chan, C. T.-M., Leung, J. S.-L., Lee, L.-K., Wong, K. C., Yam, W. C., & Siu, G. K.-H. (2022). A multi-omics investigation into the mechanisms of hyper-virulence in *Mycobacterium tuberculosis*. *Virulence*, 13(1), 1088–1100. <https://doi.org/10.1080/21505594.2022.2087304>
- Ramage, H. R., Connolly, L. E., & Cox, J. S. (2009). Comprehensive functional analysis of *Mycobacterium tuberculosis* toxin-antitoxin systems: Implications for pathogenesis, stress responses, and evolution. *PLoS Genetics*, 5(12). <https://doi.org/10.1371/journal.pgen.1000767>
- Ramakrishnan, P., Aagesen, A. M., McKinney, J. D., & Tischler, A. D. (2016). *Mycobacterium tuberculosis* resists stress by regulating PE19 expression. *Infection and Immunity*, 84(3), 735–746. <https://doi.org/10.1128/IAI.00942-15>
- Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1), W160–W165. <https://doi.org/10.1093/nar/gkw257>
- Regulski, E. E., & Breaker, R. R. (2008). In-Line Probing Analysis of Riboswitches. In J. Wilusz (Ed.), *Post-Transcriptional Gene Regulation* (pp. 53–67). Humana Press. https://doi.org/10.1007/978-1-59745-033-1_4
- Rehren, G., Walters, S., Fontan, P., Smith, I., & Zárraga, A. M. (2007). Differential gene expression between *Mycobacterium bovis* and *Mycobacterium tuberculosis*. *Tuberculosis*, 87(4), 347–359. <https://doi.org/10.1016/j.tube.2007.02.004>
- Reiss, D. J., Baliga, N. S., & Bonneau, R. (2006). Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, 7, 1–22. <https://doi.org/10.1186/1471-2105-7-280>

- Rengarajan, J., Bloom, B. R., & Rubin, E. J. (2005). Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. *Proceedings of the National Academy of Sciences*, 102(23), 8327–8332. <https://doi.org/10.1073/pnas.0503272102>
- Richards, J., & Belasco, J. G. (2019). Obstacles to Scanning by RNase E Govern Bacterial mRNA Lifetimes by Hindering Access to Distal Cleavage Sites. *Molecular Cell*, 74(2), 284–295.e5. <https://doi.org/10.1016/j.molcel.2019.01.044>
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47–e47. <https://doi.org/10.1093/nar/gkv007>
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1), 24–26. <https://doi.org/10.1038/nbt.1754>
- Rock, J. M., Hopkins, F. F., Chavez, A., Diallo, M., Chase, M. R., Gerrick, E. R., Pritchard, J. R., Church, G. M., Rubin, E. J., Sassetti, C. M., Schnappinger, D., & Fortune, S. M. (2017). Programmable transcriptional repression in mycobacteria using an orthogonal CRISPR interference platform. *Nature Microbiology*, 2(February), 1–9. <https://doi.org/10.1038/nmicrobiol.2016.274>
- Rodriguez G. Marcela, Voskuil Martin I., Gold Benjamin, Schoolnik Gary K., & Smith Issar. (2002). *ideR*, an Essential Gene in *Mycobacterium tuberculosis*: Role of IdeR in Iron-Dependent Gene Expression, Iron Metabolism, and Oxidative Stress Response. *Infection and Immunity*, 70(7), 3371–3381. <https://doi.org/10.1128/iai.70.7.3371-3381.2002>
- Rohde, K., Yates, R. M., Purdy, G. E., & Russell, D. G. (2007). *Mycobacterium tuberculosis* and the environment within the phagosome. *Immunological Reviews*, 219(1), 37–54. <https://doi.org/10.1111/j.1600-065X.2007.00547.x>
- Ruiz de los Mozos, I., Vergara-Irigaray, M., Segura, V., Villanueva, M., Bitarte, N., Saramago, M., Domingues, S., Arraiano, C. M., Fechter, P., Romby, P., Valle, J., Solano, C., Lasa, I., & Toledo-Arana, A. (2013). Base Pairing Interaction between 5'- and 3'-UTRs Controls *icaR* mRNA Translation in *Staphylococcus*

- aureus*. *PLOS Genetics*, 9(12), e1004001.
<https://doi.org/10.1371/journal.pgen.1004001>
- Rustad, T. R., Harrell, M. I., Liao, R., & Sherman, D. R. (2008). The enduring hypoxic response of *Mycobacterium tuberculosis*. *PLoS ONE*, 3(1), 1–8.
<https://doi.org/10.1371/journal.pone.0001502>
- Rustad, T. R., Minch, K. J., Ma, S., Winkler, J. K., Hobbs, S., Hickey, M., Brabant, W., Turkarslan, S., Price, N. D., Baliga, N. S., & Sherman, D. R. (2014). Mapping and manipulating the *Mycobacterium tuberculosis* transcriptome using a transcription factor overexpression-derived regulatory network. *Genome Biology*, 15(11), 502. <https://doi.org/10.1186/s13059-014-0502-3>
- Rustad, T. R., Roberts, D. M., Liao, R. P., & Sherman, D. R. (2009). Isolation of mycobacterial RNA. *Methods in Molecular Biology (Clifton, N.J.)*, 465, 13–21.
https://doi.org/10.1007/978-1-59745-207-6_2
- Ryndak, M., Wang, S., & Smith, I. (2008). PhoP, a key player in *Mycobacterium tuberculosis* virulence. *Trends in Microbiology*, 16(11), 528–534.
<https://doi.org/10.1016/j.tim.2008.08.006>
- Ryndak, M., Wang, S., Smith, I., & Rodriguez, G. (2010). The *Mycobacterium tuberculosis* High-Affinity Iron Importer, IrtA, Contains an FAD-Binding Domain. *Journal of Bacteriology*, 192(3), 861–869.
<https://doi.org/10.1128/jb.00223-09>
- Sabio y García, J., Bigi, M. M., Klepp, L. I., García, E. A., Blanco, F. C., & Bigi, F. (2020). Does *Mycobacterium bovis* persist in cattle in a non-replicative latent state as *Mycobacterium tuberculosis* in human beings? *Veterinary Microbiology*, 247(June), 108758. <https://doi.org/10.1016/j.vetmic.2020.108758>
- Saelens, W., Cannoodt, R., & Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. *Nature Communications*, 9(1), 1090. <https://doi.org/10.1038/s41467-018-03424-4>
- Sáenz-Lahoya S., Bitarte N., García B., Burgui S., Vergara-Irigaray M., Valle J., Solano C., Toledo-Arana A., & Lasa I. (2019). Noncontiguous operon is a genetic organization for coordinating bacterial gene expression. *Proceedings of the National Academy of Sciences*, 116(5), 1733–1738.
<https://doi.org/10.1073/pnas.1812746116>

- Sakurai, I., Stazic, D., Eisenhut, M., Vuorio, E., Steglich, C., Hess, W. R., & Aro, E. M. (2012). Positive regulation of *psbA* gene expression by cis-encoded antisense RNAs in *Synechocystis* sp. PCC 6803. *Plant Physiology*, 160(2), 1000–1010. <https://doi.org/10.1104/pp.112.202127>
- Sala, A., Bordes, P., & Genevau, P. (2014). Multiple toxin-antitoxin systems in *Mycobacterium tuberculosis*. *Toxins*, 6(3), 1002–1020. <https://doi.org/10.3390/toxins6031002>
- Sala Claudia, Forti Francesca, Di Florio Elisabetta, Canneva Fabio, Milano Anna, Riccardi Giovanna, & Ghisotti Daniela. (2003). *Mycobacterium tuberculosis* FurA Autoregulates Its Own Expression. *Journal of Bacteriology*, 185(18), 5357–5362. <https://doi.org/10.1128/jb.185.18.5357-5362.2003>
- Sambou, T., Dinadayala, P., Stadthagen, G., Barilone, N., Bordat, Y., Constant, P., Levillain, F., Neyrolles, O., Gicquel, B., Lemassu, A., Daffé, M., & Jackson, M. (2008). Capsular glucan and intracellular glycogen of *Mycobacterium tuberculosis*: Biosynthesis and impact on the persistence in mice. *Molecular Microbiology*, 70(3), 762–774. <https://doi.org/10.1111/j.1365-2958.2008.06445.x>
- Santa Maria, J. P., Sadaka, A., Moussa, S. H., Brown, S., Zhang, Y. J., Rubin, E. J., Gilmore, M. S., & Walker, S. (2014). Compound-gene interaction mapping reveals distinct roles for *Staphylococcus aureus* teichoic acids. *Proceedings of the National Academy of Sciences of the United States of America*, 111(34), 12510–12515. <https://doi.org/10.1073/pnas.1404099111>
- Sapriel, G., Brosch, R., & Bapteste, E. (2019). Shared Pathogenomic Patterns Characterize a New Phylotype, Revealing Transition toward Host-Adaptation Long before Speciation of *Mycobacterium tuberculosis*. *Genome Biology and Evolution*, 11(8), 2420–2438. <https://doi.org/10.1093/gbe/evz162>
- Sassetti, C. M., & Rubin, E. J. (2003). Genetic requirements for mycobacterial survival during infection. *Proc Natl Acad Sci U S A*, 100(22), 12989–12994. <https://doi.org/10.1073/pnas.2134250100>
- Sawyer, E. B., Phelan, J. E., Clark, T. G., & Cortes, T. (2021). A snapshot of translation in *Mycobacterium tuberculosis* during exponential growth and nutrient starvation revealed by ribosome profiling. *Cell Reports*, 34(5). <https://doi.org/10.1016/j.celrep.2021.108695>

- Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., ... Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 50(D1), D20–D26. <https://doi.org/10.1093/nar/gkab1112>
- Sawyer, J., Rhodes, S., Jones, G. J., Hogarth, P. J., & Vordermeier, H. M. (2023). *Mycobacterium bovis* and its impact on human and animal tuberculosis. *Journal of Medical Microbiology*, 72(11), 1–7. <https://doi.org/10.1099/jmm.0.001769>
- Schlievert, P. M., Merriman, J. A., Salgado-Pabón, W., Mueller, E. A., Spaulding, A. R., Vu, B. G., Chuang-Smith, O. N., Kohler, P. L., & Kirby, J. R. (2013). Menaquinone analogs inhibit growth of bacterial pathogens. *Antimicrobial Agents and Chemotherapy*, 57(11), 5432–5437. <https://doi.org/10.1128/AAC.01279-13>
- Schwenk, S., & Arnvig, K. B. (2018). Regulatory RNA in *Mycobacterium tuberculosis*, back to basics. *Pathogens and Disease*, 76(4). <https://doi.org/10.1093/femspd/fty035>
- Serafini, A., Pisu, D., Palù, G., Rodriguez, G. M., & Manganelli, R. (2013). The ESX-3 Secretion System Is Necessary for Iron and Zinc Homeostasis in *Mycobacterium tuberculosis*. *PLoS ONE*, 8(10), 1–15. <https://doi.org/10.1371/journal.pone.0078351>
- Serafini, A., Tan, L., Horswell, S., Howell, S., Greenwood, D. J., Hunt, D. M., Phan, M.-D., Schembri, M., Monteleone, M., Montague, C. R., Britton, W., Garza-Garcia, A., Snijders, A. P., VanderVen, B., Gutierrez, M. G., West, N. P., & de Carvalho, L. P. S. (2019). *Mycobacterium tuberculosis* requires glyoxylate shunt and reverse methylcitrate cycle for lactate and pyruvate metabolism. *Molecular Microbiology*, 112(4), 1284–1307. PubMed. <https://doi.org/10.1111/mmi.14362>
- Sesto, N., Wurtzel, O., Archambaud, C., Sorek, R., & Cossart, P. (2013). The excludon: A new concept in bacterial antisense RNA-mediated gene regulation. *Nature Reviews Microbiology*, 11(2), 75–82. <https://doi.org/10.1038/nrmicro2934>
- Shaku, M. T., Ocius, K. L., Apostolos, A. J., Pires, M. M., VanNieuwenhze, M. S., Dhar, N., & Kana, B. D. (2023). Amidation of glutamate residues in mycobacterial

peptidoglycan is essential for cell wall cross-linking. *Frontiers in Cellular and Infection Microbiology*, 13(AUG).

<https://doi.org/10.3389/fcimb.2023.1205829>

Shell, S. S., Wang, J., Lapierre, P., Mir, M., Chase, M. R., Pyle, M. M., Gawande, R., Ahmad, R., Sarracino, D. A., Ioerger, T. R., Fortune, S. M., Derbyshire, K. M., Wade, J. T., & Gray, T. A. (2015). Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLOS Genetics*, 11(11), e1005641.

<https://doi.org/10.1371/journal.pgen.1005641>

Shockey, A. C., Dabney, J., & Pepperell, C. S. (2019). Effects of Host, Sample, and in vitro Culture on Genomic Diversity of Pathogenic Mycobacteria. In *Frontiers in Genetics* (Vol. 10).

<https://www.frontiersin.org/article/10.3389/fgene.2019.00477>

Šíková, M., Janoušková, M., Ramaniuk, O., Páleníková, P., Pospíšil, J., Bartl, P., Suder, A., Pajer, P., Kubičková, P., Pavliš, O., Hradilová, M., Vítovská, D., Šanderová, H., Převorovský, M., Hnilicová, J., & Krásný, L. (2019). Ms1 RNA increases the amount of RNA polymerase in *Mycobacterium smegmatis*. *Molecular Microbiology*, 111(2), 354–372. <https://doi.org/10.1111/mmi.14159>

Singh, A., Crossman, D. K., Mai, D., Guidry, L., Voskuil, M. I., Renfrow, M. B., & Steyn, A. J. C. (2009). *Mycobacterium tuberculosis* WhiB3 Maintains redox homeostasis by regulating virulence lipid anabolism to modulate macrophage response. *PLoS Pathogens*, 5(8).

<https://doi.org/10.1371/journal.ppat.1000545>

Singh, A. K., Carette, X., Potluri, L. P., Sharp, J. D., Xu, R., Pristic, S., & Husson, R. N. (2016). Investigating essential gene function in *Mycobacterium tuberculosis* using an efficient CRISPR interference system. *Nucleic Acids Research*, 44(18). <https://doi.org/10.1093/nar/gkw625>

Singh Prabhat Ranjan, Vijjamarri Anil Kumar, Sarkar Dibyendu, & Federle Michael J. (2020). Metabolic Switching of *Mycobacterium tuberculosis* during Hypoxia Is Controlled by the Virulence Regulator PhoP. *Journal of Bacteriology*, 202(7), e00705-19. <https://doi.org/10.1128/JB.00705-19>

Singh, S. K., & Husain, S. M. (2018). A Redox-Based Superoxide Generation System Using Quinone/Quinone Reductase. *ChemBioChem*, 19(15), 1657–1663.

<https://doi.org/10.1002/cbic.201800071>

- Singh, S., Sevalkar, R. R., Sarkar, D., & Karthikeyan, S. (2018). Characteristics of the essential pathogenicity factor Rv1828, a MerR family transcription regulator from *Mycobacterium tuberculosis*. *FEBS Journal*, 285(23), 4424–4444. <https://doi.org/10.1111/febs.14676>
- Smith, A. (2017). *Using Transposon Mutagenesis of Mycobacterium bovis BCG to Identify Candidate Molecules for Novel Control Approaches for Bovine Tuberculosis*. (99514915902346) [Doctoral thesis, University of Surrey]
- Smith, A. A., Villarreal-Ramos, B., Mendum, T. A., Williams, K. J., Jones, G. J., Wu, H., McFadden, J., Vordermeier, H. M., & Stewart, G. R. (2020). Genetic screening for the protective antigenic targets of BCG vaccination. *Tuberculosis*, 124(July), 101979. <https://doi.org/10.1016/j.tube.2020.101979>
- Smith, C., Canestrari, J. G., Wang, A. J., Champion, M. M., Derbyshire, K. M., Gray, T. A., & Wade, J. T. (2022). Pervasive translation in *Mycobacterium tuberculosis*. *eLife*, 11, e73980. <https://doi.org/10.7554/eLife.73980>
- Smith, C. M., Baker, R. E., Proulx, M. K., Mishra, B. B., Long, J. E., Park, S. W., Lee, H.-N., Kiritsy, M. C., Bellerose, M. M., Olive, A. J., Murphy, K. C., Papavinasasundaram, K., Boehm, F. J., Reames, C. J., Meade, R. K., Hampton, B. K., Linnertz, C. L., Shaw, G. D., Hock, P., ... Sassetti, C. M. (2022). Host-pathogen genetic interactions underlie tuberculosis susceptibility in genetically diverse mice. *eLife*, 11. <https://doi.org/10.7554/elife.74419>
- Sohaskey, C. D., & Modesti, L. (2009). Differences in nitrate reduction between *Mycobacterium tuberculosis* and *Mycobacterium bovis* are due to differential expression of both *narGHJI* and *narK2*. *FEMS Microbiology Letters*, 290(2), 129–134. <https://doi.org/10.1111/j.1574-6968.2008.01424.x>
- Solans, L., Gonzalo-Asensio, J., Sala, C., Benjak, A., Uplekar, S., Rougemont, J., Guilhot, C., Malaga, W., Martín, C., & Cole, S. T. (2014). The PhoP-Dependent ncRNA Mcr7 Modulates the TAT Secretion System in *Mycobacterium tuberculosis*. *PLOS Pathogens*, 10(5), e1004183. <https://doi.org/10.1371/journal.ppat.1004183>
- Song, T., Song, S. E., Raman, S., Anaya, M., & Husson, R. N. (2008). Critical role of a single position in the -35 element for promoter recognition by *Mycobacterium tuberculosis* SigE and SigH. *Journal of Bacteriology*, 190(6), 2227–2230. <https://doi.org/10.1128/JB.01642-07>

- Sreenu, V. B., Kumar, P., Nagaraju, J., & Nagarajaram, H. A. (2007). Simple sequence repeats in mycobacterial genomes. *Journal of Biosciences*, 32(1), 3–15.
<https://doi.org/10.1007/s12038-007-0002-7>
- Sridhar, J., Narmada, S. R., Sabarinathan, R., Ou, H. Y., Deng, Z., Sekar, K., Rafi, Z. A., & Rajakumar, K. (2010). sRNAsScanner: A computational tool for intergenic small RNA detection in bacterial genomes. *PLoS ONE*, 5(8).
<https://doi.org/10.1371/journal.pone.0011970>
- Stiens, J., Tan, Y. Y., Joyce, R., Arnvig, K. B., Kendall, S. L., & Nobeli, I. (2023). Using a whole genome co-expression network to inform the functional characterisation of predicted genomic elements from *Mycobacterium tuberculosis* transcriptomic data. *Molecular Microbiology*, 119(4), 381–400.
<https://doi.org/10.1111/mmi.15055>
- Stupar, M., Furness, J., De Voss, C. J., Tan, L., & West, N. P. (2022). Two-component sensor histidine kinases of *Mycobacterium tuberculosis*: Beacons for niche navigation. *Molecular Microbiology*, 117(5), 973–985.
<https://doi.org/10.1111/mmi.14899>
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43), 15545–15550.
<https://doi.org/10.1073/pnas.0506580102>
- Subramaniam, S., DeJesus, M. A., Zaveri, A., Smith, C. M., Baker, R. E., Ehrt, S., Schnappinger, D., Sassetti, C. M., & Ioerger, T. R. (2019). Statistical analysis of variability in TnSeq data across conditions using zero-inflated negative binomial regression. *BMC Bioinformatics*, 20(1), 603.
<https://doi.org/10.1186/s12859-019-3156-z>
- Talwar, S., Pandey, M., Sharma, C., Kutum, R., Lum, J., Carbajo, D., Goel, R., Poidinger, M., Dash, D., Singhal, A., & Pandey, A. K. (2020). Role of VapBC12 Toxin-Antitoxin Locus in Cholesterol-Induced Mycobacterial Persistence. *mSystems*, 5(6). <https://doi.org/10.1128/msystems.00855-20>
- Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Research*, 21(12), 2213–2223. <https://doi.org/10.1101/gr.124321.111>

- Tateishi, Y., Minato, Y., Baughn, A. D., Ohnishi, H., Nishiyama, A., Ozeki, Y., & Matsumoto, S. (2020). Genome-wide identification of essential genes in *Mycobacterium intracellulare* by transposon sequencing—Implication for metabolic remodeling. *Scientific Reports*, 10(1), 5449.
<https://doi.org/10.1038/s41598-020-62287-2>
- The Gene Ontology Consortium. (2021). The Gene Ontology resource: Enriching a Gold mine. *Nucleic Acids Research*, 49(D1), D325–D334.
<https://doi.org/10.1093/nar/gkaa1113>
- Tjaden, B. (2023). TargetRNA3: Predicting prokaryotic RNA regulatory targets with machine learning. *Genome Biology*, 24(1), 276.
<https://doi.org/10.1186/s13059-023-03117-2>
- Toffano-Nioche, C., Luo, Y., Kuchly, C., Wallon, C., Steinbach, D., Zytnicki, M., Jacq, A., & Gautheret, D. (2013). Detection of non-coding RNA in bacteria and archaea using the DETR'PROK Galaxy pipeline. *Methods*, 63(1), 60–65.
<https://doi.org/10.1016/j.ymeth.2013.06.003>
- Toledo-Arana, A., & Lasa, I. (2020). Advances in bacterial transcriptome understanding: From overlapping transcription to the excludon concept. *Molecular Microbiology*, 113(3), 593–602.
<https://doi.org/10.1111/mmi.14456>
- Trinquier, A., Durand, S., Braun, F., & Condon, C. (2020). Regulation of RNA processing and degradation in bacteria. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1863(5), 194505.
<https://doi.org/10.1016/j.bbagrm.2020.194505>
- Tripathi, D., Chandra, H., & Bhatnagar, R. (2013). Poly-L-glutamate/glutamine synthesis in the cell wall of *Mycobacterium bovis* is regulated in response to nitrogen availability. *BMC Microbiology*, 13(1), 226.
<https://doi.org/10.1186/1471-2180-13-226>
- Tullius, Michael V., Nava Susana, & Horwitz Marcus A. (2019). PPE37 Is Essential for *Mycobacterium tuberculosis* Heme-Iron Acquisition (HIA), and a Defective PPE37 in *Mycobacterium bovis* BCG Prevents HIA. *Infection and Immunity*, 87(2), 10.1128/iai.00540-18.
<https://doi.org/10.1128/iai.00540-18>
- Upadhyay, A., Fontes, F. L., Gonzalez-Juarrero, M., McNeil, M. R., Crans, D. C., Jackson, M., & Crick, D. C. (2015). Partial Saturation of Menaquinone in

- Mycobacterium tuberculosis*: Function and Essentiality of a Novel Reductase, MenJ. *ACS Central Science*, 1(6), 292–302.
<https://doi.org/10.1021/acscentsci.5b00212>
- Updegrove, T. B., Kouse, A. B., Bandyra, K. J., & Storz, G. (2019). Stem-loops direct precise processing of 3' UTR-derived small RNA MicL. *Nucleic Acids Research*, 47(3), 1482–1492. <https://doi.org/10.1093/nar/gky1175>
- Urtasun-Elizari, J. M., Ma, R., Pickford, H., Farrell, D., Gonzalez, G., Perets, V., Nakajima, C., Suzuki, Y., MacHugh, D. E., Bhatt, A., & Gordon, S. V. (2024). Functional analysis of the *Mycobacterium bovis* AF2122/97 PhoPR system. *Tuberculosis*, 148(April), 102544.
<https://doi.org/10.1016/j.tube.2024.102544>
- van Opijnen, T., Bodi, K. L., & Camilli, A. (2009). Tn-seq: High-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature Methods*, 6(10), 767–772. <https://doi.org/10.1038/nmeth.1377>
- Vandal, O. H., Roberts, J. A., Odaira, T., Schnappinger, D., Nathan, C. F., & Ehrt, S. (2009). Acid-susceptible mutants of *Mycobacterium tuberculosis* share hypersusceptibility to cell wall and oxidative stress and to the host environment. *Journal of Bacteriology*, 191(2), 625–631.
<https://doi.org/10.1128/JB.00932-08>
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., ... Velankar, S. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1), D439–D444.
<https://doi.org/10.1093/nar/gkab1061>
- Vargas-Blanco, D. A., & Shell, S. S. (2020). Regulation of mRNA Stability During Bacterial Stress Responses. *Frontiers in Microbiology*, 11(September).
<https://doi.org/10.3389/fmicb.2020.02111>
- Vashist, A., Malhotra, V., Sharma, G., Tyagi, J. S., & Clark-Curtiss, J. E. (2018). Interplay of PhoP and DevR response regulators defines expression of the dormancy regulon in virulent *Mycobacterium tuberculosis*. *Journal of Biological Chemistry*, 293(42), 16413–16425.
<https://doi.org/10.1074/jbc.RA118.004331>

- Viljoen, A. J., Kirsten, C. J., Baker, B., van Helden, P. D., & Wiid, I. J. F. (2013). The Role of Glutamine Oxoglutarate Aminotransferase and Glutamate Dehydrogenase in Nitrogen Metabolism in *Mycobacterium bovis* BCG. *PLOS ONE*, 8(12), e84452. <https://doi.org/10.1371/journal.pone.0084452>
- 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. <https://doi.org/10.1016/j.tube.2004.02.003>
- Wade, J. T., & Grainger, D. C. (2014). Pervasive transcription: Illuminating the dark matter of bacterial transcriptomes. *Nature Reviews Microbiology*, 12(9), 647–653. <https://doi.org/10.1038/nrmicro3316>
- Walters, S. B., Dubnau, E., Kolesnikova, I., Laval, F., Daffe, M., & Smith, I. (2006). The *Mycobacterium tuberculosis* PhoPR two-component system regulates genes essential for virulence and complex lipid biosynthesis. *Molecular Microbiology*, 60(2), 312–330. <https://doi.org/10.1111/j.1365-2958.2006.05102.x>
- Wang, M., Fleming, J., Li, Z., Li, C., Zhang, H., Xue, Y., Chen, M., Zhang, Z., Zhang, X. E., & Bi, L. (2016). An automated approach for global identification of sRNA-encoding regions in RNA-Seq data from *Mycobacterium tuberculosis*. *Acta Biochimica et Biophysica Sinica*, 48(6), 544–553. <https://doi.org/10.1093/abbs/gmw037>
- Wang, Q., Boshoff, H. I. M., Harrison, J. R., Ray, P. C., Green, S. R., Wyatt, P. G., & iii, C., E. Barry. (2020). PE/PPE proteins mediate nutrient transport across the outer membrane of *Mycobacterium tuberculosis*. *Science*, 367(6482), 1147–1151. <https://doi.org/10.1126/science.aav5912>
- Wang, X., Monford Paul Abishek, N., Jeon, H. J., Lee, Y., He, J., Adhya, S., & Lim, H. M. (2019). Processing generates 3' ends of RNA masking transcription termination events in prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10), 4440–4445. <https://doi.org/10.1073/pnas.1813181116>
- Warman, E. A., Forrest, D., Guest, T., Haycocks, J. J. R. J., Wade, J. T., & Grainger, D. C. (2021). Widespread divergent transcription from bacterial and archaeal promoters is a consequence of DNA-sequence symmetry. *Nature*

Microbiology, 6(6), 746–756. <https://doi.org/10.1038/s41564-021-00898-9>

- Warner, D. F., Savvi, S., Mizrahi, V., & Dawes, S. S. (2007). A Riboswitch Regulates Expression of the Coenzyme B12-Independent Methionine Synthase in *Mycobacterium tuberculosis*: Implications for Differential Methionine Synthase Function in Strains H37Rv and CDC1551. *Journal of Bacteriology*, 189(9), 3655 LP – 3659. <https://doi.org/10.1128/JB.00040-07>
- Waters, W. R., Whelan, A. O., Lyashchenko, K. P., Greenwald, R., Palmer, Harris, B. N., Hewinson, R. G., & Vordermeier, H. M. (2010). Immune Responses in Cattle Inoculated with *Mycobacterium bovis*, *Mycobacterium tuberculosis*, or *Mycobacterium kansasii*. *Clinical and Vaccine Immunology*, 17(2), 247–252. <https://doi.org/10.1128/CVI.00442-09>
- Whitaker, M., Ruecker, N., Hartman, T., Klevorn, T., Andres, J., Kim, J., Rhee, K., & Ehrt, S. (2020). Two Interacting ATPases Protect *Mycobacterium tuberculosis* from Glycerol and Nitric Oxide Toxicity. *Journal of Bacteriology*, 202(16), 10.1128/jb.00202-20. <https://doi.org/10.1128/jb.00202-20>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wickham, H., François, R., Henry, L., & Müller, K. (2022). *dplyr: A Grammar of Data Manipulation*. <https://dplyr.tidyverse.org/>
- Widdison, S., Watson, M., Piercy, J., Howard, C., & Coffey, T. J. (2008). Granulocyte chemotactic properties of *M. tuberculosis* versus *M. bovis*-infected bovine alveolar macrophages. *Molecular Immunology* 45(3), 740–749. <https://doi.org/10.1016/j.molimm.2007.06.357>
- Wolf, Y. I., & Koonin, E. V. (2012). A Tight Link between Orthologs and Bidirectional Best Hits in Bacterial and Archaeal Genomes. *Genome Biology and Evolution*, 4(12), 1286–1294. <https://doi.org/10.1093/gbe/evs100>
- Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3). <https://doi.org/10.1016/j.xinn.2021.100141>
- Xin, F., Bei, X., Weishan, Z., Ning, L., Kaixia, M., & Beinan, W. (2023). Coevolution of *furA*-Regulated Hyper-Inflammation and Mycobacterial Resistance to

- Oxidative Killing through Adaptation to Hydrogen Peroxide. *Microbiology Spectrum*, 11(4), e05367-22. <https://doi.org/10.1128/spectrum.05367-22>
- Xing, D., Ryndak, M. B., Wang, L., Kolesnikova, I., Smith, I., & Wang, S. (2017). Asymmetric Structure of the Dimerization Domain of PhoR, a Sensor Kinase Important for the Virulence of *Mycobacterium tuberculosis*. *ACS Omega*, 2(7), 3509–3517. <https://doi.org/10.1021/acsomega.7b00612>
- Yao, Z., Hangya, L., Haoxiang, L., Jingyi, L., Yunfeng, Y., Feiyun, L., Wenliang, H., Zhemin, Z., Ping, W., & Shengmin, Z. (2021). Nitroreductase Increases Menadione-Mediated Oxidative Stress in *Aspergillus nidulans*. *Applied and Environmental Microbiology*, 87(24), e01758-21. <https://doi.org/10.1128/AEM.01758-21>
- Yoo, R., Rychel, K., Poudel, S., Al-bulushi, T., Yuan, Y., Chauhan, S., Lamoureux, C., Palsson, B. O., & Sastry, A. (2022). Machine Learning of All *Mycobacterium tuberculosis* H37Rv RNA-seq Data Reveals a Structured Interplay between Metabolism, Stress Response, and Infection. *mSphere*, 7(2), e00033-22. <https://doi.org/10.1128/msphere.00033-22>
- Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, 16(5), 284–287. <https://doi.org/10.1089/omi.2011.0118>
- Yu, S.-H., Vogel, J., & Förstner, K. U. (2018). ANNOgesic: A Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes. *GigaScience*, 7(9). <https://doi.org/10.1093/gigascience/giy096>
- Zahrt, T. C., Song, J., Siple, J., & Deretic, V. (2001). Mycobacterial FurA is a negative regulator of catalase–peroxidase gene *katG*. *Molecular Microbiology*, 39(5), 1174–1185. <https://doi.org/10.1111/j.1365-2958.2001.02321.x>
- Zhang, B., & Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1). <https://doi.org/10.2202/1544-6115.1128>
- Zhang, L., Hendrickson, R. C., Meikle, V., Lefkowitz, E. J., Ioerger, T. R., & Niederweis, M. (2020). Comprehensive analysis of iron utilization by *Mycobacterium tuberculosis*. *PLOS Pathogens*, 16(2), e1008337. <https://doi.org/10.1371/journal.ppat.1008337>

- Zhang, L., Kent, J. E., Whitaker, M., Young, D. C., Herrmann, D., Aleshin, A. E., Ko, Y.-H., Cingolani, G., Saad, J. S., Moody, D. B., Marassi, F. M., Ehrt, S., & Niederweis, M. (2022). A periplasmic cinched protein is required for siderophore secretion and virulence of *Mycobacterium tuberculosis*. *Nature Communications*, 13(1), 2255. <https://doi.org/10.1038/s41467-022-29873-6>
- Zhang, Y. J., Ioerger, T. R., Huttenhower, C., Long, J. E., Sassetti, C. M., Sacchettini, J. C., & Rubin, E. J. (2012). Global Assessment of Genomic Regions Required for Growth in *Mycobacterium tuberculosis*. *PLoS Pathogens*, 8(9). <https://doi.org/10.1371/journal.ppat.1002946>
- Zhou, Y., Huang, H., Zhou, P., & Xie, J. (2012). Molecular mechanisms underlying the function diversity of transcriptional factor IclR family. *Cellular Signalling*, 24(6), 1270–1275. <https://doi.org/10.1016/j.cellsig.2012.02.008>
- Zhou, Y., Sun, H., Rapiejko, A. R., Vargas-Blanco, D. A., Martini, M. C., Chase, M. R., Joubran, S. R., Davis, A. B., Dainis, J. P., Kelly, J. M., Ioerger, T. R., Roberts, L. A., Fortune, S. M., & Shell, S. S. (2023). Mycobacterial RNase E cleaves with a distinct sequence preference and controls the degradation rates of most *Mycobacterium smegmatis* mRNAs. *Journal of Biological Chemistry*, 299(11). <https://doi.org/10.1016/j.jbc.2023.105312>
- Zondervan, N. A., Van Dam, J. C. J., Schaap, P. J., Martins dos Santos, V. A. P., & Suarez-Diez, M. (2018). Regulation of Three Virulence Strategies of *Mycobacterium tuberculosis*: A Success Story. *International Journal of Molecular Sciences*, 19(2). <https://doi.org/10.3390/ijms19020347>
- Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13), 3406–3415. <https://doi.org/10.1093/nar/gkg595>