Transcription analysis of virulent strains of *Mycobacterium tuberculosis*

by
Jon Mitchell Ambler
AMBJON001

SUBMITTED TO THE UNIVERSITY OF CAPE TOWN
In fulfilment of the requirements for the degree

Doctor of Philosophy

Department of Integrative Biomedical Sciences
University of Cape Town
South Africa

Supervisor: Nicola Mulder
May 31, 2018

Abstract

Background: Despite the development of new drugs and success of social programs, tuberculosis remains a leading cause of mortality. This burden falls disproportionately on developing countries where the high burden of HIV has a potentiating effect, but may soon return to areas where it was previously brought under control as resistant strains continue to emerge. In the Western Cape, two closely related strains of the Beijing family have been isolated that provide an opportunity to study virulence in a system with relatively little noise. The aim of this project was to identify the cause of the altered virulence displayed between the two strains, and describe how the differences between the two genomes contributed to the phenotypic differences.

Results: GenGraph allows for the creation of graph genomes, and facilitated the creation of a pan-transcriptome that allowed for the mapping of gene annotations between isolates. This allowed for the mapping of reads to a more suitable Beijing family reference while interpreting the results with annotations from the H37Rv reference. We generated expression and target profiles for the known sRNA, and identified a large number of novel sRNA. Transcriptomic data from 4 different growth conditions was integrated with this sRNA data as well as variant data using the Cell pipeline. From this data we identified multiple sets of genes linked to copper sensing in MTB, including the differentially expressed MoCo operon. Increasing evidence that macrophages use copper to poison bacteria trapped in their phagosomes provides the link to virulence and pathogenicity.

Conclusions: Through the integration of data from multiple data types we were able to elucidate the most probable cause of the altered virulence found between the two isolates

in this study. We developed reusable tools and pipelines, and noted a large number of undescribed sRNA expressed in these isolates. The identification of the copper response as a chief contributor to the phenotype increases both our understanding of the isolates, and the role of the element in infection. These results will be key in guiding further investigation of the variant linked genes to identify those linked to copper homeostasis or response.

"Are you done yet?"

Ian Ambler, my beloved father, almost every day from 2014 - 2017.

Declaration

I, Jon Mitchell Ambler, hereby declare that the work on which this thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Signed by candidate

Date: 31 May 2018

Acknowledgements

We would like to thank the National Research Foundation of South Africa for their financial support of this research, as well as the University of Cape Town for use of their facilities. This work was funded by the National Research Foundation of South Africa, grant number 86934.

We would also like to thank Rob Warren, Samantha Samson and the rest of our collaborators at the University of Stellenbosch for their help in the preparation and growth of the cultures, as well as Jonathan Featherson from the Agricultural Research Council for doing the sequencing and his insight. Additional thanks to Jonathan Blackburn at the University of Cape Town, and Thys Potgieter for his work on the proteomic data.

I would like to personally thank my supervisor, Nicola Mulder, for her support and guidence during this project. She has created an amazing space at CBio where researchers have the resources and opportunities to grow, and a hub for Bioinformatics both globally and here at home in South Africa.

A big thanks also to the members of the CBio labs for their support, friendship, and enthusiasm for Thursday afternoon coffee sessions. Many of you have become close friends over the years, sharing in travels foreign lands, adventures around Cape Town, a few weddings, and plenty of beers.

And last but not least, to my friends and family, for their patience, and support, and love over the years. People often speak of the financial cost of a degree, but the greatest cost comes in the lost time together. In the weeks that I hardly emerged from my room and when I did, was too tired to be any real company. In the weekends worked and the special

times missed. So I thank you for bearing with me through the isolation, and hope to make up those lost times tenfold.

Far more important than good data, or access to compute clusters, this work would never have been completed without these people and their love and guidance over the years.

Contents

List of Abbreviations

API Application programming interface.

BLAST Basic local alignment search tool.

CDS Coding sequences.

DAVID Database for annotation, visualisation, and integrated discovery.

Elog Early logarithmic phase.

FPKM Fragments per kilobase of transcript per million mapped reads.

GFF General feature format.

GO Gene ontology.

GTF Gene transfer format.

GSA Gene set enrichment analysis.

GUI Graphical user interface.

HGT Horizontal gene transfer.

HIV Human immunodeficiency virus.

IGR Intergenic region.

LCA Last common ancestor.

MAPQ Mapping quality.

ML(C) Middle logarithmic growth (Control).

ML(T) Middle logarithmic growth (Treated).

MoCo Molybdenum cofactor.

MTB *Mycobacterium tuberculosis*.

MTBC *Mycobacterium tuberculosis* complex.

ncRNA Non-coding RNA.

NGS Next generation sequencing.

ORF Open reading frame.

PAI Pathogenicity islands.

PCA Principal component analysis.

PNA Principal network analysis.

QC Quality control.

RPKM Reads per kilobase exon per million reads.

SNP Single nucleotide polymorphism.

SRA Sequence Read Archive.

sRNA Small non-coding RNA.

Stat Stationary.

TB Tuberculosis.

TF Transcription factor.

UTR Untranslated region.

VCF Variant call format.

WHO World Health Organization

List of Figures

List of Tables

"They say a little knowledge is a dangerous thing, but it's not one half so bad as a lot of ignorance."

Terry Pratchett

1

A review of integrative approaches to understanding virulence in *Mycobacterium tuberculosis* using next-generation sequencing of coding and non-coding RNA

1.1 Introduction:

1.1.1 The origins of *Mycobacterium tuberculosis*

Mycobacterium tuberculosis is a bacterial intracellular pathogen of the family Mycobacteriaceae, and was discovered by Robert Koch in 1882, who then went on to develop a staining

technique that is still used in diagnostics today. It is part of the *Mycobacterium tuberculosis* complex (MTBC) which includes *M. tuberculosis*, *Mycobacterium bovis*, *Mycobacterium microti*, *Mycobacterium africanum*, *Mycobacterium pinnipedii*, and *Mycobacterium caprae* species. Currently it is believed that *M. tuberculosis* emerged out of *Mycobacterium prototuberculosis*, a progenitor thought to be as old as 3 million years [3], and traveled with modern humans from East Africa (40,000 years ago) as a human pathogen [3]. Two main lineages then arose from this ancestral *M. tuberculosis* 20,000 to 30,000 years ago, one lineage gave rise to the modern *M. tuberculosis* strains, and the second to the human / animal tuberculosis strains including *M. bovis* and *M. africanum* [4, 3]. This version of events puts to rest the theory that tuberculosis was originally a zoonotic disease passed to humans by close contact with livestock, and rather that *M. bovis* arose as a result of humans infecting their livestock [4]. During this time MTB has played a prominent part in the history of the human race, being found in Egyptian mummies [5], devastating Europe where it was referred to as "The consumption" [6], and being mentioned in the works of authors such as Emily Bronte's Wuthering Heights [7]. In modern times MTB has spread across the globe, and diverged into different lineages including the LAM, Haarlem, Beijing, H37, and KZN genotypes that vary in their transmissibility, drug resistance profiles, and pathogenicity [8, 9, 10]. And while the disease has subsided in much of the globe, in countries like South Africa, factors like AIDS have allowed the disease to continue to devastate the population.

Within South Africa, in a local community, hypo-virulent and hyper-virulent MTB was identified by collaborators at the University of Stellenbosch [11], and has become the focus of research owing to their potential to provide valuable insights into the biological underpinnings of virulence in this organism. In this context, virulence referring to the severity of disease the pathogen causes in the host, with the hypo- and hyper- designations assigned from experiments done in murine models by the Stellenbosch research group. The genomes of these strains have been sequenced and compared by the group, revealing 40 single nucleotide polymorphisms (SNPs), 20 of which were non-synonymous. Additionally, studies of the proteome [11], metabolome [12] and phosphorylome of the two strains have been conducted. Despite revealing a number of differences between these strains, the origin of the altered virulence remains uncertain (Rob Warren, personal communication).

The disease: The epidemiology of tuberculosis and drug resistance

Tuberculosis (TB) is the severe respiratory disease caused by *M. tuberculosis* and is one of the leading causes of mortality in the world. According to the World Health Organization (WHO) there were an estimated 10.4 million individuals who developed the disease, and 1.3 million fatalities in 2016 among HIV-negative people. Recently the majority of disease burden has fallen disproportionately on developing countries, with only 22 countries accounting for 80% of disease cases globally [13]. Despite the relatively rapid development of anti-TB drugs, the global rise of HIV and the rapid emergence of resistant strains are greatly impeding the effective eradication of MTB.

The persistence of the disease in African countries is potentiated by the human immunodeficiency virus (HIV) epidemic, where of the 1.1 million HIV-positive people who developed the disease globally, 75% were in Africa. In South Africa, an estimated 10% of the population is thought to be living with HIV (Statistics SA release P0302, 2013). In 2014 around 12% of the global incident TB cases were also co-infected with HIV (1.2 million). This proportion is particularly high in African countries where 32% of TB cases were co-infected with HIV, a figure that is as high as 50% in parts of southern Africa [13]. TB is reported as the most common cause of AIDS related death [14] and infection with HIV increases risk of latent TB reactivation 20-fold [15]. Infection with both TB and HIV has a potentiating effect as both attack the host immune systems. In addition the emergence of XDR (Extensively Drug-Resistant) strains in Kwazulu-Natal in 2006 [16] is of particular concern, both in countries with a high disease burden and countries with established health care methods as existing treatments become ineffective leading to poor treatment outcome and relapse.

Transmission occurs through the inhalation of airborne droplets containing MTB generated by coughing or sneezing of an infected person [17]. Bacteria that reach the alveoli are internalised by macrophages [18]. Once the bacteria are internalised into phagosomes, the macrophages undergo maturation during which the microbes are exposed to a combination of lytic enzymes, acidic conditions, and reactive oxygen and nitrogen intermediates in an environment that is low in the availability of free iron [19, 20]. If the patient's cell-mediated immunity is intact, activated T-lymphocytes and macrophages will begin forming granulo-

mas 2 - 8 weeks after infection [21]. MTB is able to inhibit this phagosomal maturation [22] preventing acidification of the vacuolar environment and allowing infection to persist in a latent state [23]. The hypoxic environment, such as is found within the macrophages or granuloma [24], induces multiple transcriptional responses in the pathogen. These changes include a shift to a non-replicating state through the induction of a dormancy regulon, changes in metabolism including the use of fatty acids, intensification of iron acquisition through the production of siderophores and a switch from aerobic to anaerobic respiration [25]. Additional changes in the MTB cell wall aids in its adaptability to new environments [26] and evasion of the innate immune response [27]. During infection in a murine model, the number of bacilli remain the same, and whether this is because they enter a stationary non-replicating phase of if replication and death are at the same rate is unclear *et. al.* [28, 29]. This complicates analysis of the organism *in vitro*, as virulence factors may only be active during certain growth phases, and we cannot be certain which growth phase best simulates under laboratory conditions.

At first, chemotherapy of MTB was considered impossible due to what appeared to be a impenetrable lipid-rich cell wall [30], but after the first effective compounds were identified, another challenge soon arose in the form of drug resistance. Due to the lack of a plasmid and the inability to initiate horizontal gene transfer, drug resistance in MTB occurs predominantly as a result of mutations. Some examples include Azole resistant MTB, where resistance is a result of an increase in econazole efflux and higher expression of *mmpS5-mmpL5* genes. The increase in expression of the *mmpS5-mmpL5* genes was as a result of a mutation in Rv0678, a potential transcriptional regulator of the genes [31]. Additional challenges to successfully treating the disease are found in the lifestyle of the pathogen. In a review by Gillespie [30], the author quoted multiple obstructions to successful treatment, including the low metabolic activity of the organism during dormancy, inaccessibility of drugs in the tissues of the lungs, and the presence of multiple populations in the same patient which may respond differently to treatment.

Within Cape Town, South Africa, the rapid emergence and spread of W-Beijing strains of MTB has been observed [32, 33]. This strain of MTB is remarkable in its higher disseminative ability [34] and has been noted for its superior fitness while exhibiting increased streptomycin

resistance over other streptomycin resistant strains [35]. Multi-drug resistance has been shown to occur frequently within these strains [36] and progresses in a stepwise manner with susceptible and resistant subpopulations coexisting within the patient over the course of the infection [37]. Other characteristics of this phenotype include mutations in the DNA repair genes (*mut* genes) which is thought to have resulted in an increased genomic mutation rate [38] that may have contributed to the strains' ability to adapt to changing environments [39, 40], and a mycobacterial adenine methyltransferase (*mamA*) gene that is known to be partially inactivated by a point mutation in the Beijing strains. Conversely, a second methyltransferase, HsdM is active in the Beijing strains and inactive in the Euro-American strains [41].

Mutations: The drivers of variation in the genome of *M. tuberculosis*

Since the first MTB genome sequence of the H37Rv strain was released [42], thousands of isolates have been sequenced, with the genomes of over 25 different isolates have been sequenced, annotated, and fully assembled, and publicly available [43, 44]. The genomes of MTB are characterised by high GC content [42], contain between 3,851 - 4,324 genes and are 4.43 - 4.54 Mb in length [44]. The availability of genomes allows for the identification of essential evolutionary conserved genes that are unique to prokaryotes, and make inviting targets for drug development [45]. Although it was initially thought that there was little genetic diversity amongst MTB isolates hinting that this may be an evolutionarily young organism [46], it has since been suggested that there is more diversity than originally thought [47, 48, 49].

Large sequence polymorphisms (LSPs) are one such source of diversity that has been found to be more common than previously thought [50]. The distribution of deletions in the genome appears non-random and concentrated in particular regions [51], with the size of deleted sequences varying from hundreds to tens of thousands of bp [52]. Some appear in only certain lineages, while other clusters of deletions appear in phylogenetically unrelated organisms indicating regions that are predisposed to this type of mutation. These deletions may be beneficial to the organism by reducing the load of mobile genetic elements in the genome, conferring antibiotic resistance, or enhancing transmission through increasing growth rates

or lowering the latency period [52]. It is therefore possible that genome reduction often occurs in pathogens [53, 54] driven by a combination of these beneficial outcomes.

The SAWC5527 and SAWC507 isolates from the Western Cape are part of the Beijing family (also referred to as the East-Asian lineage), whose genomes differ significantly to that of H37Rv. Some characteristics of this lineage include the constitutive over-expression of the DosR regulon [55], as well as a large 350 kb genomic duplication containing over 300 genes, including *dosR* [56], and multiple large scale chromosomal rearrangements within the W-148 genome [50]. In Stellenbosch duplication events were found in some of the locally isolated Beijing strains that were enriched for genes involved in purine and pyrimidine metabolism, DNA metabolism, and repair [57].

Horizontal gene transfer (HGT) is the movement of genetic material between species and has played an important role in the emergence of pathogenicity in members of the MBTC, despite the inability of these bacteria to initiate the exchange themselves. A study by Becq *et. al.,* [58] identified 48 regions that include 256 genes in the *M. tuberculosis* chromosome that have been acquired by HGT, with many of these genomic islands containing genes that have been identified as virulence genes. These HGT events pre-date the clonal expansion of MTB [3] and were mainly acquired from other Actinobacteria that shared the same environment, many of which were soil borne pathogens. As members of the MTBC became obligate pathogens, they no longer shared an environment with these soil borne pathogens, and as a consequence a lack of further HGT is observed among members of the MTBC since this shift to a solitary lifestyle within their hosts. These genomic islands are continuously undergoing rearrangements, and in some species only a single gene from the original integrated cluster remains [58]. Pathogenicity islands (PAI) are a class of genomic island that are defined by Hacker *et. al.,* [59] as having more than one virulence gene, a high occurrence in pathogenic strains, GC contents that differ to the rest of the organism, flanking direct repeats or insertion sequences, and are often unstable showing abnormally high rates of deletion or duplication.

A relatively modern example of HGT has been observed in *M. smegmatis* where IS*6110*, a IS3 family insertion element found exclusively within the MTBC [60], was found. This is thought to have occurred either through an intermediary species or by direct transfer from a member of the MTBC during a time that the two shared an environment [61].

This insertion element has played a significant part in the evolution of MTB, contributing to inter-isolate variation, and were once used as diagnostic markers to identify different strains [62, 42, 60]. An insertion element is a form of transposable element, which is a fragment of DNA that is able to move between different positions in the genome. This insertion element contains two partially overlapping open reading frames, *orfA* and *orfB*, which may produce the OrfAB transposase protein by way of translational frameshifting [63]. IS*6110* is capable of both replicative and non-replicative transposition [64], resulting in different copy numbers in different isolates. Most members of the MTBC contain multiple copies of IS*6110* with the exception of *M. bovis* which generally contains only one copy with limited transposition activity [64]. The movement of insertion elements like IS*6110* is another source of genotypic variation in MTB and many other organisms [65, 66, 67], and while most mutation is deleterious, occasionally it has beneficial effects. Insertion sites appear to be not entirely random leading to the creation of insertion hotspots [68, 69]. An example of this is the apparent preference for insertion of IS*6110* into PE-PGRS genes, which are thought to be surface antigens [64, 70]. The IS*6110* has also been found to have a positive regulatory ability and has been seen altering the expression of known virulence genes [71]. As members of the MTBC lack a plasmid, drug resistance must emerge by different means. IS*6110* activity provides one such avenue that can to lead to the emergence of drug resistance through the disruption of drug targets [72].

1.1.2 Modern techniques for differential expression analysis

High-throughput RNA sequencing allows researches to quantify the expression of genes under a certain set of conditions, which may then be mapped to the genome sequence to infer biological significance. In *M. tuberculosis*, there exists large differences in genes expressed during the different growth stages, with only 421 coding sequences (CDSs) representing 11% of the genome expressed with a RPKM (reads per kilobase exon per million reads) >5 during the stationary phase and 3,136 CDSs above the same threshold at the exponential growth phase, representing 78.4% of the genome [73]. In order to ensure that such observed deviations in gene expression are as a result of biological events and not errors, care must be taken in selecting the correct tools and parameters for differential expression analysis. An evaluation of

the available differential gene expression analysis methods concluded that methods that employ negative binomial modeling showed higher sensitivity and specificity [74]. This includes the tools DESeq [75], edgeR [76], and baySeq [77]. Once differential expression analysis has been conducted, an additional consideration is that the mRNA abundance is not a direct measure of protein abundance and consequently phenotype [78, 79, 80]. A myriad of factors from the presence of metabolite binding riboswitches [81], to altered tertiary structure of the protein as a result of SNPs will influence the eventual effect of altered gene expression on phenotype. These further highlight the need for a multidisciplinary approach to differential expression based studies where data from multiple biological levels is integrated.

1.1.3 The types and function of regulatory non-coding RNA (ncRNA)

In addition to the coding transcriptome of an organism large portions of transcripts produced do not code for proteins, and function in their RNA form [82]. While transfer RNAs (tRNA) and ribosomal RNAs (rRNA) are well known, the roles of 3' and 5' untranslated regions, intergenic small RNAs, and antisense transcripts in modulating gene expression is still under study [83]. Antisense transcripts have been detected both within genes, and as 3' UTR-derived antisense transcripts [73]. They base pair with the 5' region of mRNA and have been shown to alter target mRNA stability, either enhancing its degradation or forming a stabilising duplex [84]. These antisense RNAs are able to affect translation by competitively binding to the ribosomal binding site and/or start codon [85]. The 5' UTR can also contain an element known as a riboswitch, a functional structure that can alter gene expression in response to environmental signals like temperature or the binding of a metabolite [86].

In addition to ORF derived ncRNA, a large number of regulatory small RNA (sRNA) have been identified and described. These are 40 to 500 nucleotide non-coding RNA molecules that can bind proteins altering their function [87] or bind to mRNA to alter gene expression [88, 89]. These sRNA have the ability to mount a rapid response to stimuli, regulating gene expression at the post transcriptional level [90]. Many sRNA are associated with adaption to stress, which in turn relates to virulence [91]. In some organisms, sRNA are directly involved in regulating virulence through altering virulence factor levels [92] while others have a more general regulatory role responding to conditions like oxidative stress or cell wall synthesis

[93]. Some have been identified that play a direct role in pathogenicity, as is the case for Qrr1-4 sRNAs that regulate HapR in *Vibrio cholera* [94] and ArcZ, which regulates RpoS in both *Escherichia coli* and *Salmonella enterica serovar* [95, 96]. In *M. tuberculosis*, the sRNAs MTS1338, MTS0997, and MTS2823 all are present at high levels in the chronically infected lung tissue from mice [73].

In bacteria sRNA can either be *trans*-encoded or *cis*-encoded. *Trans*-encoded sRNA are located between open reading frames and bind with imperfect base pairing with their targets. The majority of the regulation by the known *trans*-encoded sRNAs is negative (Reviewed by Gottesman [97]). *Cis*-encoded sRNA are transcribed antisense to their target RNA, but can also target other RNA in a *trans* manner through imperfect base-pairing. In some organisms *trans* acting sRNA may require an RNA chaperone Hfq [98, 99] but this appears not to be the case in MTB which lacks Hfq [100]. These sRNAs alter gene expression by binding to their target RNA and either changing the stability of the RNA or affecting the binding of the ribosome by blocking binding or by altering the RNA structure to increase accessibility to the ribosome binding site in a manner similar to the previously described ncRNA. Additionally, in some bacteria, RNase III endoribonuclease has been shown to digest the dual stranded RNA that forms between sRNA and their targets, acting as a post-transcriptional mechanism to adjust mRNA levels [101, 102].

Methods for the identification of sRNA

In the past large scale studies of small RNA species were difficult and time consuming, but the development of RNA-seq platforms and *in silico* prediction has opened up this avenue of research allowing for the identification of sRNA in a variety of species. The identification of sRNA poses a series of challenges that make it more difficult than normal gene prediction. The sRNA lack sequence motifs (such a codons or ribosomal binding sites), are generally short in length, are sometimes only conserved in closely related species, and may only be expressed in certain strains [103].

The initial *in silico* identification of sRNA in a genome is done by either comparative genomics based methods [104, 105], machine learning based methods [106], RNA-seq [73], or by base composition analysis [107]. Comparative genomics based methods work under the

assumption that there is both sequence and structural conservation of sRNA in closely related genomes. As a result, these methods cannot be used to identify sRNA that may be unique to a single isolate or where few sequences are available. An example of a tool using this method is QRNA, an early structural noncoding RNA gene finder [104]. Machine learning based methods make use of a training set comprised of known sRNAs as positive samples and the rest of the genome as a negative samples to generate a model. Features are extracted that can be used to describe the samples, which are then used by neural networks, genetic algorithms, or support vector machines to generate the model. An example of which is PSoL, a machine learning method which has the interesting feature of not requiring a negative training set, only a positive one [106]. The detection of sRNA by analysis of base-composition follows the hypothesis that regions in which sRNA are found have statistically detectable differences in sequence composition such as an elevated GC content in structural RNA [108, 109]. This method was mostly used in low GC genomes, and would most likely be less effective in MTB. An alternate approach is to use the presence of known motifs associated with sRNA. An example is the tool sRNAPredict, which detects sRNA by identifying predicted Rho-independent terminators downstream from conserved intergenic regions [110]. This requires *a priori* knowledge of motifs that may not be conserved across different species. Currently identification of novel sRNAs by RNA-seq involves visual inspection of the aligned reads using a genome browser to identify regions that may represent a novel sRNA [73]. This process is time consuming and given the potential hundreds of novel sRNA in a bacterial genome, impractical.

Once a set of candidate sRNAs have been identified, validation may be done using radioactive labeling, RT-PCR, microarrays, northern blotting, co-purification with proteins, and RNA-seq. These methods vary by cost and depth of information provided. Cloning-based approaches have been used successfully, but have some limitations. The process is time consuming, and low-abundance sRNA or sRNAs that are not expressed under the culture conditions may not be detected [111, 103]. Northern blotting is a quantitative and relatively inexpensive method for identifying or validating sRNA that is widely used, but like cloning-based approaches it too can be time consuming and does not provide as much information as RNA-seq. RNA-seq, while being expensive, provides much richer information

including antisense expression levels of genes, sRNA expression levels, and identification of novel ncRNA. In addition to sRNA, RNA-seq has revealed a large number of transcripts generated from the reverse complimentary strand of ORFs [112]. The use of both experimental and bioinformatic methods in combination allows validation of sRNAs which can then be used as gold standards to improve *in silico* methods further.

Determining expression levels of sRNA

Various methods for quantification of sRNA levels are available. These include northern blotting, quantitative real-time PCR, microarrays, and high-throughput sequencing. The advantage of high-throughput methods is that unlike the aforementioned techniques, they do not require *a priori* knowledge about the sRNA sequences. A challenge to determining sRNA abundance is that the stability of sRNA in bacteria is variable, and half lives of transcripts are reported to range from less than 2 minutes to longer than 32 [113]. This also has implications for selecting a purification strategy, as more processing leads to greater deviation from the original sample composition and potentially the loss of low abundance transcripts. When sequencing the sRNA non-coding RNA of *M. tuberculosis*, Arnvig *et. al.,* [73] found that computational removal of rRNA after sequencing produced better results than physical removal, as it limits potential RNA degradation that may skew the abundance of certain sRNA. The use of high-throughput sequencing also allowed the authors to explore a wider range of non-coding RNA including the presence of anti-sense RNA. In *M. tuberculosis* the authors noted high levels of antisense RNA with 28% of the reads representing the total transcriptome mapping in antisense orientation and to IGRs (Excluding ribosomal RNAs) and 168 genes having a greater than 2:1 antisense to sense ratio. For some of the ORFs reads were evenly distributed indicating that they were most likely non-specific antisense background, but for the majority of ORFs showing significant levels of antisense transcripts the reads mapped to the 3' UTR of a nearby gene in the opposite orientation [73].

Known sRNAs found in the MTBC

The first experimental evidence of sRNA in *M. tuberculosis* was produced by Arnvig *et. al.,* [93] in which the sRNAs were identified and their expression between different growth phases

and environmental stresses characterised. Three of the trans-encoded sRNA (B11, G2, and F6) were artificially over-expressed leading to severe phenotypic effects, and in the case of B11 resulted in a complete lack of colony growth. While the regulatory targets of the *cis*-encoded sRNA were discussed, the targets of *trans*-encoded sRNA were not investigated leaving the significance of their expression unclear. A followup study by Arnvig *et. al.,* [73] used RNA-seq to go deeper into the sRNA repertoire of *M. tuberculosis*. Novel intergenic sRNA were identified in MTB that were unlinked to CDSs and did not fall within identifiable ORFs. These were given a designation of "MTS" and numbered according to the nomenclature used in the TIGR annotation of intergenic regions.

One of these sRNAs, MTS2823, which is most abundant during the stationary phase, was selected for further analysis [73]. Overexpression of the sRNA during the exponential phase led to a decrease in expression of a large number of genes, in particular those involved in the methyl citrate network. Unfortunately because the targets of MTS2823 are unknown, it is unclear whether the genes are differentially expressed as a direct result of sRNA binding, or as a consequence of the sRNA affecting mRNAs that code for regulators of these genes. The authors noted that MTS2823 demonstrated functional homology to 6S RNA, raising the possibility that it may not be a true sRNA.

The presence of sRNAs in strains varies greatly, with some being highly conserved between different strains and others appearing to be unique to certain lineages. Di Chiara *et. al.,* [111] identified 34 novel small RNAs (sRNAs) in *M. bovis* BCG, 15 of which were also found in *M. smegmatis* and 12 that were also found to be conserved in a wide range of mycobacterial species. The sRNA found in both non-pathogenic strains like *M. smegmatis* and pathogenic strains are most likely to regulate conserved cellular functions, while the sRNA found only in the pathogenic strains may be related to virulence. Of particular interest was the sRNA Mcr11 found between two genes involved in cAMP metabolism that was shown to be under different regulatory control between MTB and BCG. In BCG the expression responds to both growth phase and a hypoxic environment, while in MTB expression is only growth phase dependent. This highlights the diverse functionality of sRNAs in different species, and that even if they are conserved, they may not have the same regulatory profile.

In another study which investigated the response of *M. tuberculosis* to treatment with

isoniazid, differential expression (based on a threshold of a change greater than 2 standard deviations from the mean intensity) of 14 small RNAs was observed [114], further demonstrating the utility of RNA sequencing as a technology that can provide insight into the sRNA response to changing environmental conditions and pressures. While it is clear that these sRNA are responding to stimuli and the regulation of the response, it is still unclear whether they are functioning by the mechanism described and if the resultant downstream changes in gene expression are as a direct result of sRNA binding or a secondary effect. It is likewise possible that some of the identified sRNA are perhaps not true sRNA. One example is mcr7, which though thought to be a sRNA, may encode a protein involved in bacterial response to low pH [112, 115]. In order to get a clear view of the mechanisms by which these small molecules are acting, the targets they are directly interacting with need to be identified.

Methods for predicting the targets of sRNA

In order to fully understand the role that sRNA play in the cell, researchers have had to develop methods to identify the targets with which the sRNAs are interacting. While experimental validation of sRNA targets provides the best evidence for interactions, it is time consuming and has varying degrees of accuracy [116]. One method is by "fishing", whereby if the sRNA interacts with a protein like Hfq, the Hfq can be His-tagged and the Hfq, sRNA, mRNA complex purified. This was used to identify an ATP-binding cassette (ABC) permease as the target of the RydC sRNA in *E. coli* [117]. Other methods include ribosome profiling, reporter gene assays, binding site mutagenesis, microarrays, RNA probing, gel-shift, and proteomics [103].

In silico methods for identifying sRNA targets have a higher throughput, but a lower accuracy. Pain *et. al.,* [118] classified the sRNA target prediction methods into three categories. There are those that employ an alignment like method that searches the genome for reverse compliments of a query RNA sequence, inter-RNA tools that assess the interactions between sRNA and mRNA using a nearest neighbor thermodynamic model, and independent fold approaches that consider both the binding energy of the sRNA to the target RNA, but also the energy required to get the sequences in a single stranded conformation in which the

interaction can take place. The ability of the sequences to interact is rated as a sum of these two requirements. Additionally, tools that make use of combinations of these methods also exist. A review of the tools identified CopraRNA [119] as the best preforming tool, but the requirement of sequence conservation data for the sRNA limits its use to cases where this is available. In the absence of sequence conservation data, IntaRNA [119], RNAplex [120], and RNAup [121] were the best performing tools in that order. The authors pointed out that one of the main challenges to validating sRNA target detection methods was the lack of gold standard test sets. Even in *E. coli*, the bacteria in which most sRNA studies have been conducted, the majority of recorded sRNA targets are unconfirmed. An additional potential confounding factor is that these models exclude the possibility that an unidentified chaperone protein is involved in the process, as is the case with Hfq in some organisms. In order to fully understand how sRNA affect the phenotype of MTB, we need to integrate information on which sRNAs are being produced, what the targets of those sRNAs are, and what effects they are having on those targets.

1.1.4 Methods for the integration of heterogeneous data: Systems biology

Traits such as pathogenicity are the result of complex interactions between multiple systems within an organism. Whole genome sequencing, RNA-seq, sRNA identification and target prediction provide us with data on different levels of these interactions, but in order to understand how a variant at a genome level affects the linked components in the system, we need an integrative approach to analysing data. Functional genomics approaches have been successfully employed in the elucidation of protein functional relationships from a combination of sources, generating molecular functional annotations for 3,698 of the 4,195 proteins identified in MTB isolate CDC1551 [122].

This starts with the integration of heterogenous data into a relational structure that can be mined for cause and effect interactions. One such data structure is a network. Networks provide a mathematical representation of interactions, comprised of nodes and edges. Nodes represent units in the network, which can be genes, proteins, or variants,

while edges represent interactions between these units and can have attributes like weights that represent the magnitude of that interaction. Edges can also directed or undirected, representing a flow of information. In this way two nodes connected by an undirected edge represent two proteins that physically interact while directed edges represent interactions like those between a TF and the target gene where the TF node influences the gene expression, but the gene does not influence the TF expression (Unless the network contains a feedback mechanism). The data used to create these networks can be obtained from experiments, accessed from databases including STRING [123] and KEGG [2, 124, 125], and even from the literature using text mining methods that can extract functional relationships [126].

Generated networks such as these have been used to identify drug targets, recognised as highly connected hubs within the network [127]. An example being the analysis of regulatory networks of MTB during hypoxia which identified changes in lipid content and metabolic pathways, where the Rv0081 transcription factor was identified as a central regulatory hub in this response [128]. These networks also allow researchers to differentiate between primary and secondary effects in a network, which are often obscured when considering only a single layer such as gene differential expression. This way of structuring data has allowed for the development of new tools including principal network analysis (PNA) that can identify subnetworks based on gene expression data over multiple conditions [129] and new approaches for the modeling of transcriptional regulation (Reviewed by Smolen *et. al.* [130]).

Many tools exist for generating, manipulating, and analysing networks including the Python package NetworkX [131] and the software package Cytoscape [132]. Both provide a platform that is able to import data into a network which can then be visualized, used to generate models, or identify molecular and genetic interactions. Cytoscape differs in that it provides this functionality in the form of apps that can be downloaded from the included app store or created by the user providing a polished workflow for working with the most common datasets, where the use of NetworkX requires the creation of custom functions for importing data, integration and analysis. The disadvantage of Cytoscape is that the development of apps requires knowledge of the Java programming language. NetworkX provides greater flexibility and the ability to rapidly develop new analysis methods using the large number of scientific modules available in python. But the inclusion of an API for Cytoscape and shared

file formats allows for both to be used in unison, with custom analysis done by NetworkX and tasks like visualisation done in the more visually inclined environment of Cytoscape. Many other tools are available including the R Bioconductor package BioNet [133] providing a fertile environment for the development of novel methods.

Graph genomes: The reference genome 2.0

Another use for graphs has been found in the representation of genomes. The current standard format for genome storage is as linear sequences stored in a fasta file or variant call format (VCF). This sequence is often a consensus from a set of sequences that collectively represent anything from an individual isolate such as *Mycobacterium tuberculosis* H37Rv, to an entire species, in the case of the human genome assembly hg19. Although they have served their purpose up until now, in the age of pan-genomes and microbiome studies these representations have become limiting in terms of the file space they occupy, the functionality they provide, and their ability to represent population scale variation. Projects that require the sequencing of entire populations like the human microbiome project [134] and the FIND Tuberculosis Strain Bank [135] generate large amounts of mostly repetitive data and have the potential to benefit from compression methods that remove redundancy.

In light of these challenges, a shift from linear to graph based representation of sequences began. Graphs are already used by genome assemblers such as ABySS [136] and Velvet [137], both of which make use of De Bruijn graphs, and are now making their way into downstream analysis as a new way of representing multiple individual sequences in a single data structure. The representation of multiple sequences in a compressed de Bruijn graph was proposed by Marcus *et. al.,* [138], and was later improved and implemented by Beller & Ohlebusch [139]. Methods like these allow for the creation of pan-genomes; a single structure that represents all variation between individuals in a defined clade, which offer a myriad of advantages over the use of a single reference genome.

A few methods exist for creating these graph genomes [140, 138], the commonality between many of them is that stretches of sequences are represented as nodes in a graph, with the edges forming a path through the graph, though the methods used to create these graphs, the type of metadata they contain, and the downstream functionality varies. Tools

like vg use a genome and variant file to generate a variant graph [141] against which reads may be mapped. More recent tools such as PanTools [142] provide the means to create a pan-genome from multiple annotated genomes which can then be compared. A significant feature of some of these tools including PanTools is the ability to account for large non-collinear rearrangements. Most of these existing tools are open-source though some, including the Seven Bridges genome graph toolkit (https://www.sbgenomics.com/graph/), are proprietary.

The representation of a species as a single genome graph has the potential to alleviate many of the challenges that arise from using a single canonical reference (H37Rv in the case of *M. tuberculosis*) and has been gaining traction in recent times (Reviewed by Novak *et. al.* [143]). As MTB contains various large structural changes in the genome [50], and differs in gene content from species to species [144], the use of a single reference genome graph would improve read mapping and prevent important genes that may not be found in H37Rv from being excluded from analysis.

1.1.5 Taking the lead in the arms race

While we are continually discovering novel treatments and refining detection methods and models, the battle against TB is far from over. Even with the success of the WHO DOTS programs raising cure rates above 80% and reducing dropout rates by nearly 10% [145], TB maintains strongholds in various generally poorer countries, including South Africa. These strongholds are of importance to countries where TB has been eradicated as they are the incubators from which highly resistant strains are emerging, and may soon begin to return to regions where it had been previously brought under control. As a result of this emergent drug resistance, medicine is in an arms race with pathogens where we are mostly acting in response to changes in the pathogens. In order to get ahead of their evolution we need a more integrated understanding of how changes at a genome level translate to phenotypic changes, a better understanding of the forces that drive the evolution of the genome, and what changes occurred over time on the road to becoming a modern resistant pathogen.

This begins with the integration of data from different levels including interactions within the organism, between the organism and the host immune system [19, 21], and with co-

infections like HIV [14]. Better understanding of host pathogen interaction by use of new techniques like 3-dimensional culture [146] and the use of bioinformatic and NGS techniques to deconvolute ncRNA regulatory networks will also be needed to create a more complete picture. Databases are being created that store strains which are screened for drug resistance with a broad geographic and lineage diversity [135], which may provide larger datasets that allow us to use machine learning and other methods that require training sets and benefit from the use of genome graphs. Integrating these highly connected datatypes has the potential to create a flood of data from which we will be unable to differentiate noise from signal. This requires the development of innovative new techniques in fields like functional genomics that will allow us to identify the most relevant interactions in the hive of interconnecting pieces. These advancements will not only allow us to better treat TB, but the generalisable lessons learned will allow us to better respond to any unforeseen emergent diseases that will inevitably arise over the course of human existence.

1.1.6 Project aims and motivation

The availability of the two *Mycobacterium tuberculosis* isolates S507 and S5527 has provided the opportunity to study how changes at the genomic level effect virulence, and how those changes disseminate through the biological network. With this in mind, the aim of this project is to identify the cause of the altered virulence displayed between the isolates at a genomic and transcriptomic level, and identify the potential biological mechanisms through which their divergence at a genomic level produced the phenotypes. This will require a systems level approach, including a better understanding of sRNA mediated regulation, a relatively novel post-transriptional regulatory layer, particularly for MTB, as well as the development of methods and systems to integrate the newly generated data with existing knowledge. This in turn requires the development and testing of tools and pipelines that aid in the interpretation of the data, and take advantage of new paradigms including the use of graph based representations of genomes.

This research into MTB sRNA we will further our understanding of the regulatory systems in MTB, and aid future studies where these sRNA may be key to explaining regulatory perturbations. This includes describing the altered regulatory systems of isolates S507 and

S5527, where by identifying the pathways and genes involved in the virulence phenotype could potentially high-light novel drug targets, identify processes involved in determining virulence, and improve our understanding the evolutionary drivers related to virulence in MTB.

"The pen is mightier than the sword if the sword is very short, and the pen is very sharp."

Terry Pratchett

2

GenGraph toolkit: for the simple generation and manipulation of genome graphs

2.1 Introduction

The current standard format for genome storage is as linear sequence stored in a fasta file or variant call format (VCF). This sequence is often a consensus from a set of sequences that collectively represent anything from an individual isolate such as *Mycobacterium tuberculosis* H37Rv, to an entire species, in the case of the human genome assembly hg19. There is also a lack of uniformity in the field of tuberculosis research, where different studies use different reference genomes or different versions of annotations making comparison between studies difficult. Even in the reference genome, H37Rv, the number of genes varies between sources with 4173 genes reported on Mycobrowser and only 4,008 (3,906 protein coding) in the

NCBI's reference genome NC_000962.3. Although they have served their purpose up until now, in the age of pan-genomes and microbiome studies these representations have become limiting in terms of the file space they occupy, the functionality they provide, and their ability to represent population scale variation.

An additional challenge faced in the modern setting of high-throughput sequencing and large -omics projects is one of data storage. This is particularly the case in microbiome sequencing projects where entire populations of organisms are sequenced [134], or projects such as the various human genome sequencing projects around the globe. In these projects large amounts of mostly redundant data is generated and stored, and where there is redundancy there is great potential for compression.

In light of these challenges, a shift from linear to graph based representation of sequences began. Graphs are already used by genome assemblers such as ABySS [136] and Velvet [137], both of which make used of De Bruijn graphs, and are now making their way into downstream analysis as a new way of representing multiple individual sequences in a single data structure.

The representation of multiple sequences in a compressed de Bruijn graph was proposed by Marcus et. al [138], and was later improved and implemented by Beller & Ohlebusch [139]. Methods like these allow for the creation of pan-genomes; a single structure that represents all variation between individuals in a defined clade, which offer a myriad of advantages over the use of a single reference genome.

A few methods exist for creating these graph genomes [140, 138], the commonality between many of them is that stretches of sequences are represented as nodes in a graph, with the edges forming a path through the graph The methods used to create these graphs, the type of metadata they contain, and the downstream functionality varies. Tools like vg (variant graph) use a genome and variant file to generate a variant graph [141] against which reads may be mapped. More recent tools such as PanTools [142] provide the means to create a pan-genome from multiple annotated genomes which can then be compared. A significant feature of some of these tools including PanTools is the ability to account for large non-collinear rearrangements. Most of these existing tools are open-source though some including the Seven Bridges genome graph toolkit (https://www.sbgenomics.com/graph/) are

proprietary.

While there has been great progress in the early days of graph based representations of sequences, there remains ample room for further growth and development for applications. There is a need for a generalisable approach and set of tools that can allow the development of new applications, and a platform to test and refine the graph based sequence structure. To this end the GenGraph toolkit was created. In this chapter we outline the structure of the genome graph, the methods and tools used in its creation, and show an example of a down stream application. We present some metrics relating to the reduction in file size achieved as well as hi-lighting some of the areas of future development. All of the tools and methods are presented with a transparent, modular, and open-source ethos to facilitate development and adoption of graph based methods in a practical and easily implementable manner.

2.2 Materials and methods

2.2.1 Modular function

GenGraph is written in Python, and is freely available as OpenSource software. Detailed descriptions of the methods, functions, and conventions are included in the supplementary data as well as in the GitHub repository. Only common existing filetypes have been used in order to facilitate adoption and prevent fractionation, with the only key dependency required being the Python NetworkX [131] package.

In order to promote community development of the GenGraph toolkit, the code is highly modular. The process of creating a graph genome has been broken down into multiple core functions, each with a defined input and output that allows them to be developed individually without jeopardising the stability of the entire workflow. Unit tests have been implemented to maintain the integrity of the codebase. Where many of the current tools use a specific method for generating the graphs, GenGraph can use any current or future alignment tool that produces a parsable standard output. This allows the toolkit to evolve and improve with time, as well as utilise alignment tools that are best suited to the dataset at hand and the latest advancements in alignment including GPU acceleration.

2.2.2 Structure of the graph

GenGraph creates sequence graphs that are directed multi-graphs. Nodes in the graph represent blocks of aligned identical sequences, and have attributes including a unique node identifier, a list of isolates represented within the node, the nucleotide sequence of the node, and the relative start and stop positions of the sequence in the node for each isolate. Edges in the graph are directed, representing the concatenation points between the end point of the sequence in a predecessor node to the start point of the sequence in a successor node, and are labeled with the set of isolates whose path through the graph they represent. One of the challenges of transitioning from a linear sequence to a graph structure is determining the relative position of bases in the graph genome. While this is simple when working with a string, once the sequences are aligned and converted to a graph structure determining the location of a feature such as a gene based on a set of coordinates can be challenging. As such in the context of the alignment a nucleotide may be position 2,321 for isolate 1, position 442 for isolate 2 etc. Storing the relative start and stop positions of the sequence for each isolate as node attributes allows for unambiguous coordinate mapping, and the use of existing annotations from the original linear sequences. The nodes can contain a reverse-compliment of a sequence, represented by negative start and stop values.

2.2.3 Alignment

In GenGraph, there are two parts to the creation of a genome graph from a set of linear sequences. These steps are global alignment and a local realignment (Figure 2.1).

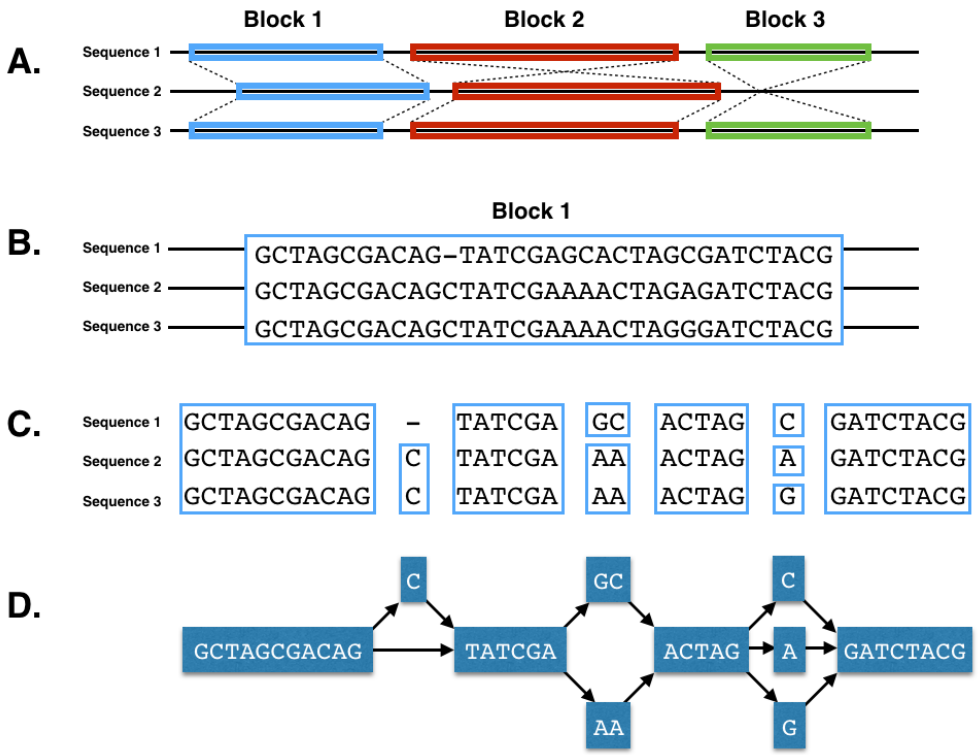


Figure 2.1: An overview of the GenGraph algorithm. (A) Initial global non-linear multiple sequence alignment. (B) Secondary local multiple linear sequence alignment. (C-D) Identical blocks in the alignment are represented as single non-overlapping nodes in the graph.

Global alignment

To account for large-scale potentially non-colinear evolutionary events such as chromosomal rearrangements and inversions, an initial global sequence alignment is conducted. Currently GenGraph uses progressiveMauve [147] for this step, though the use of other alignment tools is facilitated. The backbone file produced by progressiveMauve represents large regions of sequence homology. It is this backbone file that is used in the creation of the initial graph structure, the format of which is described in the progressiveMauve documentation.

Local alignment

The resultant structure graph then undergoes a local realignment step, where the sequences of the original genomes represented by each node are extracted from the original input

sequence files to undergo local multiple sequence alignment. The aligned fasta file produced is then converted to a sub-graph, that then replaces the original node the sequences were derived from. Any local multiple sequence alignment tool that produces a standard fasta alignment file may be used. Currently, Muscle [148], Mafft [149] and Clustal Omega [150] are supported.

The final graph objects created by GenGraph may be exported as GraphML, XML, or as a serialized object, though various other formats may be added in future. Using existing graph formats allows exported sequence graphs to be visualized in commonly available programs such as Cytoscape [151]. GenGraph creates a report file containing information such as the number of nodes and edges in the graph, the average in and out degree of the nodes, the total sequence length of all the nodes in the graph and the density of the graph. This information can be used to monitor how graphs change as more genomes are added as well as the relationship between the number of features and the graph size.

2.2.4 Toolkit

Once the graph is created, multiple tools and functions exist that allow direct analysis using the graph data structure. These tools and functions may be combined or improved to allow for more complex operations to be carried out.

Use case: MTB pan-transcriptome

As an example of how tools within the GenGraph toolkit may be used together to perform more complex functions, a composite function was created that allows a rudimentary pan-transcriptome to be extracted from the pan-genome graph. The function uses the GTF files associated with the input genomes to identify the position of features in the graph. The similarity of overlapping features is calculated and a homology matrix is generated (Figure 2.2). Genes that were above a 95% level of similarity were deemed homologous, their sequence extracted, and used to create a file in fasta format that may be used by current alignment tools for read mapping. As only a single version of the gene region may be represented in the fasta file, when exporting the pan-transcriptome the sequence of a reference isolate

was used. Genes unique to a single or set of isolates were also included, making this an exhaustive set of genes found within the provided MTB genomes. The core genome of the set of genomes may likewise be extracted. Six MTB isolates (CCDC5180, CDC1551, F11, H37Ra, H37Rv, and W148) were used in the creation of the graph from which the pan-transcriptome was created. Using bwa [152], RNA reads from NCBI's Sequence Read Archive (SRA), https://www.ncbi.nlm.nih.gov/sra/SRX798220[accn] from a CDC1551 isolate were aligned to the pan-transcriptome, as well as to the virtual transcriptomes of the individual isolates. The individual transcriptomes were created by extracting only the coding regions from the genome using the associated annotation file for each of the isolates (CCDC5180, CDC1551, F11, H37Ra, H37Rv, and W148) and creating a fasta file against which the CDC1551 reads could be aligned. This removed the possibility that the other isolates have better mapping due to reads aligning to intergenic regions.

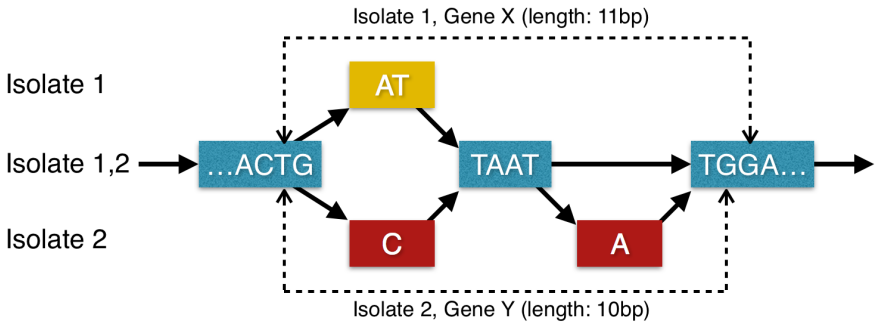


Figure 2.2: Overview of the similarity assessment. Given a graph created from two isolates and their respective annotation files, two genes X and Y may be compared to one another and their similarity quantified. Currently the similarity of gene X to gene Y is calculated by the cumulative length of the shared nodes (blue, 7bp) divided by the length of the query gene X (11bp) giving a score of 63%

Use case: Cladogram construction

As the genome graph represents an aligned form of the sequences where each node can be seen as an evolutionarily conserved block, the structure of the graph can be used to calculate pairwise similarity between sequences. In normal phylogenetic analysis a table is created containing the pairwise differences between the aligned sequences, which is then used to create a phylogenetic tree. Here, we calculate the distance between two sequences as a

ratio of nodes shared between the sequences to the total number of nodes for the sequence. This means that for two sequences that are identical, they would share 100% of their nodes. If the sequences share 95 nodes, and each have 100 nodes in total, they share 95% of their nodes and are 95% similar. The generated similarity matrix represents the distance between two sequences as a ratio of nodes shared between the sequences to the total number of nodes for the sequence. This provides an estimate of the phylogenetic distance of the sequences to one another and can be used to generate the cladogram using SciPy's dendrogram function.

Use case: Ancestral genome

To demonstrate how weighted path algorithms may be employed to detangle the evolutionary history of sequences, a tool for the reconstruction of a rudimentary ancestral genome was created using a genome graph and the functions provided by GenGraph representing the last common ancestor (LCA) of the isolates in question. The algorithm traverses the graph along the path that is taken by the majority of the sequences. This means that while traversing the graph from node A, given a choice of two nodes (B and C) each linked by an outgoing edge from A, where node B contains the sequence at that position of 5 isolates, and node C only 3, it will traverse to node B. Once the graph has been traversed in this manner, the sequences found in the nodes along the path are concatenated into one single sequence representing the potential ancestral genome. The nodes are weighted based on the similarity matrix generated above so that nodes containing closely related sequences contribute a lower weight to the path, alleviating the effect of oversampling of certain isolates. This means that in the previous example, if node B had contained 5 isolates that are all 99.99% similar, it will not be used over node C containing 3 more distantly related isolates.

2.3 Results and discussion

Here we present some of the properties of the graphs generated using the methods implemented in the GenGraph toolkit. The genomes of various isolates of *Mycobacterium tuberculosis* downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/genome/) were used in the graph generation (Table 2.2).

With multiple sequence alignment being the current bottleneck in the creation of genome graphs, the scalability of GenGraph is dependent on the ability of the latest alignment tools. The time taken to generate the graphs increases in a linear fashion, influenced by the number of sequences being aligned, their length, and their similarity. By breaking down the genomes into partially pre-aligned blocks, GenGraph is able to align multiple long genomes in segments and with the current version of mafft able to align up to 30,000 sequences can be aligned in a block, though this has not been tested.

2.3.1 On the graph data structure

Performance wise, GenGraph was able to create a genome graph containing 5 MTB genomes on a 2012 i7 Macbook Pro with 8 GB ram using Mafft in 53 minutes and 10 genomes in 2 hours and 44 minutes. For smaller genomes, 300 HIV-1 genomes were aligned and converted to a genome graph in 35 minutes. From testing we see the scalability of GenGraph is dependent on the ability of the latest alignment tools, and is dependent on the number of sequences being aligned, their length, and their similarity though in general we observe a linear increase in genome graph generation time as the number of sequences increase. Graph generation represents the most computationally intense and time consuming process, while downstream analysis benefits from the use of the data in an aligned form. Various features in the graph including SNPs and large inversions were represented. The graph was exported to GraphML format, which was then imported into Cytoscape for quick visualisation of genomic features such as large deletions, SNPs, or conserved regions. The graph genomes generated are able to incorporate the input sequences into a single graph, including the MTB W-148 isolate that contains large scale chromosomal rearrangements [50]. This ability to accommodate large structural events is a distinguishing feature between GenGraph and vg.

2.3.2 Data compression using the graph structure

One of the tantalising promises of graph genomes is of smaller file sizes as a result of the reduction in redundancy. The major caveat to this, is that the metadata required to represent polymorphisms also contributes to the eventual file size. Despite GenGraph's currently early

level of optimisation, the exported graphML file was significantly smaller when compared to the individual constituent genome fasta files (Table 2.1). As each feature in the graph has an effect on the size and complexity of the graph (in terms of the number of nodes and edges), a decrease in the realised compression is observed as sequences diverge. This eventually leads to a divergence threshold, above which the meta information required to store the variants of the additional sequences negates the size reduction gained by minimising sequence redundancy (Figure 2.3). This threshold differs depending on factors including the export file type and nature of the meta information. Additionally, when quantifying the compression ability of genome graphs an important consideration is that the graph structure represents not just a sequence file but also an alignment file and potentially a variant file.

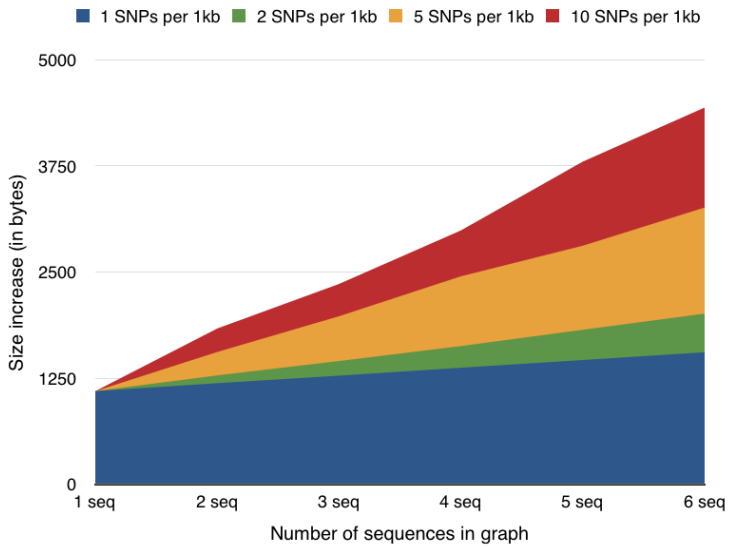


Figure 2.3: Increase in file size of an exported graph in GraphML format excluding the contribution by the GraphML structure. As additional sequences are added to the graph, the file size increases by a factor related to the similarity of the sequences and the data required to store the differences. Only the resultant increase in file size by the addition of sequence to the graph structure is presented.

Table 2.1: Increase in file size per genome added to the graph.

Number of genomes	1	2	3	4	5	6
File size	4,5Mb	5,9Mb	7,6Mb	8,5Mb	11Mb	13Mb
Number of nodes	0	3,690	8,106	9,320	13,264	15,355
Number of edges	0	4,886	10,868	12,485	17,823	22,296

2.3.3 Toolkit

Use case: MTB pan-transcriptome

Due to the variance in genome size and composition between strains, as well as the limited availability of well annotated reference genomes, researchers often have to choose between aligning reads to the closest related genome or to the available reference genome for the organism. While aligning reads to the closest related genome maximises mapping, it lacks detailed annotation that would allow interpretation of function, for the use of a reference genome, the converse is true. If the closest isolate to the sequenced organism is not known, it can be equally difficult to select a reference. A solution to this problem is the use of a pan-transcriptome that increases the overall mapping without sacrificing on functional data from annotations.

To demonstrate a practical application of a graph genome using some of the tools available in the GenGraph toolkit, a pan-transcriptome was created and used as a reference to align RNA sequencing reads from a MTB CDC1551 transcriptomics dataset obtained from the SRA. When compared to local sequence alignment based homology searches, the method employed by GenGraph is able to identify homologues with a greater accuracy, especially for highly similar genes such as the PE/PPE gene family. This is a result of the tools ability to consider overlapping features in a global alignment, where not only their similarity is considered but also their position in the genome.

The resultant pan-transcriptome contained 3,910 genes that were present in all isolates and 319 accessory genes that were either missing, truncated, or showing less than 95% sequence similarity. When compared to alignment to the individual transcriptomes, an increase in mapping between 3.6% (717,787 reads) and 8.79% (1,720,734 reads) percent was

observed against CDC1551 and H37Rv respectively (Table 2.2). Despite multiple genome rearrangements in some isolates of *M. tuberculosis* [56, 50], the relatively recent emergence of the pathogen [46] has resulted in genomes that are fairly similar at the gene composition level. Though the variance in the number of genes between isolates is moderate, with F11 having only 3,959 genes and CDC1551 as many as 4,189, even small differences in MTB genomes are shown to have a large impact on the pathogenicity and response to treatment. In the case of MTB, the use of H37Rv as the reference therefore, means that when sequencing strains from other genotypes a significant number of genes are not included in the study. This rudimentary example demonstrates a subset of GenGraph's aims realised, in that the use of a single linear reference genome can result in the loss of what could potentially be the most informative data, where the use of a genome graph is an exhaustive representation of the organisms in question.

Table 2.2: Comparison of mapping statistics for CDC1551 reads mapped to the pan-transriptome and a subset of the genomes from which the pan-transriptome was created including CDC1551.

Genome	Graph genome	CDC1551	F11	H37Rv
Accession number	N/A	NC_002755.2	NC_009565.1	NC_000962.3
length (in bp)	4,487,683	4,403,838	4,424,435	4,411,533
File size	7,4Mb	4.3M	4.3M	4.3M
Transcriptome				
Number of genes	4,229	4,189	3,959	3,999
Aligned reads	13,576,748	12,858,961	11,890,328	11,856,014
Percentage aligned reads	69.4%	65.7%	60.8%	60.6%

Use case: Cladogram construction

The generated cladogram accurately represents the relationship between the isolates as described in previous phylogenetic studies (Figure 2.4). Though this method does not include methods such as bootstrapping and position weight matrices, it demonstrates the ability of the toolkit to take advantage of the aligned data structure. An advantage of the graph structure was that the large chromosomal inversion present in isolate W-148 does not pose

a problem for this method and is treated as a single evolutionary event. Large deletions and chromosomal inversions could be problematic for alignment based methods, depending on the location of the deletion and the methods used.

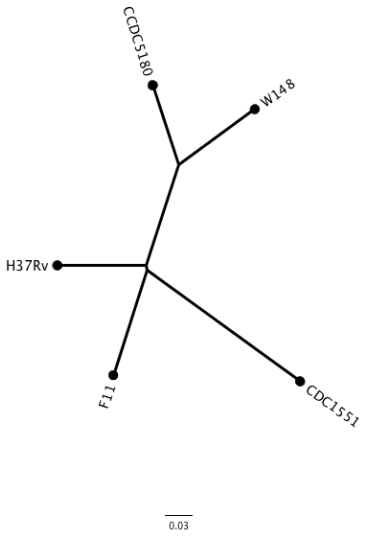


Figure 2.4: Cladogram of select MTB species. Isolates from the Beijing lineage (CCDC5180 and W-148) are shown clustered together with the other lineage 4 strains likewise clustered.

Use case: Ancestral genome

An ancestral genome allows us to better understand the evolutionary history of the sequences and place the variants we observe in context. The resultant genome is 4,3Mb in size and appears to have the correct structure when aligned to the CDC1551 genome, lacking the chromosomal rearrangement seen in isolate W-148. The sequence is exported as a fasta file as well as represented as a path in the original graph genome and provides metrics such as the proportion of isolates that follow the ancestral path. This heuristic approach, though not a fully fledged ancestral genome reconstruction algorithm, is an example of how a graph traversal algorithm can be developed to take advantage of the genome graph structure.

2.3.4 Current and future developments

Additional functionality of the toolkit that is currently under testing includes the ability to extract variants from the graph and export them into VCF files, the ability to add variants to the graph from a VCF file, and the ability to use the graph as a reference for raw read mapping. Additional testing is underway for improving the speed of the tool and the size of the generated graphs.

2.4 Conclusions

The GenGraph toolkit provides a simple method to create and utilise genome graphs, as well as providing a framework for further development by being highly modular and simple to use. It provides tools that allow real-world usability of graph genomes without requiring major alteration of existing pipelines, protocols, or training which are often some of the greatest barriers to the adoption of a new method. The tool is scalable from small viral genomes to bacterial genome on desktop computers, with further testing for large genomes already underway. The ability to create a pan-transcriptome against which reads may be aligned from the generated graphs acts as a demonstration of these characteristics. And though shy of it's potential, the current lossless compression of genomes is significant, and aids in the growing problem of ever increasing datasets. In this initial version of the toolkit many aspects were identified as bottlenecks with potential for improvement and are already being addressed. As GenGraph's design philosophy allows for these changes to be easily implemented, this should see the toolkit evolve rapidly. By making the code available on GitHub, we hope that GenGraph will be adopted, adapted and applied as the community requires.

"'Multiple exclamation marks,' he went on, shaking his head, 'are a sure sign of a diseased mind.'"

Terry Pratchett, Eric

3

The development of the Cell pipeline

3.1 Introduction

In this chapter we outline the Cell pipeline, which was created and used for the management of next generation sequencing data in this project. The use of pipelines increases reproducibility, reduces time spent on repetitive tasks, and can be used in subsequent projects saving time in the future. As the complexity of projects increases with the growing number of datasets and data types, pipelines are required to do more than simply automate processes. The Cell pipeline addresses this growing complexity problem by passing the results to a reporting tool named Holmes. Holmes structures the results into networks that can be traversed by an algorithm, and uses them to create reports that amalgamate data from various sources, high-lighting interesting interactions and anomalies. Both the Cell pipeline and Holmes are written in python, and contain a suite of functions and wrappers for existing

tools that orchestrate the analysis and reporting of NGS data.

3.2 Materials and methods

The cell pipeline is composed of two parts (Figure 3.1). The preprocessing steps take the raw reads and create the output files normally associated with differential expression analysis or variant calling. The post-processing is done by Holmes and involves data amalgamation and reporting.

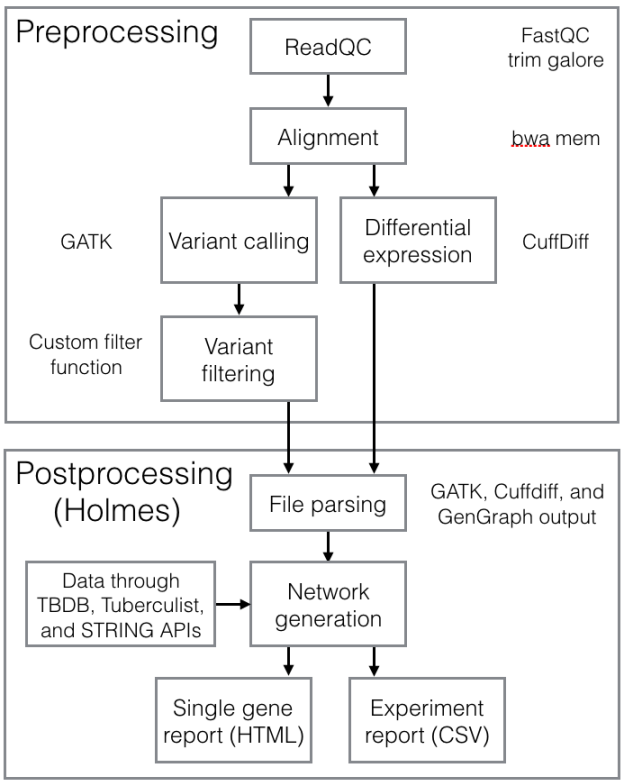


Figure 3.1: The components in the Cell pipeline are modular, and can be swapped for other tools or updated to add new functionality.

3.2.1 Preprocessing

Read quality control: filtering and trimming

Initial quality control (QC) of raw sequencing reads is done using FastQC [153], and the results from this step are summarised in an output table. Individual inspection of each

report should be conducted as there are a variety of problems that can occur at this stage that can not yet be reliably corrected in an automated fashion. The results from FastQC are used to inform parameter selection for read trimming. This is carried out by Trim Galore! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), a wrapper for cutadapt [154] and FastQC, that provides adapter trimming and filtering based on the overall read length and quality, and per-base quality trimming. The FastQC report generated for the filtered and trimmed reads can then be compared to the initial report, and if the quality is acceptable the reads are passed onto the main pipeline.

The paths to the filtered and trimmed reads are placed in a sample sheet that contains information on the isolate, and condition to which they pertain. The parameters for a run are set depending on the type of analysis, and include the reference genome to be used, the annotation file, and an experiment descriptor. Optional parameters include setting a preferred alignment tool (BWA-MEM by default) or the number of threads available for the analysis.

Read alignment

In order to select a default sequence alignment tool we compared two of the most popular available, BWA-MEM (maximal exact matches), and tophat2. As this pipeline would be used primarily for Mycobacteria, a test set of RNA sequencing reads from six samples from the MTB isolates to be used in this project were aligned to the H37Rv genome. Before alignment, the reference fasta files are indexed, and a sequence dictionary is created. Once reads are aligned, the alignments are filtered by SAMtools (http://samtools.sourceforge.net). Unmapped reads are removed, along with reads where the mapping quality falls bellow a quality threshold (default = 30). The read alignments are then converted to bam format, sorted and indexed. The filtered, sorted and indexed bam files are then moved into an alignment directory and the intermediate files removed to recover disk space. A report file of the alignment is then generated using picard tools (http://broadinstitute.github.io/picard/) and summarised by the pipeline in the report file. The pipeline is also able to use pan-transcriptomes generated by GenGraph as a reference, and is able to make use of the gene homology matrix in the report generation phase.

Differential expression analysis

By default, the pipeline will go through the provided annotation file and retrieve any annotation for features that need to be masked. This includes sequences that should have been removed during sequence preparation such as rRNA and tRNA that may have excessive coverage and skew normalisation efforts. These features are placed in a new mask file in gff3 format and is used by Cuffdiff. The sequences placed in this mask file can be set by the user, or a specific mask file may be provided.

The sample sheet can contain multiple conditions, and the user may either select a pairwise comparison of specific conditions, or conduct an all against all comparison. Depending on what is required, a run specific sample sheet is generated, and used by Cuffdiff to conduct the differential expression analysis. This allows for flexibility in the experimental design, with pairwise comparisons generally having less statistical power, but are more robust against one outlier sample skewing the normalisation and vice versus for the all against all experimental design. Quality control of the differential expression analysis results is done in R using the cummeRbund package which conducts clustering analysis and produces visualisations including PCA plots. These results require human interpretation, and cannot be automated during this time.

Variant calling

Variant calling in the Cell pipeline is done by the GATK HaplotypeCaller. The advantage of the HaplotypeCaller is that is capable of local *de-novo* assembly of regions where high sequence variation makes read mapping and haplotype calling difficult, after which it can conduct SNP and indel calling. The resultant variants are then filtered based on depth, mapping quality, and the number of alternate alleles at the locus using a custom tool included in the pipeline. If two isolates are being compared and aligned to a common reference genome, variants that are common to both isolates relative to the reference are filtered out leaving the variants that distinguish the isolates from one another. This variant file is then available to be used in the amalgamation and reporting phase.

3.2.2 Post processing

Generating an information network

Up to this stage, the pipeline used standard practices for variant calling and differential expression analysis. Once the variants have been called, and the gene expression levels determined, the pipeline passes the results to Holmes, that integrates the results with information from additional sources. The generated files are parsed and converted to directed networks using the python NetworkX package. Each gene is represented as a node of type 'gene', containing attributes such as the gene name, the level of expression under the different conditions, and the gene position within the genome. Information on homology is obtained from the GenGraph output. The genes are linked to one another in the order that they would be found in the genome. The network is exported and can be visualised using programs like Cytoscape. Using Cytoscape's styling tools the nodes can be coloured based on the log2 fold change in expression, or change their width as a function of the gene length.

The inclusion of data from sRNA experiments is also supported, and sRNA are included in the network as node type 'sRNA'. By parsing the outputs of *in silico* target prediction software, edges are created from the sRNA nodes to their targets. Variants are included in this network as nodes with the type 'variant', creating edges between variants and the gene nodes that they fall within or in proximity to depending on a distance set by the user. This allows the user to select a gene from the network and identify any variants within the coding sequence or in close proximity that may be causing a change in the expression profile. Protein-protein interaction data from the STRING database is retrieved via the API, parsed and integrated into the network as edges between interacting proteins. Data on the nature of these interactions is stored in the edges in can include information on the type of interaction and the weight. Information on transcription factor binding sites from ChIP-seq experiments made available on the tuberculosis database [43] are used to recreate regulatory networks.

Generating reports using Holmes

Holmes uses network traversal to generate reports for either a particular condition or for a gene. Certain node types are considered "influencers", this includes TFs, sRNA, and

variants. Information in the network is directional, meaning that a transcription factor car regulate a target gene, but the expression of the gene does not effect the TF (unless a feedback mechanism exists). The algorithm identifies influencers that are linked by incoming edges to the query node (Figure 3.2). If these nodes represent transcription factors, the expression value is assessed to see if it is also differentially expressed. The algorithm then repeats the process for the influencer nodes if they are TFs or sRNA, checking if they have any incoming edges from nodes that are differentially expressed or variants. Variant nodes represent endpoints for the traversal, as mutations are currently the most readily available cause for differential expression. Because variants are linked by edges to genes they fall within or are proximal to, variants that effect gene expression by altering the sequence of a TF or by changing a TF binding site of a gene are both identified. The level of recursion is set at 3, but can be increased. For the gene reports, the node representing the gene is located in the network. From this node, the expression values, function, and homology information is obtained. The tool then investigates the neighbourhood around the gene node, including adjacent genes. If neighbouring genes are found differentially expressed the algorithm will repeat this process for the neighbours, and include them in the report. If a variant is found a few genes upstream of an operon it can be detected by this method. The algorithm then extracts this subgraph and uses it to generate a html report. Relevant information is summarised and links to additional information included. For summaries of entire experiments the pipeline first extracts all genes that are found to be differentially expressed and uses them as a starting point for network traversal. Relevant connected nodes are identified using the same algorithm as discussed above. The results are summarised in an exported table in csv format.

3.3 Results and discussion

Pipelines are efficient ways to conduct analyses that are run often in a research environment, reducing errors and aiding in reproducibility. The tools used by the pipeline are interchangeable, making it simple to add new alignment software, variant callers, or functionality.

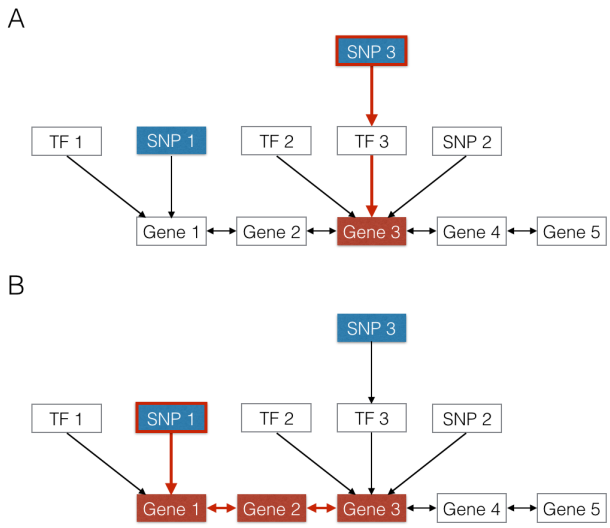


Figure 3.2: The Holmes algorithm is able to identify the neighbouring nodes of the query gene (Gene 3) and extract the nodes that are likely to have an effect on the gene expression. Differentially expressed genes are coloured in red with variants coloured blue. A: The differential expression is as a result of a mutation in the TF that regulates gene 3. In this scenario the mutation altered the TF binding affinity but not the TF expression. B: The gene is part of an operon that includes Genes 1-3. All the genes are differentially expressed as a result of a mutation in the upstream TF binding site.

3.3.1 Preprocessing and quality control

Selecting the best parameters for read QC is one of the phases of the pipeline that should not be automated. While tools can trim reads based on base quality and other known metrics, sometimes sequencing can produce strange results that only a trained individual can detect. After QC is completed, the pipeline carries the analysis through to the report generation phase. In the comparison of BWA-MEM and Tophat2 it was found that BWA-MEM was superior, aligning on average 0.2% more reads (Table S7.2) and was therefore used as the default alignment tool. This should be reconsidered on a per project basis, especially if the organism is a eukaryote or if new tools become available.

Generating reports using Holmes

Holmes provides additional functionality after processing by integrating the results into a network that can be traversed by the Holmes algorithm. As projects grow and become

more complex, it is becoming impossible for researchers to investigate every interaction individually. While the ability of a researcher to interpret the results and apply critical thought will always be required, tools that shape the results into a more palatable form, highlighting the most probable causes for the differences in sets, will become increasingly necessary.

The gene reports provide a summary of information about a gene of interest in one clear html document that can be viewed in any web browser (Supplementary file S7.1, S7.2, S7.3, S7.4, S7.5, S7.6) while the experiment CSV table displays information from multiple sources in a clear and easy to interpret manner (Table 7.6). The exported network GraphML file is viewable with Cytoscape and allows the researcher to further explore the data, applying different styles, or extracting sub-networks that represent interactions of interest (Figure 3.4) or genes that may be effected by a nearby variant (Figure 3.3).

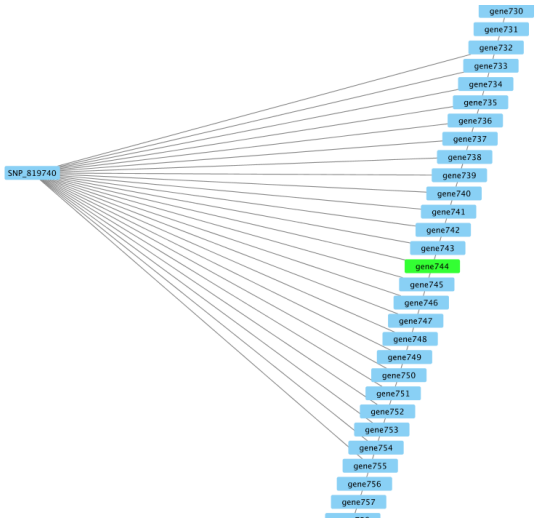
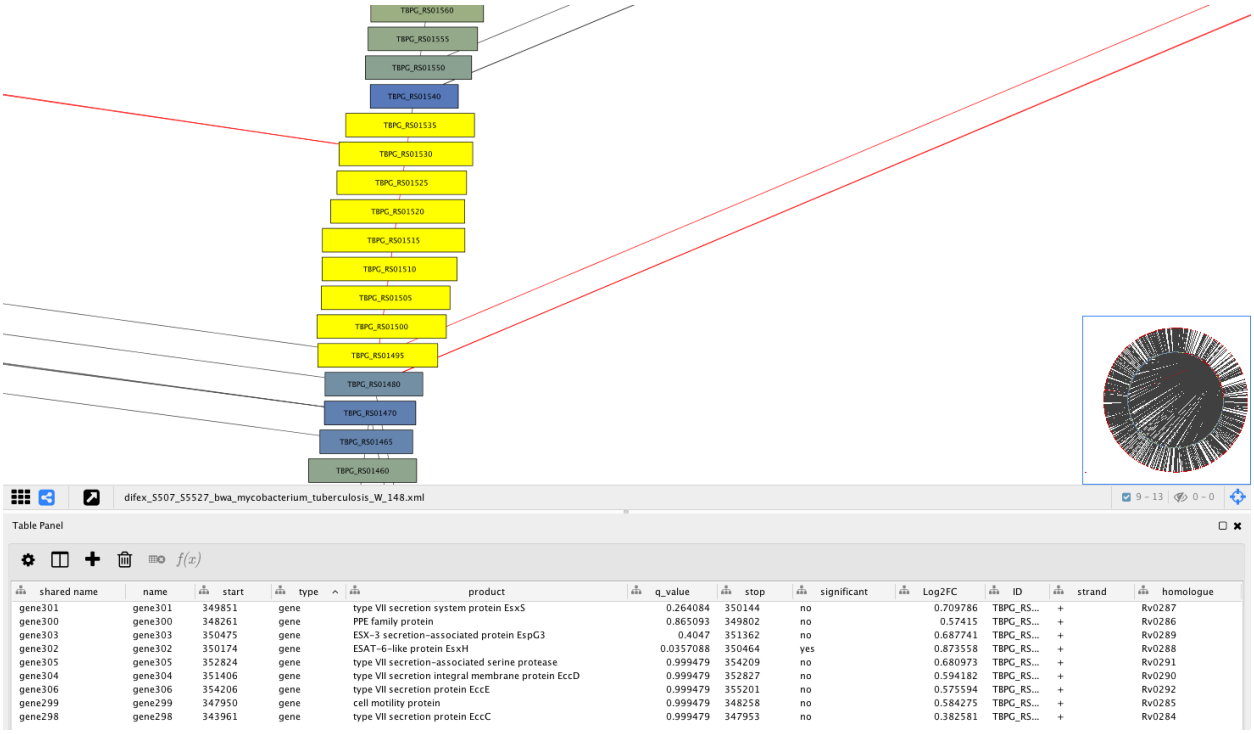


Figure 3.3: Nodes that represent variants can be linked not only to the gene in which they are located, but also to nearby genes. The distance at which a variant node is linked to a gene can be set by the user to identify variants that effect distant TF binding sites or only those that fall within the coding region of a gene.



shared name	name	start	type	product	q_value	stop	significant	Log2FC	ID	strand	homologue
gene301	gene301	349851	gene	type VII secretion system protein EsxS	0.264084	350144	no	0.709786	TBPG_RS...	+	Rv0287
gene300	gene300	348261	gene	PPE family protein	0.865093	349802	no	0.57415	TBPG_RS...	+	Rv0286
gene303	gene303	350475	gene	ESX-3 secretion-associated protein EspG3	0.4047	351362	no	0.687741	TBPG_RS...	+	Rv0289
gene302	gene302	350174	gene	ESAT-6-like protein EsxH	0.0357088	350464	yes	0.873558	TBPG_RS...	+	Rv0288
gene305	gene305	352824	gene	type VII secretion-associated serine protease	0.999479	354209	no	0.680973	TBPG_RS...	+	Rv0291
gene304	gene304	351406	gene	type VII secretion integral membrane protein EccD	0.999479	352827	no	0.594182	TBPG_RS...	+	Rv0290
gene306	gene306	354206	gene	type VII secretion protein EccE	0.999479	355201	no	0.575594	TBPG_RS...	+	Rv0292
gene299	gene299	347950	gene	cell motility protein	0.999479	348258	no	0.584275	TBPG_RS...	+	Rv0285
gene298	gene298	343961	gene	type VII secretion protein EccC	0.999479	347953	no	0.382581	TBPG_RS...	+	Rv0284

Figure 3.4: A window from Cytoscape showing a region that represents a possible operon where one of the genes was found to be differentially expressed. The nodes width and colour is scaled based on the log2 fold change in expression, and edges linking them to variants or TFs can be seen. In the table panel various attributes of the gene nodes is visible including the functions. Using the circular layout, Cytoscape can arrange the genes in a circle, representing the structure of a bacterial chromosome.

3.4 Conclusions

The Cell pipeline provides an ordered structure for the analysis of next generation sequencing data and reporting tools that not only amalgamate data, but also intelligently highlight probable causes of differential expression in a system. The pipeline is able to take advantage of pan-transcriptomes generated by GenGraph, as well as up to date information from databases via the API. As the number of data layers increase, it will become increasingly hard for researchers to manually interpret results, making the development of smart pipelines imperative. The Cell pipeline will be used in chapter 5 for the comparison of two MTB isolates to identify the cause of an altered virulence phenotype.

"If only we had laboratories to produce self-replicating scientists, to explore all the worlds. Ah, but we do! They're called university campuses."

Terry Pratchett

4

Identification and differential expression of small RNA in two closely related *Mycobacterium tuberculosis* isolates

4.1 Introduction

Gene regulatory networks are already highly complex and yet incomplete. The addition of sRNA and other non-coding RNA regulatory elements to this network allows us to explain regulatory cascades that were previously unexplained by the currently known regulatory elements. But with this additional layer, there is an increase in complexity, making the networks harder to navigate. The application of next generation sequencing technologies allow us to investigate how changes in the levels of sRNA effect the other components of this

network as well as identify previously unknown sRNA. Previous research has shows that the stimuli that these sRNA respond to differs, as well as the targets that they are thought to regulate.

In this chapter we aim to identify the sRNA that are differentially expressed between isolates S507 and S5527, and assess the consequence of their altered expression by generating profiles for the sRNA that describe the processes they are regulating and the conditions under which the sRNA are active. As we are using next-generation sequencing technologies and *in silico* methods, this requires us to determine which protocol is optimal for small RNA sequencing sample preparation to achieve an accurate representation of the sRNA abundance in the samples. Additionally we investigate what portion of sRNA is yet to be discovered in MTB by using the small RNA sequencing data to identify potential novel sRNA in isolate W-148.

We find that when comparing the S507 and S5527 isolates, many of these sRNA are differentially expressed, and share certain characteristics including transcription factors and metabolic processes. The combined effects from these sRNA indicate that the more virulent S5527 may have a tempered dormancy response, existing in a generally more active state. Additionally from an explorative search for novel sRNA we find that the known sRNA may be an underestimate of the total sRNA in MTB, and that further *in vitro* experiments are required to ascertain the full compliment of sRNA.

The results of this chapter will compliment the comparison of the two strains by providing a more complete regulatory network, and additionally high-lighting metabolic pathways or regulatory elements that may be effected.

4.2 Materials and methods

4.2.1 Identifying known sRNA in H37Rv and W-148

A list of known sRNA identified in *Mycobacteria tuberculosis* (MTB) was obtained from various publications [112, 93, 73] and their genomic locations of the sRNA were obtained from TubercuList [155]. The sRNA positions were relative to H37Rv, and using these co-

ordinates, the sRNA sequences were extracted and added to an data file for each sRNA for target prediction downstream. In order to create an annotation file for the sRNA in the W-148 isolate, the nucleotide BLAST was used to identify the position of the H37Rv sRNA sequences in the W-148 genome.

4.2.2 Selecting the optimum protocol for the purification of sRNA out of total RNA for sequencing

Sample collection and growth was conducted by members of Rob Warren's research group based at the University of Stellenbosch, with the sequencing being done by Jonathan Featherson at the Agricultural Research Council in Pretoria. Before sequencing of the total sample set, three different treatments were tested at a smaller scale with the aid of Jonathan Featherson. Treatment with only 5' RNApolyphosphatase, treatment with 5' RNApolyphosphatase and T4 Polynucleotide Kinase, and treatment with 5' RNApolyphosphatase and T4 Polynucleotide Kinase with additional ATP. The samples were run on an Illumina HiSeq 2500 using version 4 SBS chemistry (2x125bp) and the small RNA preparation kit RS-200-0012. The reads were trimmed with trim_galore removing low quality base pairs and reads as well as adapter trimming and mapped to the H37Rv genome using BWA-MEM [152]. Mapping quality was assessed using qualimap and the number of reads mapping to the known sRNA sites determined using samtools and the annotation file containing both known sRNA and gene annotations from the H37Rv genome. As different numbers of reads were produced for each condition, the FPKM values of each sRNA was used as a indication of relative coverage.

4.2.3 sRNA sequencing protocol and experimental design

Each isolate was sampled under under four conditions, early logarithmic phase growth (Elog), stationary phase (Stat), middle logarithmic growth (ML(C)), and middle logarithmic growth treated with 5 mM hydrogen peroxide for 6 hours (ML(T)), each with three biological replicates resulting in a total of 24 samples (Figure 4.1). The growth curves were determined by normal OD600 readings. This concentration of hydrogen peroxide resulted in the highest number of differentially expressed genes in a study conducted by Voskuil textitet. al. (2011)

[156]. For the removal of rRNA from the total RNA samples the Truseq stranded mRNA library preparation kit (RS-122-2101) with the Bacterial Ribozero kit (MRZMB126) was used. For the sRNA work our collaborators used the small RNA preparation kit RS-200-0012. The samples were sequenced on an Illumina HiSeq 2500 using version 4 SBS chemistry (2x125bp), with approximately 10 million reads per sample loaded into the lane. The reads were trimmed, filtered, and aligned using the same protocol as described during the sRNA purification protocol selection. The reads were aligned both to the H37Rv genome for differential expression analysis and to the Beijing W-148 genome for identification of novel sRNA.

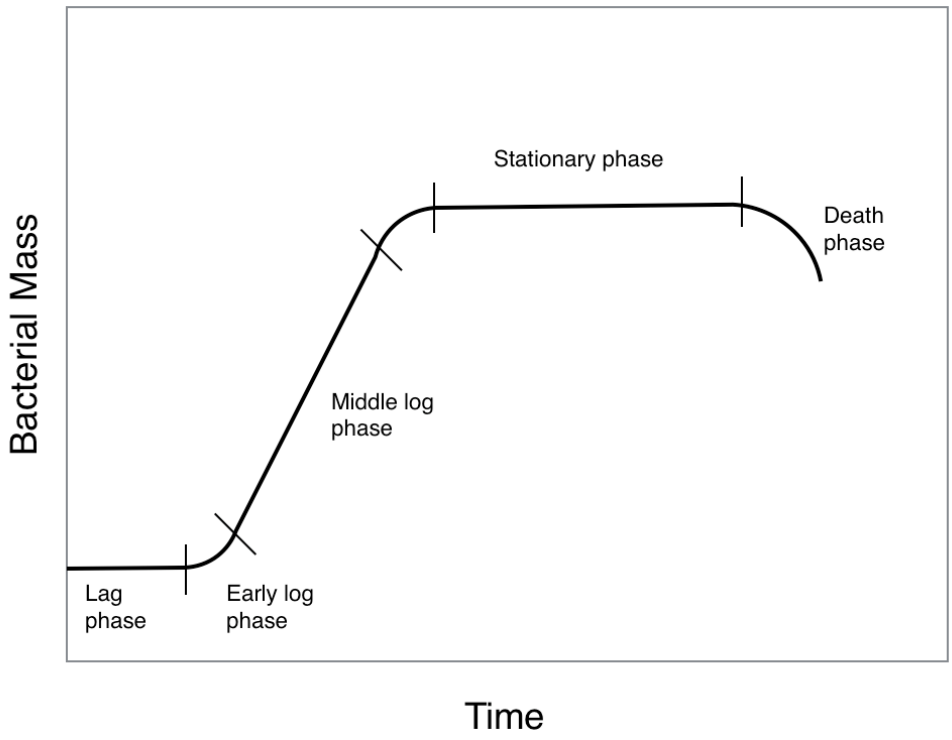


Figure 4.1: In this study isolates were sampled during the early logarithmic phase (Elog), middle logarithmic phase (ML(C) / exponential phase), and the stationary phase (Stat) of growth as well as a H_2O_2 treated ML phase sample (ML(T)). These phases were determined by measuring optical density with a spectrophotometer to determine the number of bacteria in suspension. This was measured at intervals so a rate of change could be calculated and the growth phase determined.

4.2.4 Identification of novel sRNA

In order to identify novel sRNA, a tool was developed that scans intergenic regions of the genome for windows larger than 50bp that have a read depth greater than twice that of the mean. Reads were aligned to the W-148 genome using BWA-MEM, and the regions scanned for each of the 24 samples to ensure sRNA that are only expressed under certain conditions are detected. These regions were then exported to individual GFF3 files, which were then merged into one exhaustive consensus annotation file containing all possible novel sRNA.

Potential novel sRNA that occur as a result of multi-mapping reads were identified by nucleotide BLAST searches of the sRNA sequences to the whole W-148 genome. The sRNA candidates with sequences that overlapped both coding and integenic regions were excluded. The average counts across all samples for each candidate sRNA was calculated, and the sRNA ordered by this value. The sRNA candidates with a read depth less than 200 were excluded from further analysis. A local nucleotide BLAST database was created using the candidate sRNA from W-148, and the known sRNA sequences from H37Rv were queried against it to identify homologues. All candidate sRNA sequences were then screened for possible protein products by use of the NCBI BLASTX platform. Candidates were annotated accordingly with any significant hits, and broadly classified as having no protein matches, having partial or matches to hypothetical proteins, and as matching a known protein. In order to determine the strandedness of the sRNA, a bam file was converted to a stranded form using SAMtools then converted to a pandas data frame and the region visualised.

4.2.5 Differential expression analysis of sRNA

The differential expression analysis of sRNA was conducted using the Cell pipeline, as detailed in chapter 2, with BWA-MEM used for the sequence alignment and Cuffdiff for the differential expression analysis. The reads were aligned both to the H37Rv genome using the known sRNA annotation file for the isolate, and to the W-148 isolate genome using the high-confidence list of novel sRNA we identified for this strain. The results were analysed using the R cummeRbund software package that is part of the Bioconductor toolkit and custom python data analysis scripts.

4.2.6 Predicting the targets of sRNA

Three sRNA prediction tools were tested, IntaRNA, RNAPredator, and TargetRNA2. The results of each were compared to determine the level of overlap between the results, the usability of the tool, the richness of information that they provide, and the current literature where the different methods were compared. The targets as determined by interRNA were used for further analysis. The number of significant targets for each sRNA was determined and a gene set enrichment analysis (GSA) performed on this set using the database for annotation, visualisation, and integrated discovery (DAVID). This was used to generate a profile for each sRNA in order to identify specific cellular processes the sRNA may be regulating. Unfortunately the W-148 genome is not supported by IntaRNA at this time, and sRNA target prediction could only be conducted relative to H37Rv. The expression data for the sRNA was combined with the expression data for their targets (Detailed methods found in chapter 5) and added to the sRNA profile.

4.3 Results and discussion

Sequencing of the sRNA using the selected treatment protocol allowed for the detection of novel sRNA in W-148, and well as the expression profiling of the known sRNA found in H37Rv and their targets. With the high level of coverage used we observed a large number of differentially expressed sRNA with several standing out as being strongly linked to a particular condition or isolate. When comparing the two strains S507 and S5527, 8 sRNA were differentially expressed between the two isolates with a log2FC greater than 0.5 across all conditions. By using *in silico* methods for sRNA target prediction, the gene targets of each sRNA were identified, linking the sRNA expression profiles with the expression profiles of their targets. Additionally, the sRNA sequencing data was used to identify 152 candidate sRNAs in the two strains when using the W-148 genome as a reference. Of these, 56 sRNA passed preliminary quality control and represent probable sRNA.

4.3.1 Testing sRNA preparation protocols

As a defined protocol for the sequencing of sRNA in bacteria has not been established, three different treatments were compared. By including both sRNA and coding sequence annotations, the specificity of the treatment for sRNA could be determined. The treatment by 5' RNApolyphosphatase produced the best result, with the reads mostly appearing to originate mostly from sRNA and not mRNA (Table 4.1). Difficulty in the mapping of the second read in the pair was noted, with only between 2.2% and 4.8% of the R2 reads mapping, possibly due to the short length of the fragments being sequenced.

Table 4.1: Results of different sRNA purification methods. FPKM: Fragments Per Kilobase of transcript per Million mapped reads.

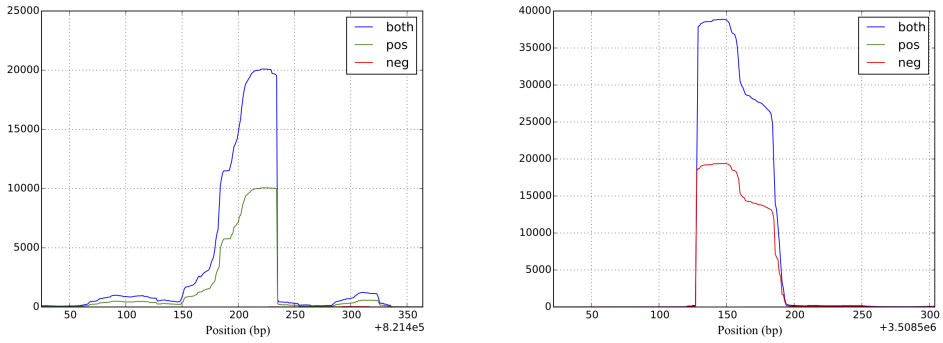
Treatment	Mean coverage of sRNA	Mean coverage all genes	Total coverage	sRNA total coverage	Mean FPKM for sRNA	Percentage mapping to only sRNA	sRNA total coverage normalised to data input
5' RNApolyphosphatase	5,184	130	323,104	165,903	554,913	51%	157,846
T4 Polynucleotide Kinase	1,640	932	776,858	52,492	169,591	16%	52,492
T4 PK + ATP	3,300	125	269,813	105,616	65,195	39%	17,197

4.3.2 Identification of novel sRNA

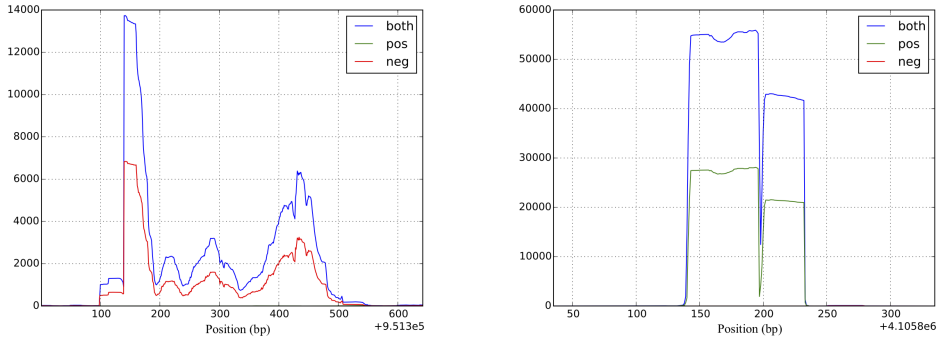
Using the reads mapped to isolate W-148, 228 candidate novel sRNA were identified. After removing those with an average read count less than 200 (to account for background signal), 152 remained. Of this set, 28 matched with known proteins, and are likely the result of incorrect annotations or derived from mRNA, and another 68 had either partial matches to known proteins, or to hypothetical / uncharacterised proteins. This left 56 sRNA that were promising candidates for novel sRNA in W-148. These sRNA varied in length and coverage and demonstrated some interesting features including the signs of sRNA degradation in sRNA_42 and sRNA_187 (Figure 4.6). Of these identified sRNA, only 9 had identifiable homologues of known sRNA in H37Rv by BLAST (B55, B11, C8, Mcr3, Mcr7, Mcr11, MTS2975, MTS1338, MTS2823), even though all known sRNA previously described in H37Rv were confirmed to be present in W-148 using BLAST. Upon visual inspection of the reads mapping to these regions, the reason the remaining 21 of the 30 known sRNA were not detected by this method

was due to low coverage (10 of the sRNA), because they overlapped coding regions (another 10 of the sRNA), and because they were too short (Mcr5 is less than 44bp). Refinements to the novel sRNA detection method in terms of minimum coverage and tolerance for coding region overlap will improve the sensitivity in future iterations.

Differential expression analysis comparing the expression of the 56 newly identified sRNA identified 23 differentially expressed between S507 and S5527, 6 of which were differentially expressed in more that one condition. As a per-condition breakdown, 6 were differentially expressed during ML(C), 2 during ML(T), 1 during Stat, and 20 during Elog. Included was sRNA_211, a homologue of MTS2823 in H37Rv. This sRNA was increased in isolate S5527 during the Elog phase, and is consistent with the observation in H37Rv for MTS2823. These results indicate that the currently known sRNA in MTB may only be a fraction of the total population, and further experimental validation of these candidate sRNA is required both to complete that set and to test the accuracy and sensitivity of sRNA sequencing based methods for sRNA identification.



(a) sRNA_42 **(b)** sRNA_187



(c) sRNA_47 **(d)** sRNA_210

Figure 4.2: Plots showing per-base coverage of the candidate sRNA during the Elog phase of growth replicate 1. (A,B) The sRNA_42 is found on the positive strand while the sRNA_187 is found on the negative strand. The directional tapering of coverage may be as a result of varying degrees of sRNA degradation. (C) sRNA_210 (B11 in H37Rv) showing a possible variant resulting in a gap in the sequence coverage. (D) sRNA_47 is over 500bp in length, and shows inconsistent coverage. A quirk in the nature of bam sequence format resulted in the "both" coverage line becoming double the total coverage and will be corrected in future versions of the tool.

4.3.3 Prediction of sRNA targets and profiling

A recent review of sRNA target prediction tools was conducted by Pain *et. al.*, [118] and three tools were suggested for sRNA target prediction. Of these, IntaRNA showed the greatest accuracy and most informative output. One limitation of these tools is the limited support for genomes, apart from the usual reference strains. Only one isolate, H37Rv, was

supported, and target prediction for W-148 sRNA could not be conducted at this time. It is also apparent from the review that even with the improvements in the latest generation of *in silico* sRNA target prediction methods, false targets may have been called [118]. This should always be a consideration when interpreting results of this nature, particularly regarding the effects of multiple testing.

An example is that both *cis* (including ASdes, ASpks) and *trans* encoded sRNA are predicted to have similar numbers of targets. This non-specific targeting of *cis* encoded sRNA is not entirely congruent with the current model of their action, where they are thought to primarily regulate the gene to which they are anti-sense to. This may be as a result of shortcomings in the models employed by target prediction tools, or it may be the case that these sRNA strongly bind to the gene they are anti-sense to and weakly regulate other mRNA through *trans*-sRNA like interactions.

The term clustering for the IntaRNA sRNA target prediction tool allowed us to identify particular cellular functions that the sRNA are regulating, from which we may begin to determine what cellular processes the sRNA are regulating and how MTB uses them to respond to stimuli (Table 4.2). By combining the expression data from the sRNA and their predicted targets, the conditions under which the sRNA are expressed can be deduced (Figure 4.3), and a more complete profile for these sRNA can be generated (Table 4.3). These profiles allowed us to better interpret the differential expression observed between conditions or between the two isolates, providing context and potential outcomes to the altered sRNA expression.

Table 4.2: Target GSA profiles of the sRNA predicted by DAVID. *The number of sRNA targets is shown at two different p-value cutoffs, 0.05 and 0.01.

sRNA	Top ontologies found in the list of targets	*Number of sRNA targets
F6 (Synonyms: Mcr14, Mpr13, MTS0194)	Mono-oxygenase, oxidoreductase, hexachlorocyclohexane degradation	94 / 8
B55 (MTS0479)	Amino-acid biosynthesis, cofactor binding, oxidoreductase, transmembrane	15 / 1
B11 (Synonym: Mpr19)	ATP Binding and nucleotide binding	176 / 18
G2	Nucleoside binding adenyl nucleotide binding	181 / 40
C8 (Synonyms: Mcr6, 4.5S)	Cofactor binding, transmembrane protein	133 / 19
ASdes (Complementary genes: desA1, Rv0824c)	Nucleotide binding, ATP binding	180 / 33
ASpks (Complementary genes: desA2, Rv1094)	ATPase	192 / 53
AS1726 (Complementary gene: Rv1726)	Cation binding (iron in particular), transmembrane	200 / 46
AS1890 (Complementary gene: Rv1890c)	Nitrogen compound biosynthetic process, vitamin biosynthesis, amino-acid biosynthesis	151 / 33
ncrMT1302		
Mcr3 (Synonym: Mpr7)	Propanoate metabolism, valine, leucine, and isoleucine degradation	171 / 34
Mcr5	Transmembrane	115 / 16
Mcr7	DNA recombination and replication, integral to membrane	191 / 48
Mcr10	Palmitate, cell membrane, cation binding	162 / 44
Mcr11 (Synonym: MTS0997)	ATPbinding nucleotide binding, cell membrane	192 / 42
Mcr15	ATPase, nucleotide binding	140 / 26
Mcr16	Transmembrane, nucleotide binding, ATP-binding	152 / 34
Mcr19	Peptidase activity, hydrolase, ATP binding, nucleotide binding	141 / 18
Mpr5	Co-factor binding, vitamin binding, regulation of transcription	198 / 38
Mpr6	Purine ribonucleotide biosynthetic process	192 / 40
Mpr11	Purine ribonucleotide biosynthetic process, ATP-binding	163 / 44
Mpr12	Transferase, cell membrane	166 / 32
Mpr17	Regulation of transcription	116 / 19
Mpr18	DNA binding, transcription regulation	149 / 34
MTS1082	Transmembrane, peptidase activity	164 / 44
MTS2975	Cation binding	147 / 15
MTS0858	Transmembrane	132 / 17
MTS1338	Nucleotide binding, ATP-binding	140 / 32
ncrMT3949	Phosphorylation, ATP-binding, co-factor binding.	162 / 40
ncrMT1234	Protelysis, terpenoid backbone biosynthesis	137 / 24
MTS2823	Nitrogen compound biosynthetic process, vitamin biosynthetic process	186 / 49

4.3.4 Growth condition specific sRNA

Certain sRNA are known to respond to different conditions, including in response to different growth phases [112]. As part of the sRNA profiling, we report the following sRNA that appear to be growth phase dependent.

Notable sRNA with increased expression during the stationary growth phase:

It is known that sRNA play a role in dormancy and responding to starvation in MTB [93, 73, 111]. Under these conditions, the cell enters the stationary phase as resources

Table 4.3: The number of sRNA targets that are differentially expressed when comparing different conditions. Cells are shaded where the sRNA is significantly up (blue) / down (red) regulated in condition 1 vs condition 2. The number of sRNA target genes that are significantly up or down regulated between the conditions is also displayed. As an example, the sRNA B11 has lower expression in Stat when compared to ML(C), and in isolate S507, 5 of its targets show increased expression and 1 shows decreased expression.

sRNA	ML(C) vs Stat				Elog vs ML(C)				Elog vs Stat				ML(C) vs ML(T)			
	S507		S5527		S507		S5527		S507		S5527		S507		S5527	
	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg
F6	4	0	6	1	3	1	4	3	9	4	9	7	1	0	1	0
ncrMT3949	5	0	4	1	2	2	2	2	10	8	11	4	0	2	1	0
MTS1338	4	2	5	0	4	3	4	1	9	6	10	3	0	1	2	0
Mcr11	6	1	6	1	1	0	1	0	9	11	13	8	0	0	5	0
G2	4	3	4	3	2	4	5	2	8	12	10	8	0	1	1	2
Mcr7	4	0	4	2	4	3	2	5	13	12	10	7	0	0	0	0
Mcr5	2	1	2	0	1	5	2	6	5	12	6	9	0	1	1	0
Mcr3	8	1	9	0	4	4	6	3	14	5	16	6	1	0	1	0
AS1890	2	1	1	0	0	4	0	4	7	7	6	8	0	1	1	0
B11	5	1	8	2	2	4	5	1	14	9	16	7	1	1	4	2
MTS1082	4	0	8	0	5	2	6	3	11	8	14	5	0	2	0	2
ASpks	6	1	8	0	2	1	4	3	17	8	14	11	0	3	0	0
MTS0858	2	1	2	1	2	4	2	4	10	8	9	7	0	4	0	3
B55	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	1
MTS2975	4	1	3	2	6	5	5	4	11	8	8	6	0	3	1	0
ASdes	1	0	1	1	0	5	0	3	8	10	6	6	0	1	1	1
Mpr18	7	1	6	2	1	2	2	3	12	6	14	7	0	0	3	2
Mpr11	3	0	5	0	4	0	4	0	9	3	8	4	0	0	1	0
Mpr12	2	0	2	2	2	2	2	3	8	10	13	6	0	0	1	1
Mpr17	4	0	4	0	3	1	2	3	11	8	8	5	0	2	1	0
Mcr19	3	0	4	1	3	2	3	1	6	8	4	8	0	0	0	0
Mcr15	5	0	6	1	4	1	7	1	14	8	13	4	0	3	2	2
Mcr16	8	0	8	1	6	3	7	2	20	8	20	4	1	2	4	0
Mpr5	2	0	3	2	2	2	3	0	8	9	10	7	0	0	3	2
Mpr6	4	1	3	1	5	2	8	3	13	9	14	12	0	0	1	1
AS1726	5	1	5	3	3	1	5	3	8	9	11	8	0	1	1	1
Mcr10	1	1	0	2	1	7	2	5	4	8	6	8	0	0	1	1
ncrMT1234	3	1	3	1	4	2	2	3	10	5	13	4	0	0	2	0

become scarce and the rate of cell death matches the rate of new cells being formed. This phase is also associated with the expression of various virulence factors and of genes involved in the β-oxidation of fatty acids [157]. In our results the sRNAs F6, MTS1338, and mcr11 appear highly expressed under these conditions.

Many of these sRNA (F6, MTS1338, mcr11) have already been reported as active during stationary phase growth [93, 73, 111, 158], an exception being MTS2823, previously reported to be highly expressed during stationary phase, but found to be strongly expressed during the early logarithmic phase in our datasets.

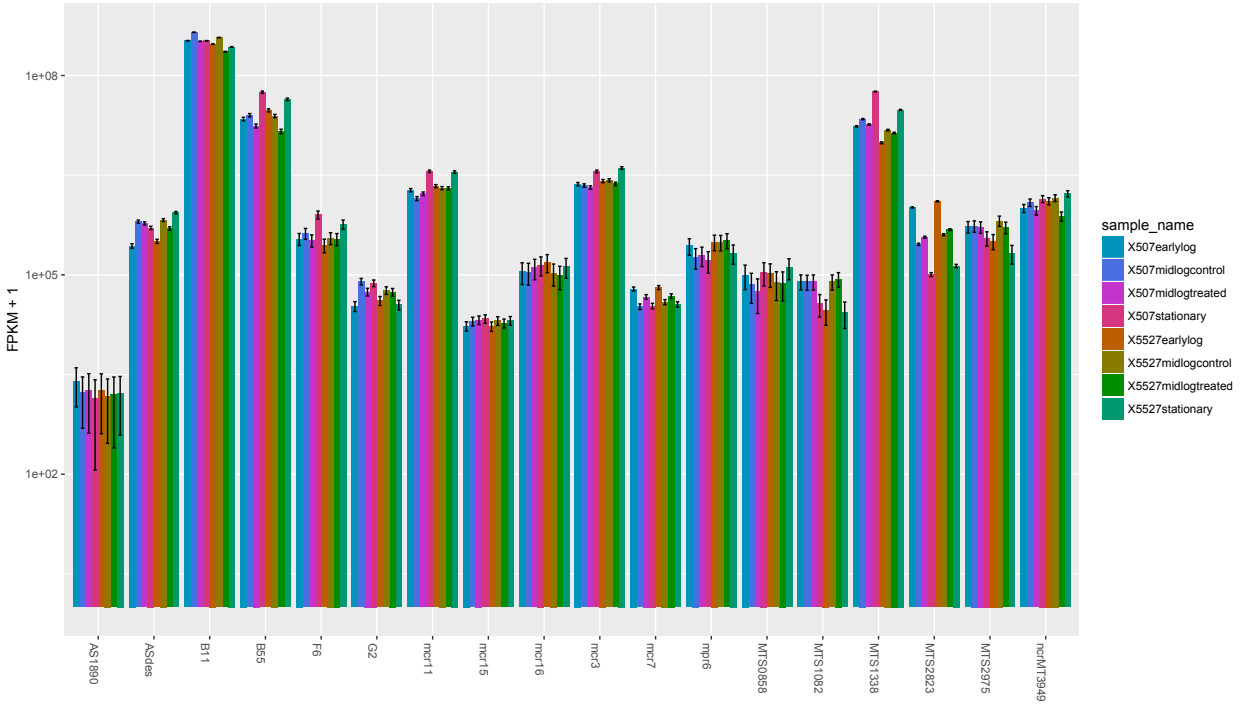


Figure 4.3: The expression levels of the known sRNA included in this study with consistent stable expression across replicates.

We found two of the sRNA (B55 and mcr3) targeted Fad genes (*fadD31*, *fadE2*, *fadE17*, *fadD13*) that are involved in fatty acid metabolism, specifically lipid degradation. None of these Fad genes showed significant differential expression, suggesting that their levels may be mediated at the post-transcriptional level by the sRNA. F6 was also found to target genes involved in fatty acid metabolism, but this time biosynthesis, and includes *agpS* (a possible alkyldihydroxyacetonephosphate synthase), *fadD28* (a fatty-acid-AMP synthase), and *cyp128* (a heme-thiolate monooxygenase which can oxidize fatty acids).

Virulence factors are also reported to be expressed under stationary phase conditions, and both mcr3 and F6 included targets linked to virulence. For F6 this includes *vapB27* (involved in virulence, detoxification, and adaptation), and in the case of mcr3, *vapC19* (a possible toxin producing gene), *esxU* (a secreted virulence factor), and *mpt70* (a secreted immunogenic protein producing gene) were predicted targets.

In addition to starvation, F6 expression has also been linked to H_2O_2 [93]. F6 appears to target three monooxygenase genes (*cyp128* and *cyp141* which are part of the heme-

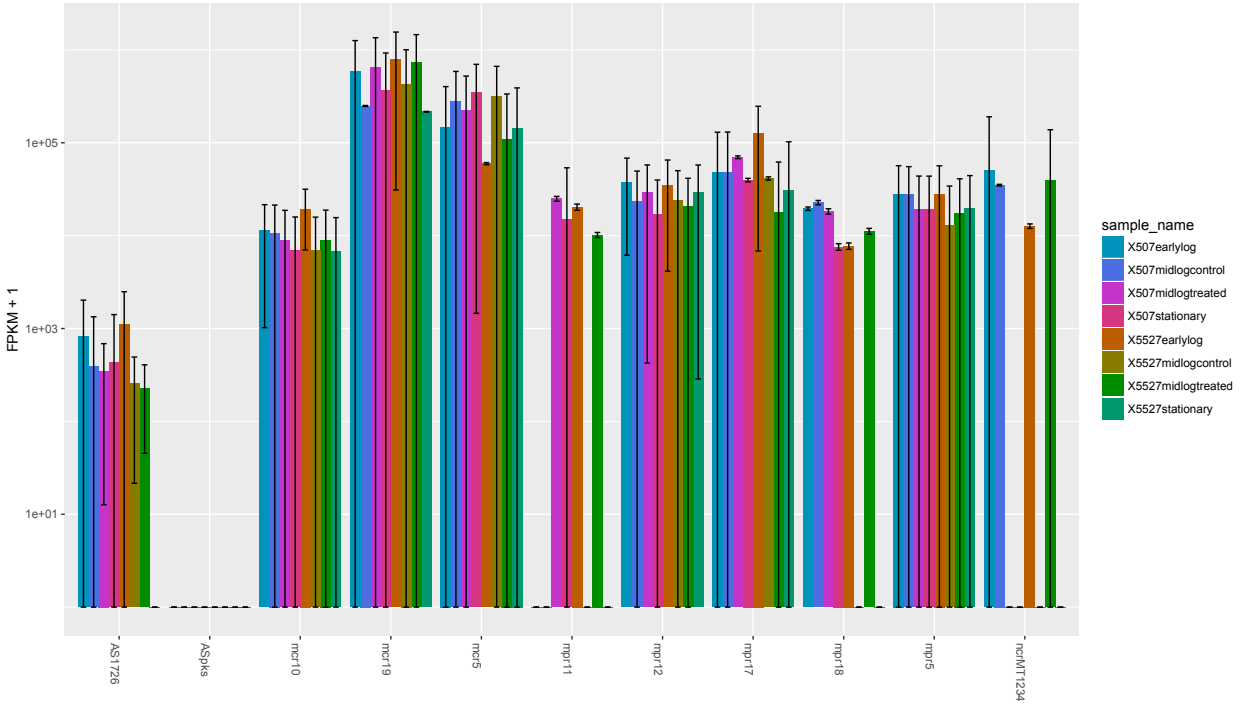


Figure 4.4: The expression levels of the known sRNA included in this study with inconsistent expression across replicates. The variance in these samples was high, due to no reads mapping in some samples or conditions, and in the case of ASpks, no reads mapping at all.

thiolate monooxygenases, and Rv0892, a probable monooxygenase) and a peroxidase *bpoA*, all involved in oxidation reduction reactions. Other stress response genes are targeted by F6 and include *sigF* and *lexA*. The sigma factor *sigF* has been shown to bind to F6 in CHIP-seq experiments, and is a MTB stress response transcription factor expressed during the stationary phase [159]. The second gene, *lexA*, is involved in the regulation of nucleotide excision repair and sos response. LexA is a repressor of a number of SOS response genes, making the result of an increase in F6 a decrease in LexA, and consequently expression of SOS response genes.

The remaining sRNA mcr11 and MTS1338 appear to have a far more general profile of targets, but appear enriched for general GO terms including ATP-binding. Some targets of MTS1338 include an anion transporter ATPase, a molybdopterin molybdenumtransferase and a GTP cyclohydrolase, where mcr11's targets include a phosphate starvation-inducible protein PSIH, a inorganic polyphosphate / ATP-NAD kinase, and a transmembrane ATP-

binding protein ABC transporter.

Notable sRNA with increased expression during the early logarithmic growth phase:

During the early log phase of growth, cells undergo shifts in metabolism, coming out of dormancy and gearing towards higher energy use and biosynthetic processes for cellular division [160]. During this phase, the repair of damage done during the stationary phase is conducted [161]. The two sRNA that show the greatest expression during this phase are mcr7 and MTS2823.

The sRNA mcr7 has been identified as a regulator for the TAT secretory system in MTB, binding to *tatC* in a manner that blocks ribosomal binding [162]. Mcr7 is itself regulated by PhoP (Rv0757 / TBPG_RS16745), with a phoP mutant shown to have a complete lack of mcr7 expression [162]. PhoP regulates 30 genes directly including other transcription factors [162] and is closely linked with virulence, with a mutation in phoP contributing to the attenuated virulence phenotype seen in H37Ra [163]. In our datasets we see that *phoP* has slightly increased expression during the early log phase, but not significantly. It is possible that as the principle inducer of mcr7 expression, even slight fluctuations in *phoP* expression result in major mcr7 expression changes or that there is further post transcriptional regulation of *phoP* in MTB. Top targets of mcr7 predicted by interRNA include a cluster associated with DNA replication, recombination and repair. These include two possible resolvase genes that act to prevent the co-integration of foreign DNA into the chromosome, thus maintaining genome integrity. This sRNA was also seen to respond to treatment with H_2O_2 during ML(T) growth in our datasets, during which oxidative stress could potentially lead to DNA damage. The remaining top targeted genes appear not linked by any common process, but includes *glpQ1* which is flanked by the sRNA ncrMT3949 (downstream) and MTS2975 (upstream) found on the opposite strand. The expression of these two sRNAs appears not to be linked to that of mcr7. It has been suggested that mcr7 is not truly an sRNA, and is annotated as two hypothetical proteins in CDC1551 (MT2466 and MT2467) [112] and as two acid and phagosome regulated proteins (AprA and AprB) in H37Rv involved in the mycobacterial response to low pH.

The sRNA MTS2823 is found between Rv3661 and Rv3662c (both reported to possibly play a regulatory role in cellular differentiation and involved in virulence, detoxification, adaptation according to TubercuList) and is found at decreased levels during the stationary phase in our datasets. This sRNA has been associated with infection, stationary phase, and low pH [73, 111], and was predicted by IntaRNA to target multiple membrane proteins, including a copper transport gene mctB. Additional potential target genes include *nadE* (Rv2438c) which is involved in biosynthesis of NAD, *ribH* (Rv1416) a probable riboflavin synthase beta chain RibH, and *parA* (Rv3918c) a probable chromosome partitioning protein ParA. The over expression of this sRNA has been observed to result in widespread down regulation of energy metabolism genes similar to what is observed during the transition from exponential growth to stationary phase [73].

Notable sRNA with increased expression during the middle logarithmic growth phase:

The middle logarithmic phase or exponential phase is a period characterised by cell doubling. During this phase we observed two sRNA that were highly expressed, ASdes and B11.

The sRNA ASdes is one of the *cis* encoded sRNA that also shows potential for *trans* interaction. It is found antisense to *desA1* (Rv0824c), an essential acyl-ACP desaturase which is responsible for the conversion of saturated fatty acids to unsaturated fatty acids. MTB also contains a second homologue of desA1, desA2 (Rv1094), which is also recognised by ASdes in this *trans* manner [93]. The two targets desA1 and desA2 are reported to be repressed during the stationary phase [164], while in our data this was true for *desA1* in isolate S507, in isolate S5527 *desA1* showed lower expression in the Elog phase and ML(T) compared to the Stat phase. The second target, *desA2* showed decreased expression in the stationary phase for both isolates in our results.

The second sRNA active during the ML(C) growth phase was B11. This sRNA was reported to respond to H_2O_2 in MTB, and is involved in intracellular survival during the early stages of infection [112]. It is located between Rv3660c and Rv3661, and has a putative SigA promoter immediately upstream of the 5' end [93]. In our results we see that lowered expression of this sRNA during the stationary phase of growth results in many of its targets

becoming significantly upregulated, possibly indicating that this sRNA is largely responsible for their regulation during this phase (Table 4.3).

Notable sRNA differentially expressed in response to treatment with hydrogen peroxide:

Treatment with H_2O_2 in culture elicits a similar response to NO in MTB [156] and approximates the phagosomal environment *in vitro*. The highest fold change between the treated and untreated samples was observed for B11, B55, and G2 (Table 4.4). They showed a decrease in abundance during treatment, with B11 and B55 having being previously reported to react to treatment by H_2O_2 [93, 165]. B55 is found at the end of what is possibly an operon that contains two vap genes *vapB28* (Rv0608 / TBPG_RS03150) and *vapC28* (Rv0609 / TBPG_RS03155), though no changes in the expression of either of these genes was observed as a result of H_2O_2 treatment indicating B55 is under independent regulatory control. An additional sRNA reported to respond to H_2O_2 levels is F6, which showed a significant decrease in abundance during treatment for isolate S507 but not for S5527, where it showed only a slight decrease in our datasets.

The sRNA not previously identified as responding to H_2O_2 treatment include mcr7, and ncrMT3949. The sRNA mcr7 showed an increase in expression during treatment, whose predicted targets are involved in mechanisms to maintain genomic integrity. Its involvement during a period of oxidative stress brought on by the H_2O_2 is therefore fitting. Specific genes targeted by mcr7 with functions relating to oxidative stress include a carbon monoxide dehydroxinase large subunit gene (Rv0373c). The second unreported sRNA ncrMT3949 had decreased expression in the treated sample that was more pronounced in isolate S5527 (Table 4.4). The targets include Esat-6 genes, a stress response protein GrpE, and three oxidoreductase genes including *fadB* and *nuoK*, all of which would see increased protein levels as a result of the lowered ncrMT3949 sRNA levels. This sRNA is downstream from a Bacterioferritin BfrB (Rv3841) gene which encodes a protein that stores iron in a non-toxic, readily available form. The *bfrB* gene is seen to be induced by hypoxia [166], and thus treatment with H_2O_2 would generate the opposite response. Our observations support this, and suggest ncrMT3949 and BrfB are co-expressed in response to oxidative stress levels.

The expression of G2 appears to differ between the two isolates, with decreased expression of G2 only observed in S507 in response to treatment with H_2O_2. Additionally, G2 is reported to be up regulated during the exponential / middle log phase [93], which is in accordance with our results for S5527 but not for S507, where increased expression of G2 was observed in S507 during stationary phase growth (Figure 4.3). In this way the S5527 isolate follows the sRNA expression patterns seen in H37Rv but not S507. It was noted that a possible SigC promoter was identified upstream of G2 [93], which is required for lethality in mice [167]. This could indicate that G2 is linked to pathogenicity, and contributes to the phenotypic differences between the two strains.

Generally, despite the number of significantly differentially expressed sRNA as a result of the H_2O_2 treatment, most of the sRNA did not show large fold changes, and those that were observed were consistent with previous studies. When comparing the responses of the two strains, the most notable differences were in the expression of G2, which is linked to pathogenicity, and ncrMT3949 which showed the greatest response and notable links to oxidative stress response in the genes it was co-expressed with, in the case of *brfB*, and the genes that it was predicted to target.

4.3.5 sRNA found to be differentially expressed between isolate S507 and S5527

When comparing the sRNA expression profiles of the two isolates, a large proportion showed a significant differential expression due to the high coverage provided by the reads. Here we discuss the sRNAs that showed the greatest change in expression between the two isolates (Table 4.6). As there are only a few sRNA, which were sequenced with high coverage with little variance between samples, the p-values were mostly all less than 5e-05, which is the minimum p-value reported by cuffdiff (see release notes: http://cole-trapnell-lab.github.io/cufflinks/releases/v2.1.0/). As a result, the q-values were identical in many cases.

Table 4.4: The log2 fold changes of sRNA differentially expressed as a result of treatment with H_2O_2. Cells high-lighted in the sRNA column show sRNAs differentially expressed in both of the isolates. Cells high-lighted in the fold change column show a log2 fold change greater than 0.5.

sRNA	Fold change	q-value
S507		
ASdes	-0.0865	0.00883
B11	-0.453	0.000135
B55	-0.540	0.000135
F6	-0.353	0.00754
G2	-0.521	0.000135
MTS1338	-0.272	0.000135
MTS2823	0.354	0.000135
mcr11	0.241	0.000135
mcr3	-0.0980	0.0134
mcr7	0.485	0.000135
ncrMT3949	-0.398	0.000135
S5527		
ASdes	-0.417	0.000135
B11	-0.705	0.000135
B55	-0.770	0.000135
MTS1338	-0.148	0.000135
MTS2823	0.254	0.000135
mcr3	-0.177	0.000135
mcr7	0.298	0.000135
ncrMT3949	-0.917	0.000135

sRNA that were found with lower expression in isolate S5527

- **MTS1338 (ncRv11733)** This sRNA showed decreased expression in Stat, Elog, and ML(C) phase growth in isolate S5527. It has been reported to accumulate to high levels

Table 4.5: Reported sRNA regulatory events during which the sRNA show increased expression.

sRNA	Observed condition	Reported condition	Reference
MTS0194 (F6)	Stationary phase	SigF, starvation, H_2O_2, low pH	[93]
MTS0479 (B55)	H_2O_2 repressed	H_2O_2	[93]
MTS0997 (mcr11)	H_2O_2 induced, stationary phase	Stationary phase, infection	[111]
MTS1310 (G2)	Reduced by H_2O_2	Exponential phase	[93]
MTS1338	Stationary phase	DosR, hypoxia, infection	[73]
MTS2822 (B11)	H_2O_2 repressed, middle log phase	H_2O_2	[165]
MTS2823 (mpr4)	H_2O_2 induced, early log phase	Stationary phase, infection, low pH	[73, 111]
MTS2975		Exponential phase	[73]

during the stationary phase [158] with its accumulation dependent on the DosR transcriptional regulator [73]. As a dormancy related sRNA, its presence at decreased levels in the hyper-virulent S5527 strain could indicate that the S5527 isolate is generally in a more metabolically active state. Or at the very least, the pathways that the targets of the sRNA belong to are more active. These predicted targets include many membrane proteins including an anion transporter ATPase and a ESX-4 secretion system protein EccC4 as well as genes that form part of the folate and biopterin biosynthesis pathways *moeA1* (molybdopterin molybdenumtransferase 1) and *folE* (GTP cyclohydrolase I) (Figure 4.6a). The location of the sRNA itself is running in the antisense orientation amongst a set of probable transmembrane and hypothetical proteins, providing little additional insight into its regulation.

- **G2 (MTS1310)** Over-expression of this sRNA was shown to prevent growth of MTB [93], and with the sRNA targets enriched for terms including transcriptional regulators, nucleoside binding activity and membrane localization, it is clear that the altered expression of this sRNA likely has far reaching effects for the cell (Figure 4.6b). The lowered expression may lead to an increase in the levels of the transcription factors

Table 4.6: The top differentially expressed sRNA (with a log2 fold change greater than 0.5) between the two isolates for each of the conditions. The transcription factors that annotated binding sites from ChIP-sequencing near to the sRNA obtained from the Tuberculosis database (TBDB) are listed. sRNA with a negative Log2FC are expressed at lower levels in S5527.

sRNA	Condition 1	Condition 2	Log2FC	q-value	TFs
MTS1082	507 early log	5527 early log	-1.421	0.000135	Rv0303, TrcR, Rv3597c, Rv0081
G2	507 stationary	5527 stationary	-1.082	0.000135	Rv0081, Rv0324, CsoR
MTS1338	507 stationary	5527 stationary	-0.921	0.000135	DevR
MTS1338	507 early log	5527 early log	-0.821	0.000135	DevR
MTS2975	507 stationary	5527 stationary	-0.759	0.000135	Rv0302, Rv0081, CsoR, Lsr2
MTS2975	507 early log	5527 early log	-0.721	0.000135	Rv0302, Rv0081, CsoR, Lsr2
MTS1338	507 ML control	5527 ML control	-0.549	0.000135	DevR
B11	507 ML treated	5527 ML treated	-0.516	0.000135	14 TF including Rv3249c, CsoR, Rv0081, Rv1353c
mcr11	507 ML control	5527 ML control	0.522	0.000135	Lsr2, Rv0081
mpr6	507 ML control	5527 ML control	0.746	0.000265	TrcR, Rv0023, Rv0081, Lsr2
mpr6	507 ML treated	5527 ML treated	0.753	0.000907	TrcR, Rv0023, Rv0081, Lsr2
ASdes	507 stationary	5527 stationary	0.757	0.000135	Rv2250c

they target. One such target is *pknG* (Rv0410c) a serine/threonine-protein kinase that is linked to glutamate / glutamine levels [168]. A mutant deficient in pknG showed delayed mortality in mice and reduced growth particularly during the stationary *phase in vitro*, which is the same phase that this sRNA is decreased in S5527. Assuming this is a true target of the sRNA that would result in increased levels of PknG in

S5527 relative to S507 in the stationary phase. The sRNA is additionally linked to SigC regulation as mentioned before, and showed lower expression in response to H_2O_2 treatment in isolate S507, but not for isolate S5527. As both pknG and SigC levels have been linked to mortality in mice [167], it is possible that SigC regulates G2, which in turn targets *pknG*, and disruption of G2 expression is a contributor to the phenotypic differences between S507 and S5527.

Increased expression of G2 in isolate S507 during the Stat phase of growth vs Elog may have led to the suppression of some of its target genes. When comparing the expression of genes between the two isolates between these two conditions, we see 2 of its targets becoming induced and four less being repressed (Table 4.3) during this comparison where G2 is most markedly dissregulated between the isolates (Table 4.6).

- **MTS1082** The profile of MTS1082 indicates that it is normally found in decreased levels during Stat phase growth (Figure 4.3) where some of its targets show significant increases in expression when compared to their expression in the ML(C) phase (Table 4.3). The two isolates differ significantly only during the Elog phase, where the expression of MTS1082 in S5527 is less than half of that in S507.

  Both G2 and MTS1082 showed lowered levels of expression in the hyper-virulent isolate S5527. As they are likewise reported to show lowered levels of expression in response to treatment with antibiotics (isoniazid) [114] and links to oxidative stress response (*sigC* for G2 and *lexA* for MTS1082), it is possible that the expression of these two sRNA is linked by a common regulator. Although the sRNA do not appear to share TFs, a noted similarity between the G2 and MTS1082 regulators was that both Rv3597c / TBPG_RS18940 (*lsr2*, an iron-regulated H-NS-like protein) for MTS1082 and Rv0967 / TBPG_RS15635 (*csoR*, a copper-sensitive operon repressor) for G2 were sensitive to cation levels.

  Another known TF that is reported to bind to MTS1082 is the SOS response transcriptional regulator LexA (Rv2720 / TBPG_RS13150) [91]. LexA is a transcriptional repressor involved in the response to DNA damage, and though *lexA* showed slightly increased expression in S5527 in all samples except middle log phase growth when

treated with H_2O_2, the change was not significant. LexA was also shown to bind to MTS2823, which showed slightly increased expression in S5527, though once again this change was not significant. Considering that the change in expression of MTS1082 was the opposite to that of MTS2823 and the low fold changes observed, it is unlikely that the observed changes in MTS1082 is as a result of altered LexA activity between the two isolates.

- **MTS2975** This sRNA is found between Rv3843c and Rv3844, and has predicted Rv0302 (TetR/AcrR-family TF), Rv3597c (*lsr2*) and Rv0967 (*csoR*) TF binding sites in close proximity, similar to G2 and MTS1082. This sRNA is reported to be active predominantly during the exponential growth phase [73], which is in accordance with what we have observed in addition to activity during the middle logarithmic phase of growth. The targets of this sRNA indicate cation binding as a common feature in the enriched group of 8 genes including a ferredoxin FdxD and a bacterioferritin BfrB which is involved in the storage or iron, and found upstream from ncrMT3949 which is located approximately 2,000bp upstream from MTS2975 (Figure 4.6d).

- **B11 (Mpr19, MTS2822)** This sRNA lies between Rv3660c (septum site determining protein [169]) and Rv3661 (phosphoserine phosphatase), two genes thought to play a regulatory role in cellular differentiation that are orientated facing outward from each other with B11 nested between them, proximal to 14 different TF binding sites as reported by TBDB, with that repertoire including Rv3597c (*lsr2*) and Rv0967 (*csoR*) once again. The targets of this sRNA have annotated functions including ATP binding, ATPase activity, transcription regulators, and metal binding, and includes multiple transmembrane proteins (Figure 4.6e). B11 was confirmed not to be co-transcribed with Rv3660c and overexpression of B11 even under a weak promoter leads to cell death. Expression of B11 in *M. smegmatis* leads to the development of deformed cells in culture [93]. In our data, Rv3660c is highly expressed during the Stat phase and the ML(C) phase, while B11 appeared to be upregulated in the ML(C) phase in both isolates, but with a significant decrease in expression of B11 in isolate S5527 during the ML(T) phase, during which Rv3660c shows an increased level of expression in S5527.

The location of the sRNA between two genes involved in cellular differentiation likewise indicates that this sRNA is linked in some manner to cellular differentiation.

sRNA with increased expression in S5527

- **mcr11 and mpr6** Both of these sRNA are specifically up in S5527 during ML(C) and ML(T) growth (Though for mcr11 the increase during ML with H_2O_2 is not significant). The mcr11 sRNA lies between Rv1264 (adenylyl cyclase) and Rv1265 (involved in mycobacterial intracellular survival) and is active during macrophage infection and the stationary phase [111] as observed in our results. Little else is known about the function of this sRNA, with the predicted targets showing enrichment for fairly general terms such as ATP binding, nucleotide binding, ribonucleotide binding (Figure 4.6a). The sRNA mpr6 is co-transcribed and co-regulated with *sigE* and is found between *sigE* (Rv1221) and anti-sigma factor *rseA* (Rv1222) with RseA negatively regulating *sigE* [111] and expression of *sigE* regulated to some extent by SigH [167] a heat shock response protein. The role of SigE is the activation of the SigE regulon, which encodes genes that are also active during macrophage infection. Notable targets of mpr6 include genes involved in purine biosynthesis and ATP synthesis (Figure 4.6b). The common link between macrophage survival and expression pattern of these two sRNA indicate they may be part of the same regulatory network, though the implication of their increased expression in S5527 during stationary phase growth is unclear.

- **ASdes** This cis-encoded sRNA is found anti-sense to Rv0824c, an acyl-carrier protein desaturase (DesA1) [112] and has been associated with lipid metabolism, specifically mycolic acid biosynthesis in *M. smegmatis* [93, 170]. While these cis-encoded sRNA are generally thought to regulate the gene to which they are in the anti-sense orientation, they may also effect other related genes through the trans-regulatory system, in this case another acyl-carrier protein desaturase DesA2 (Rv1094). The sRNA was also predicted to potentially target genes with nucleotide binding, FAD binding, and oxidation reduction ontology annotations, and includes antitoxin *vapB29* and *vapB17* genes (Figure 4.6c). Whether these "off-target" interactions occur and induce a result

(a) MTS1338 **(b)** G2 (MTS1310) **(c)** MTS1082

(d) MTS2975 **(e)** MTS2822

Figure 4.5: Highest ranked cluster of ontologies for the targets of the sRNA with lowered expression in isolate S5527 as reported by IntaRNA.

in a change in protein abundance requires further investigation.

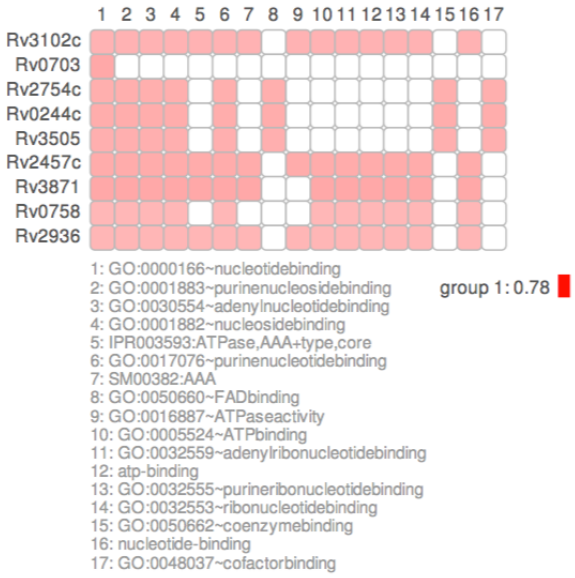
4.3.6 Common themes found in sRNA differentially expressed between the isolates

In the down-regulated sRNA we see that MTS2975 and G2 are both reported to be active during the exponential phase (ML(C)) (Table 4.5), and share Rv0081, Rv0967 (CsoR) TF binding sites. From our results the sRNA B11 likewise shows strong expression in the ML(C) phase, and shares both the Rv0081 and CsoR TF binding sites proximal to its start site. While Rv0081 shows no differential expression between the two isolates in any condition, the *csoR* gene is expressed at a significantly lower level in S5527 under multiple conditions. This result implicates CsoR as the regulator of these three sRNA and the driver of their differential expression between the two isolates.

The remaining sRNA down regulated in S5527 (MTS1082 and MTS1338) do not appear to have any notable common regulatory links. MTS1082 does not appear to share any of the known reported TFs except for Rv0081 which is not differentially expressed, but as discussed it shares a response to isoniazid with G2 indicating a probable regulatory link facilitated by a mechanism that is currently not apparent. This sRNA may however be regulated by Lsr2, which shows slightly decreased expression in S5527, but not enough to be considered significant.

The sRNA up regulated in S5527 both share the Lsr2 and Rv0081 TFs except for ASdes. The regulation of ASdes may be different because it is a *cis*-encoded sRNA as opposed to the other two which are both found in the 5' and 3' untranslated region between two genes. Both mpr6 and mcr11 were differentially expressed during the ML(C) phase, indicating that the regulatory mechanism that is altered in S5527 for these sRNA is active at that time. As the expression of *lsr2* is only slightly decreased in S5527, and *Rv0081* expression does not appear to differ between the two strains, the cause of the difference in expression is not immediately clear. The location of *mpr6* between *sigE* and *rseA* on the same strand is notable, as it potentially links the expression of this sRNA to that of SigH, which is known to be involved in the response to oxidative stress and heat shock [171]. As *sigE* is known to be regulated by

(a) mcr11 **(b)** mpr6



(c) ASdes

Figure 4.6: Highest ranked cluster of ontologies for the targets of the sRNA with increased expression in isolate S5527 as reported by IntaRNA.

SigH [172], and *sigH* showed significantly lower expression in S5527, the organisation of these genes suggests that SigE, SigH, and Lsr2 are connected to the expression of these sRNA but the extent to which they contribute to the observed differences in expression between the two isolates remains unclear and warrants further investigation.

Lsr2 is an iron-regulated global transcriptional regulator required for adaptation, specifically to changing oxygen levels and virulence that is also found to be up-regulated by SigE, Rv2827c, and Rv0232[173]. Interestingly another sRNA mpr17 is found between *lsr2* and *clpC* on the opposite strand. Though the expression of mpr17 was seen is some samples, it was absent in the replicates of others (Table 4.4). As a result of the high level of variance the expression profile this sRNA could not be confidently ascertained. A commonality between the regulators Rv0081 and Lsr2 (Rv3597c) is that both display binding to whiB3, an iron-sulfur (Fe-S) cluster containing regulatory element that responds to the dormancy signals NO and O_2 [174]. And while Rv0081 is not differentially expressed, this once again draws attention to pathways involved in dormancy and oxygen response, and is a core part of a regulatory subnetwork described by Galagan *et. al.,* [128] that links hypoxia, lipid metabolism, and protein degradation, and that contains many of the regulatory elements high-lighted in our results including Lsr2, WhiB3, and SigE.

4.4 Conclusions

Understanding the relationship between regulators in a biological system and their targets is a complex task, particularly when the targets are uncertain and we are unable to distinguish primary effects as a result of direct interaction with the regulatory element from downstream effects. With the addition of non-coding RNA to the regulatory system, this complexity grows. The starting point to unraveling this problem is the identification of all the players involved.

With no established method for the automated identification of novel sRNA using small RNA-sequencing, we developed a tool that could identify novel sRNA in non-model isolates including W-148. We identified 152 potential novel sRNA in the W-148 strains (expressed in the two isolates) including 56 that represented highly probable sRNA for this isolate, 9 of

which were identified as homologues of those described in H37Rv. As many of these sRNA were found to be differentially expressed between the two isolates, this demonstrates that the currently identified sRNA may only be the tip of the iceberg, though these results will need to undergo laboratory confirmation to assess the accuracy of the method.

In order to understand the roles of the different known sRNA in MTB, we identified the mRNA that they are interacting with using *in silico* methods, and generated expression profiles for both the sRNA and the targets at different growth phases. These profiles were then used to better understand the differential expression of sRNA between the two isolates S507 and S5527.

As part of these profiles we identified which sRNA are expressed chiefly under certain conditions, including the sRNAs ncrMT3949 and B55 that are involved in the response to H_2O_2, and thus adaption to the macrophage environment. Many of these sRNA are already associated with known growth stages and environmental stimuli, and were also found to be expressed under the same conditions in our results. We identified 8 sRNA differentially expressed between the two isolates (with a log2FC greater than 0.5), and noted the probable involvement of a number of regulatory elements involved in dormancy and oxidative stress. One of the probable drivers of the differences in sRNA expression profiles between the two isolates was the TF gene *csoR* that is also differentially expressed between the two isolates and predicted to regulate 3 of the down regulated sRNA.

Many of the regulators linked to these sRNA (SigE, SigF, DosR, WhiB3, Lsr2, Rv0081) are described as being part of a regulatory interaction network that links hypoxia, lipid metabolism, and protein degradation [128]. Together with the lowered expression of sRNA including G2 and MTS2975 which are normally highly expressed during the exponential phase and lowered expression of MTS1338 which is linked to dormancy also suggested that the more virulent S5527 isolate has a tempered dormancy response, existing in a generally more active state.

While the results of the small RNA sequencing has already identified players in the regulatory differences between the two strains, they are only one set of players in the regulatory network. Additionally, it is unknown what level of sRNA differential expression is required to have an impact on their targets. A 2 fold change in sRNA abundance may translate to a

10 fold change in the abundance of their targets protein abundance or have little to no effect at all. In the next chapter these results will be combined with gene expression data and variant data to attempt to identify the root cause of the differences in expression observed at the different levels, and assess whether the changes in abundance observed have had a noticeable effect.

"I think perhaps the most important problem is that we are trying to understand the fundamental workings of the universe via a language devised for telling one another where the best fruit is."

Terry Pratchett

5

Differential expression of genes in two closely related *Mycobacterium tuberculosis* isolates

5.1 Introduction

The isolation of isolates S5527 and S507 in the Western Cape provided an interesting opportunity to study the systems that determine how virulent different strains of MTB differ to their more benign counterparts. The isolates studied are closely related, with few variants observed between them, none of which were found to affect known virulence drivers in MTB. In order to find the basis of the altered virulence observed between the two isolates, a more system wide approach had to be adopted. With the availability of data from the genome, transcriptome (coding and non-coding), and regulatory networks for the two isolates, the final integration and interpretation of the systems may take place. The development of a

method to incorporate annotation data across isolates using genome graphs as described in chapter 2 proved invaluable in overcoming some of the technical and biological obstacles that come with isolates that differ significantly at a genomic level.

In this chapter we discus the genes found to be differentially expressed between the two isolates, and place them within the context of the results from the previous chapters in order to find a common theme that permeates the dataset.

The first thread of this common theme appeared as a group of genes involved in the production of a molybdenum cofactor, which were found to be differentially expressed between the two isolates for all the growth conditions considered. This in turn led to the investigation of genes that are involved in copper response in MTB, and finally to the hypothesis that the difference in the isolates virulence phenotypes is most likely the result of their different ability to respond to phagosomal copper overload, a mechanism employed by macrophages to kill MTB within its phagosomes.

5.2 Materials and methods

5.2.1 Sample collection, experimental design, and sequencing

Sample collection and growth was conducted by members of Rob Warren's research group based at the University of Stellenbosch, with the sequencing being done by Jonathan Featherson at the Agricultural Research Council in Pretoria. The conditions compared and the treatments applied were identical to those selected for the sRNA sequencing in the previous chapter (Section 4.2.3). For the removal of rRNA from the total RNA samples the Truseq stranded mRNA library preparation kit (RS-122-2101) with the Bacterial Ribozero kit (MRZMB126) was used. The samples were sequenced on an Illumina HiSeq 2500 using version 4 SBS chemistry (2x125bp), with approximately 10 million reads per sample loaded into the lane. The sequencing data was then sent to our labs for analysis.

5.2.2 Read filtering and trimming

Initial read quality was assessed using fastQC [153]. Adapter trimming and removal of low quality reads was done using trim_galore (0.4.0) with the minimum length cutoff 20bp, the quality phred score cutoff set to 20 (For Quality encoding type ASCII+33), and the maximum error rate set to 0.1. Following this, the file locations and sample details for the reads were amalgamated into a sample file and passed to the Cell pipeline. For the read alignment within Cell, BWA-MEM was used, while Cuffdiff was used for differential expression analysis.

5.2.3 Confirming the samples are correctly labeled

In order to confirm that the RNA sequencing samples represented the same isolates as the whole genome sequencing data used for variant calling and genome assembly, RNA sequencing reads from two of the samples were aligned to the genome of isolate W-148 using BWA-MEM, and SNP calling was conducted with GATK. As the samples did cluster by isolate, the main concern was that the labelling of isolates could have been reversed, one representative sample for each isolate was selected, sample19 for isolate S5527 and sample7 for S507. The variants were then compared to the whole genome sequencing variants from S507 and S5527 when mapped to isolate W-148.

5.2.4 Alignment to the reference genomes

The Cell pipeline (Discussed in Chapter 3) was used to analyse the data and provide downstream integration with existing knowledge to contextualise the results. This pipeline manages the read alignment tools, quality control, differential expression analysis and file organisation. To determine which reference genome provided the best read mapping the samples were aligned to H37Rv, CDC1551, W-148, and a pan-transcriptome generated by the GenGraph toolkit discussed in Chapter 2. The pan-transcriptome generated contained 4 genomes (H37Rv, CDC1551, W-148, CCDC5180) and used a sequence homology cutoff of 95% shared identity. GenGraph was also used to create a homology matrix to allow mapping of annotations across species.

5.2.5 Differential expression analysis using CuffDiff

Differential expression analysis was then conducted using the tool Cuffdiff (v2.2.1), as part of the Cell pipeline. For library normalisation, a geometric means method was used where FPKMs and fragment counts are scaled via the median of the geometric means of fragment counts across all libraries, and a pooled cross-replicate dispersion estimation method. Additionally, the CuffDiff bias detection and correction algorithm was used, as well as the correction algorithm for reads mapping to multiple locations in the genome. As there were rRNA and tRNA sequences still present in the samples after the library preparation, a mask file containing these reads was created and used by Cuffdiff to exclude these reads from transcript abundance estimates. Cuffdiff calculates both p-values (of the reported test statistic) and q-values (a false discovery rate adjusted p-value of the test statistic), with genes that had a q-value less than 0.05 considered differentially expressed. Tertiary analysis and visualization was done using the Cummberbund and ggplot2 packages in R, and included principle component analysis (PCA) to confirm samples were clustering correctly by condition and strain, generation of heat maps and plotting of individually differentially expressed genes.

5.2.6 Variant calling vs the W-148 genome

The Cell pipeline was used to conduct variant calling using whole genome sequencing reads from SAWC5527 and SAWC507 mapped against the W-148 genome. The variant files were then filtered using vcftools (0.1.12b) [175] with additional filtration carried out by the Cell pipeline.

5.2.7 *In silico* detection of transcription factor binding sites using FIMO

In order to detect the binding of TFs in regions that differ between the two strains, we used the tool Find Individual Motif Occurrences (FIMO) [176], and a list of TF binding motifs from a genome-wide TF binding study conducted by Minch *et. al.* [177]. In the case of the Moa3 operon, a region 200bp upstream from the start of the *moaA3* gene was scanned for known motifs.

5.2.8 Integrating the results into networks: Cell

As part of the Cell pipeline, the tool Holmes integrates data from various sources and amalgamates it into an interaction network from which various outputs can be produced. This includes linking variants to genes that are found differentially expressed between the isolates, and targets of predicted sRNA that were found in Chapter 4. The results from the sRNA analysis were integrated into the network, first indicating which differentially expressed genes had an associated sRNA that was predicted to regulate it, then assessing whether that sRNA was likewise differentially expressed. The GenGraph homology matrix allowed mapping of annotations between isolates, allowing us to take advantage of the rich annotation of H37Rv, while mapping to the more closely related isolate W-148.

The results were visualised using Cytoscape, or exported as tables. Possible causes for altered expression could then be identified by selecting the nodes representing the differentially expressed genes, then selecting the neighbours of those nodes that may include connected nodes that represent proximal variants. The set of selected nodes could then be extracted to create a new network which highlights the link between variants and differentially expressed genes, and other genes in the region which may represent an operon.

5.3 Results and discussion

5.3.1 Quality control: Read trimming and filtering of RNA sequencing reads

Normalisation is the process by which technical bias is removed while introducing as little noise as possible. Using the trim_galore (0.4.0) wrapper script for cutadapt-1.8.1, Illumina adapter reads were successfully detected and removed while trimming and filtration of the reads resulted in an overall improvement in the read quality (Figure 7.13). On average 1.81% of reads were removed for either poor quality or length bellow the minimum threshold. The use of certain samples (samples 2, 11, 14, 23) for differential expression analysis was found to result in all genes reporting infinite FPKM values across all genes. The reason for this is currently still unknown and these samples were excluded.

5.3.2 The effects of using different reference genomes

One of the most important decisions when conducting read alignment is the selection of the reference genome. This choice has both immediate and downstream consequences on the analysis, and should be given adequate thought. Within the MTB complex the H37Rv genome is by far the most well annotated and complete genome. Unfortunately, it also has fewer genes than most of the other genomes (3,906 in H37Rv and 4,075 in W-148). The two isolates that are the focus of this project are most closely related to the W-148 isolate, a member of the Beijing cluster, which has a genome that is larger in size by 4,133bp and contains 134 additional genes. As a consequence, alignment of the filtered reads to the W-148 genome produced better results for all samples (Table 7.4) with on average 0.21% more reads mapping to the W-148 isolate. This result also confirms that the isolates' sequences are more closely related to the W/Beijing genotype than to the H37Rv strain. The additional benefit in the use of the W-148 genome is that we are better able to integrate variant data downstream, and there is less likely to be structural changes that may obscure important details such as the co-expression of genes within the same operon or the position of regulatory regions.

5.3.3 Confirming the correct samples

Alignment of the RNA sequencing reads and whole genome sequencing reads to the W-148 reference and variant calling confirmed that the isolates were the same and the labels were not reversed between the two datatypes. Sample 7, the isolate S507 representative shared the greatest number of variants when aligned to W-148 with the sequencing reads for the S507 isolate and likewise for sample 19, the S5527 representative (Table 5.1). The samples were selected randomly as representatives for the isolates.

5.3.4 Quality control: Final settings and results

Quality control for differential expression analysis involves assessing whether there is any observable bias in the data as a result of batch effects, sample preparation, sequencing technology, or of the normalisation.

Table 5.1: Comparison of shared variants between samples and isolates. Shared variants occur when both samples (whole genome sequencing and RNA sequencing) identify a variant at the same position in the W-148 reference genome.

Sample	Number of shared variants
S5527 and sample19	11
S507 and sample19	0
S5527 and sample7	1
S507 and sample7	12

The density profiles and dispersion plots of the samples after normalisation were fairly similar and there were no samples that appeared as significant outliers (Figure 7.7). When looking at the clustering of the samples in the isolate versus isolate comparison, we see that often the random variance within the samples is enough to obscure the groupings (Figure 7.8 and 7.10) while in the condition vs condition analysis, the samples correctly separate by condition in most of the cases (Figure 7.9a). The exceptions to this are in the cases of the hydrogen peroxide treatment versus the control, and the early log phase growth versus middle log phase growth comparison for isolate S502, where one of the early log phase growth samples appeared to separate from the rest of the samples (Figure 7.9b), particularly when considering the PCA plot (Figure 7.11a). In other comparisons the abnormal S507 sample did cluster correctly (Figure 7.8 and 7.11b) and was thus retained as is was not enough to justify the loss of statistical power that would result from it's exclusion.

5.3.5 Differentially expressed genes between conditions

Understanding the changes in gene expression at different growth stages of MTB is needed to better contextualise the differential expression of genes between the two isolates. Considering a wide range of conditions is also important as the genes responsible for the difference in phenotype may only be differentially expressed under certain conditions, such as within macrophages or during exponential growth.

When comparing the isolates the number of genes differentially expressed under different conditions varied greatly, with only two genes found that were uniquely differentially

expressed during ML(C) while ML(T) had the most at 34 (Figure 5.1). This could indicate that the isolates are relatively similar during normal exponential growth, but react quite differently to oxidative stress / treatment with H_2O_2.

In order to understand the context of the genes found differentially expressed between the two isolates, the conditions under which they are normally expressed must first be understood. We first compare the expression of genes between different conditions, identifying genes that are more active during exponential growth, those that trigger the dormancy response, and those expressed in response to treatment with H_2O_2.

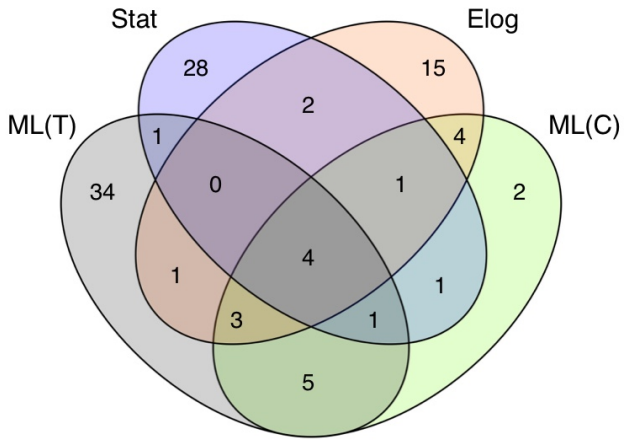


Figure 5.1: A Venn diagram showing the number of differentially expressed genes between the two isolates under the four experimental conditions. This was created by taking the sets of differentially expressed genes between the two isolates for each of the conditions and determining the overlap between the sets when comparing conditions.

Early logarithmic growth vs middle phase logarithmic growth

When comparing Elog growth to ML(C) growth 97 genes were differentially expressed in isolate S507 and 149 genes were differentially expressed in isolate S5527. In total 188 unique genes were differentially expressed, of which 58 were in both the isolate S507 and S5527 datasets, 32 of which were up regulated during ML(C) growth and 26 were down regulated. In the shared set of genes, Panther identified 3 ontology terms that were significantly enriched, notably those involving gene regulation and expression (Table 5.2) with protein methylation and regulation of transcription from RNA polymerase II showing the greatest

level of enrichment.

Table 5.2: Gene set enrichment analysis results produced by Panther: Elog vs ML(C). The column "found in MTB" refers to the number of proteins found in H37Rv with this ontology. "In data set" is the number of proteins in the query dataset that have this ontology. The "expected" value is the number of proteins with the given ontology expected when randomly selecting proteins from the total dataset. The "fold enrichment" refers to the fold difference in the number observed proteins with the given ontology in the current dataset. The "+/-" designation refers to whether this fold change is positive or negative. The p-vale is the significance of this fold change.

Biological Process	Found in MTB	In dataset	Expected	Fold Enrichment	x/-	P value
Protein methylation (GO:0006479)	13	3	0.08	36.10	+	9.00E-03
Regulation of transcription from RNA polymerase II promoter (GO:0006357)	22	3	0.14	21.33	+	4.19E-02
Transcription, DNA-dependent (GO:0006351)	110	6	0.70	8.53	+	7.66E-03

Some of the notable genes up regulated during ML(C) growth belong to known pathways including the leucine biosynthesis genes *leuD* and *leuC*, and the pyrimidine metabolism gene *mmsA* (Methylmalonate-semialdehyde). The leucine biosynthetic pathway is essential for the growth of *Mycobacterium tuberculosis* [178] while the *mmsA* gene is involved in valine and pymiridine metabolism and binds fatty acyl-CoA [155]).

Within the shared set of genes down regulated in ML(C) growth, three putative intergrase / transposase producing genes homologous to Rv1765A, three membrane proteins homologous to Rv1216c, and four transcriptional regulators were identified. Intergrase genes originate from bacteriophages, and were most likely integrated into the genome during infection and are now consequently expressed along with their proximal genes. Unfortunately the function of the genes homologous to Rv1216c is currently unknown.

Some of the transcriptional regulators include 3 orthologues of HTH-type transcriptional regulator PrpC (From provided annotations in the W-148 annotation file, *prpR* ortholog) and the transcriptional regulator WhiB1, that were both highly expressed during the Elog phase. The transcriptional regulator PrpC is involved in regulating pathways for the utilisation of fatty acids from the host as carbon sources [179] while WhiB1 is a well studied regulator that has been shown to repress the essential chaperone protein GroEL2 [180, 178] and is said to contain a *NO* sensitive [4Fe-4S] cluster [181]. The GroEL2 protein in turn is a potent

inducer of cytokine synthesis, and is an important virulence factor in tuberculosis for this ability to modulate host immune response [182].

The transition from early to middle log phase growth appears to be characterised by these changes in expression of regulatory genes and newly available food sources (as in the case of the *leuD*, *leuC*, and *mmsA* genes) and suppression of dormancy related genes such as WhiB1. The function and significance of many of the remaining genes is unknown and the expression of the intergrase genes is most likely to be simply a result of their integration close to regulatory regions involved in the transition between these two phases.

Middle phase logarithmic growth vs stationary phase

When comparing the ML(C) growth to the Stat phase, 151 genes were differentially expressed in isolate S507, 345 in isolate S5527, and a total of 392 unique genes were differentially expressed in the combined set. Of the 104 genes differentially expressed in both isolates, 68 genes showed an increase in expression during the middle phase and 36 showed a decrease in expression. In this set of shared differentially expressed genes, there is an enrichment of ontological terms related to stress response and metabolic reprioritisation (Table 5.3).

Notable pathways that contained genes showing increased expression in the Stat phase cultures include some involved in the metabolism of pyrimidine and pyruvate (pathway P02771 and P02772 respectively), and the degradation of aminobutyrate.

Genes in pathways involving leucine and ATP synthesis (P02749 and P02721) showed decreased levels of expression in Stat phase cultures, and included a homologue of the down regulated ATP synthesis gene *atpG* (Rv1309) known to be involved in the MTB response to starvation [183]. Many of the other ATP synthesis genes showed similarly lowered expression during the Stat phase in our results, and were also found to be differentially expressed in a study by Betts *et. al.,* [183] in the response of MTB to nutrient starvation.

Early logarithmic growth vs stationary phase

When comparing the Elog growth to the Stat phase 691 genes were differentially expressed in isolate S507, 762 in isolate S5527, and a total of 979 unique genes were differentially expressed in the combined set. Of the 474 genes differentially expressed in both isolates, 265

Table 5.3: Gene set enrichment analysis results from Panther: ML(C) vs Stat. The column "found in MTB" refers to the number of proteins found in H37Rv with this ontology. "In data set" is the number of proteins in the query dataset that have this ontology. The "expected" value is the number of proteins with the given ontology expected when randomly selecting proteins from the total dataset. The "fold enrichment" refers to the fold difference in the number observed proteins with the given ontology in the current dataset. The "+/-" designation refers to whether this fold change is positive or negative. The p-vale is the significance of this fold change.

Biological Process	Found in MTB	In dataset	Expected	x/-	Fold Enrichment	P value
protein folding (GO:0006457)	20	6	.26	+	23.23	3.05E-05
response to stress (GO:0006950)	67	6	.87	+	6.94	2.65E-02
response to stimulus (GO:0050896)	123	8	1.59	+	5.04	2.19E-02
cellular amino acid metabolic process (GO:0006520)	398	16	5.14	+	3.11	5.48E-03
primary metabolic process (GO:0044238)	1746	45	22.54	+	2.00	6.34E-05
metabolic process (GO:0008152)	1980	51	25.57	+	1.99	5.33E-06

genes showed an increase in expression during the middle phase and 209 showed a decrease in expression.

Analysis of overrepresented gene ontologies of the genes differentially expressed in both isolates using Panther showed a significant enrichment for genes involved in fatty acid and lipid metabolism indicating shifts in energy source utilisation (Table 5.4). Other ontologies were too general for speculation though the porphyrin-containing compound metabolic process is notable.

There was an increase in expression of genes belonging to pathways for the metabolism of pyrimidine, pyruvate and vitamin D in the Stat phase of growth relative to the Elog phase. The significance of vitamin D synthesis seems to be in its role as a regulator of lipid metabolism in MTB [184] and thus modulating the use of different energy sources. The gene in question *fprB* (Rv0886) is a probable NADPH:adrenodoxin oxidoreductase, and is only found to have lowered expression in the Elog phase. Conversely, though *bkdA* (Rv2497c) and Rv2913c (linked to pyruvate and pyrimidine metabolism respectively) also show increased expression during Stat phase, it is because of their very high expression specific to the Stat phase.

Genes involved in ATP synthesis and methionine biosynthesis showed decreased levels of

expression in the stationary phase, once again including the ATP synthase genes, and the *metB* gene that is involved in methionine biosynthesis.

The comparison of the gene profiles of bacterial cells during Elog phase growth and Stat phase contained the largest number of differentially expressed genes and effected pathways. The Elog represents a time where cells are transitioning to a highly replicating state, where as Stat phase represents a time of reduced resources where replication and cell death as a result of starvation are at an equilibrium and are thus expected to have significantly different profiles. The majority of pathways affected are for metabolic pathways and energy production, as would be expected when comparing the gene expression profiles at these two growth phases defined as times of an increase and depletion of available energy sources.

Differentially expressed genes between treated and untreated samples

In response to treatment with H_2O_2, the MTB samples showed only a few genes with significant differential expression. Isolate S506 showed only 13 genes differentially expressed, where as isolate S5527 showed 38, with 6 genes being found differentially expressed in both of those datasets (Table 5.5).

A gene annotated as a heat-shock protein (Hsp20 / *hsp*) similar to Rv0251c in H37Rv was identified as significantly differentially expressed between the two conditions. It is thought to be involved in the initiation step of translation at high temperature and possibly as a molecular chaperone, and has been found to bind to the 30S ribosomal subunit, and is also induced by oxygen after a time of anaerobic growth [185] and part of a set of general damage-associated response genes found to be up-regulated during prolonged exposure to intracellular stress [186]. Interestingly, the gene is significantly increased in response to H_2O_2 in S507, and significantly decreased in response to H_2O_2 in S5527. The expression of this gene has been linked to several regulatory elements including SigE, HspR, HrcA, and PhoP [187, 185, 155] specifically it is positively regulated by SigE and negatively by HspR.

The remaining genes were all found to have lowered expression in the sample treated with H_2O_2. These include a lysine ϵ-aminotransferase (*lat*) gene (Rv3290c / TBPG_RS03370), an alarmone that in the experiments by Duan *et al.,* [188] was observed to be up-regulated during hypoxia and nutrient starvation. This was partially contrary to our results. On

Table 5.4: Gene set enrichment analysis results from Panther: Elog vs Stat. The column "found in MTB" refers to the number of proteins found in H37Rv with this ontology. "In data set" is the number of proteins in the query dataset that have this ontology. The "expected" value is the number of proteins with the given ontology expected when randomly selecting proteins from the total dataset. The "fold enrichment" refers to the fold difference in the number observed proteins with the given ontology in the current dataset. The "+/-" designation refers to whether this fold change is positive or negative. The p-vale is the significance of this fold change.

Biological Process	Found in MTB	In dataset	Expected	x/-	Fold Enrichment	P-value
Homeostatic process (GO:0042592)	118	30	6.61	+	4.54	1.46E-09
Biological regulation (GO:0065007)	135	30	7.56	+	3.97	3.49E-08
Porphyrin-containing compound metabolic process (GO:0006778)	54	11	3.02	+	3.64	3.09E-02
Regulation of nucleobase-containing compound metabolic process (GO:0019219)	76	13	4.26	+	3.05	4.74E-02
Catabolic process (GO:0009056)	223	37	12.49	+	2.96	7.90E-07
Fatty acid beta-oxidation (GO:0006635)	163	25	9.13	+	2.74	8.67E-04
Fatty acid metabolic process (GO:0006631)	324	40	18.14	+	2.20	3.82E-04
Lipid metabolic process (GO:0006629)	515	58	28.84	+	2.01	4.30E-05
Nitrogen compound metabolic process (GO:0006807)	486	53	27.21	+	1.95	3.63E-04
Cellular amino acid metabolic process (GO:0006520)	398	42	22.29	+	1.88	8.29E-03
Cellular process (GO:0009987)	879	87	49.22	+	1.77	1.16E-05
Primary metabolic process (GO:0044238)	1746	170	97.77	+	1.74	6.42E-13
Metabolic process (GO:0008152)	1980	181	110.87	+	1.63	1.97E-11

the one hand our data showed that the expression of *lat* was down-regulated in the H_2O_2 treated sample when compared to the control, but during the Stat phase we observed the highest expression levels of the *lat* gene, in accordance with the reported nutrient starvation response. Others include a methyltransferase (Rv1405c / TBPG_RS06275) that along with *whiB7* and *hsp*, is linked to macrophage invasion [186], and has been shown to respond to stress, including acid shock, where it showed an increase in expression [189, 190]. The expression of *whib7* in this context is regulated by PhoPR [187], and was found to contain a non-synonymous mutation in the S507 isolate. This mutations did not appear to effect the

expression in the S507 isolate in any significant way.

The two ABC transporter genes are most likely co-regulated and represent the nucleotide binding domain (Rv1687c) and the membrane-spanning domain (Rv1686c) of an ABC transporter. The substrate being transported is uncertain, but the two subunits are reportedly similar to NosF in *Pseudomonas stutzeri* which is involved in copper transport and processing [191, 192].

The final gene is a transcriptional regulator Rv0678, that potentially regulates the mmpS5 and mmpL5 genes [31], both of which show decreased expression as a result of treatment with H_2O_2 in both isolates, though the change is not significant. These two genes (mmpS5 and mmpL5) are membrane proteins that are part of an efflux system that is related to resistance to azole [31]. The link between this gene and treatment with H_2O_2 remains unclear, but may be understood once the regulators of Rv0678 are known.

The observed differences in the responses of the two isolates to treatment with H_2O_2 may be the leading contributor to the increased virulence seen in S5527 as S5527 showed a greater response in terms of the number of differentially expressed genes. This was also seen in the expression of *hsp*, which will be discussed further when comparing the expression levels of genes between the strains. Although genes linked to stress were detected, it was interesting to note the lack of any of the WhiB-like or DosR / DevR genes that are generally associated with oxidative stress and treatment with H_2O_2 [174, 193, 194], though the expression of DosR / DevR was lower in the H_2O_2 treated samples for both isolates but not enough to be significant. It is possible that other genes involved in this response were likewise underpowered statistically, or that even small fold changes in their expression are sufficient to mount a response. It has been shown in previous studies that differing concentrations of H_2O_2 and length of exposure lead to different genes being differentially expressed. For example the expression of recA was induced over 4 fold by 5 and 10 mM treatments of H_2O_2 but only 0.8 and 0.5-fold at 50 and 200 mM. This was also seen for radA which had vastly lowered fold changes in response to treatment [156]. These results make it apparent that a single time point and concentration may not be enough to profile the full response of MTB to H_2O_2, and that the response is possibly made up of multiple phases of expression.

Table 5.5: Significantly differentially expressed genes found when comparing samples treated with H_2O_2 and those without.

Gene name in W-148	H37Rv homologues	Log2(FC) S507	Log2(FC) S5527	Function	Genbank
TBPG_RS01325	Rv0251c	-1.27 (S507)	1.02 (S5527)	Heat-shock protein *hsp* (hsp20, hrpA, acr2)	WP_003900838.1
TBPG_RS03370	Rv3290c	-1.13 (S507)	-1.44 (S5527)	L-lysine-epsilon aminotransferase *lat*	WP_003900004.1
TBPG_RS06275	Rv1405c	-1.28 (S507)	-1.05 (S5527)	Methyltransferase	WP_003407297.1
TBPG_RS07665	Rv1686c	-1.13 (S507)	-1.45 (S5527)	ABC transporter permease	WP_003898974.1
TBPG_RS07670	Rv1687c	-1.35 (S507)	-1.32 (S5527)	Multidrug ABC transporter ATP-binding protein	WP_003898975.1
TBPG_RS17170	Rv0678	-1.24 (S507)	-1.01 (S5527)	Transcriptional regulator	WP_003403442.1

5.3.6 Differentially expressed genes between isolates

In order to thoroughly investigate the cause of altered virulence between the two strains, differential expression of genes at different growth phases must be considered, as the phenotype may be as a result of a gene whose expression is only detectable in one of the phases. This is true of our results, as of the 102 genes found differentially expressed between the isolates, many are only differentially expressed under a particular condition and only 4 were found differentially expressed under all conditions. When comparing gene expression between the isolates on a per condition basis, the number of differentially expressed genes ranges from 21 to 49 (Table 5.6). The most notable genes was a cluster of gene involved in molybdenum cofactor (MoCo) biosynthesis. This cluster of genes (gene636-gene639) was found to have decreased expression in S5527 under all conditions (Except for gene637 in ML(T)) and is discussed in detail in a later section.

Table 5.6: Summary of the number of genes found to be differentially expressed between the two isolates under different conditions, with isolate W-148 used as the reference genome.

Condition	Total	Up regulated in S5527	Down regulated in S5527
Early log	30	23	7
Mid log control	21	16	5
Mid log treated	49	17	32
Stationary	38	37	1

Early log genes of interest

During this growth phase, 6 clusters of genes were found, 5 of which had all genes down-regulated in S5527 while 1 had all genes up-regulated (Table 7.5). There were also 3 transcriptional regulators differentially expressed, and 3 variants associated with the differentially expressed genes including one 834bp from one of the clusters (gene3697, gene3700 and gene3701) that codes for PE-PGRS family proteins. Other differentially expressed genes such as gene774 and gene3102 have SNPs 54bp, 2102bp away respectively. These genes are likely to be involved in the altered phenotype if the mutation effects their expression. The first is gene744, an acetyl-CoA carboxylase biotin carboxyl that showed a decrease in expression in the more virulent isolate S5527. The second is gene3102, a FmdB family transcriptional regulator that has been identified as one of the genes involved in regulatory mechanisms in response to nitrogen limitation in *Mycobacterium smegmatis* [195, 196].

Another interesting feature of this dataset is the three genes linked to cations, specifically cadmium. This includes gene1850 (*cmtR* / Rv1994c) which is a cadmium-lead-sensing ArsR-SmtB Repressor [197] that showed a decrease in expression in isolate S5527, a cation diffusion facilitator (CDF) gene1887 (Rv2025c) that has increased expression in S5527, and a cadmium inducible protein gene2547 (CadI / Rv2641) that has lower expression in S5527. These genes point to a difference in either cation sequestration, sensing, or concentration control between the isolates.

Middle log (control) genes of interest

When comparing the gene expression of the isolates during the ML(C) phase, apart from the cluster of Moa genes, three observations stand out (Table 7.6). Firstly we see that the expression of the *hsp* (Hsp20) gene previously mentioned when comparing gene expression between ML(C) and ML(T) is significantly decreased in isolate S5527. This result is mirrored by the aforementioned increase in expression during ML(T) (Table 7.7). The second is the EsxB (Rv3874), a secreted virulence factor that is required for pathogenesis in *Staphylococcus aureus* and MTB [198] that showed increased levels of expression during both ML(C) and ML(T) conditions in isolate S5527. This gene is possibly repressed by a heat shock

protein transcriptional repressor HrcA (Rv2374c) which, though not found to be differentially expressed, is also linked to the heat shock response, same as *hsp*. Proteomics done by the Stellenbosch research group identified the EsxA proteins as being under-represented in the hyper-virulent strain [11]. While we observed esxB being expressed at a higher level in the hyper-virulent strain, esxH (during Stat and ML(C)) and esxI (during ML(T)) was expressed at lower levels in the hyper-virulent strain. Other ESAT-6 like genes including esxA were not differentially expressed when comparing the two isolates at other growth phases, and no sRNA were identified that target the mRNA of this gene. No reason for this discrepancy is immediately apparent from our data. Finally, both gene3254 (TBPG_RS16270) and gene3253 (TBPG_RS16265) showed decreased expression in S5527 during both Elog and ML(C). TBPG_RS16265 is annotated as a hypothetical protein, but the sequence matches that of a lipoprotein LpqS that forms part of a regulon specific to virulent mycobacterial species and controlled by a copper sensing repressors, RicR (Rv0190 / TBPG_RS01010) [199, 200]. This copper sensing repressor gene *ricR* is a paralogue of the copper metalloregulatory repressor gene *csoR* and was significantly downregulated in S5527 during the Elog phase, and also showed decreased expression during the ML(C) phase. RicR is also part of a copper sensitive operon [201]. This condition specific expression (lower in S5527 during Elog and ML(C)) is also seen for *mymT* (Rv0186A), a metallothionein that protects the cell from copper toxicity that is part of this same gene cluster as *ricR*, and is reported to be up-regulated by copper, cadmium, and compounds that generate nitric oxide or superoxide [202]. Another gene in this region is *ilvD* (Rv0189c) which encodes a dihydroxy-acid dehydratase. While this gene is not differentially expressed, it contains an iron-sulfur cluster [4Fe-4S] that is attacked by copper preventing cluster assembly in *E. coli* [203], which is most likely why it is in a region containing a copper protective metallothionein *mymT* gene. These genes all add to the number of cation sensing genes seen to have altered expression between the two isolates, though bringing particular focus to copper.

Middle log (treated) genes of interest

Though this condition had the greatest number of genes differentially expressed between the two isolates, a large number of them were hypothetical proteins. What this does reveal, is

that the two strains respond quite differently to oxidative stress as induced by H_2O_2 (Table 7.7). The only genes to have a proximal SNP in this dataset are gene2521 (a homologue of Rv2616) and gene2522 (a homologue of Rv2617c, a transmembrane protein), both were found to be up-regulated in the hyper-virulent strain. The cause of the altered expression is unknown, as the SNP is over 1,000bp away, and is thus unlikely to have been the reason for the change in expression, and the genes predicted to regulate them do not show changes in expression. Notable genes found in this dataset with links to pathogenicity were a type B diterpene cyclase (gene3563, WP_003417905.1, TBPG_RS17815) and a diterpene synthase (gene3564, WP_003417908.1, TBPG_RS17820) that were both found to be down-regulated in the hyper-virulent strain. Diterpenes have been studied in the past for their potential role in promoting phagolysosome maturation arrest [204].

Stationary phase genes of interest

The stationary phase contains the largest set of adjacent differentially expressed genes (gene3558 - gene3565) found to be down-regulated in the hyper-virulent strain (Table 7.8). These include the previously mentioned diterpene synthesis related genes found in the H_2O_2 treated samples. The other genes appear to be likewise involved in metabolism, including trehalose-phosphate phosphatase that is located in the cell wall and induces humoral and cellular immune responses in the host [205]. Though considered to be significantly differentially expressed, many of these genes have very low expression compared to other genes in the region (Figure 5.2) specifically gene3557, gene3562, and gene3563 have a higher read coverage while the remaining genes of the cluster show low levels of expression. Despite this, these low expression genes have little inter replicate variance, and are still significantly differentially expressed.

5.3.7 The molybdenum cofactor genes and amalgamation

The only genes consistently differentially expressed under every condition is a cluster of genes belonging to the same operon (Figure 5.3) that are involved in molybdenum cofactor (MoCo) biosynthesis. Molybdenum (Mo) is an essential micro-element for nearly all organ-

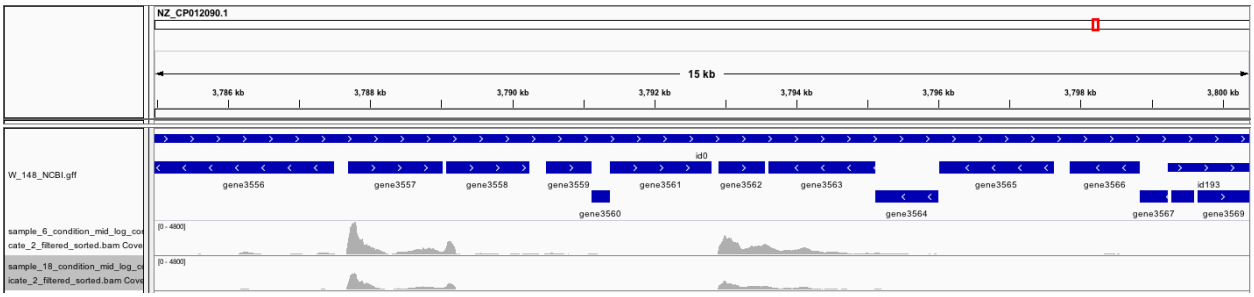


Figure 5.2: The varied read coverage of the differentially expressed gene3558 - gene3565 region as found in W-148.

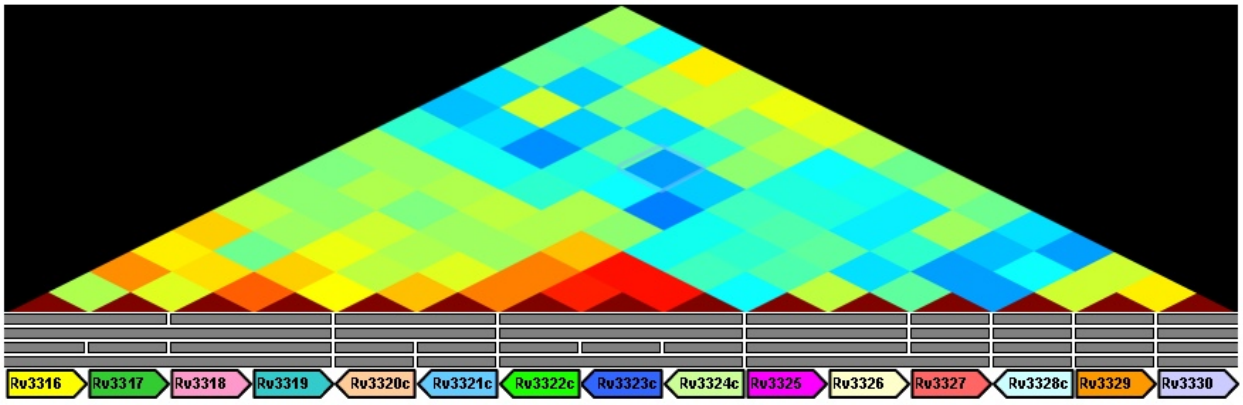


Figure 5.3: The Moa3 operon in H37Rv containing the differentially expressed genes (having a q-value less than 0.05) generated by the TBDB operon browser.

isms including MTB, where is it found in these Mo containing cofactors which are in turn used by enzymes including nitrate reductase, carbon monoxide dehydrogenase (CO-DH), biotin sulfoxide reductase, as well as enzymes involved in the initial step of degradation of some pyridine derivatives [206, 207, 208]. The synthesis of MoCo requires GTP, linking the pathway to folate biosynthesis, which also draws from the pool pf GTP (Figure S7.12).

Structure of the Moa genes

These genes form part of a segment of DNA obtained by lateral gene transfer [209]. The expansion of the MoCo genes in the members of the MTB complex was part of the transition from an environmental generalist to a obligate pathogen [210]. As they allowed the organism to better survive in an oxygen starved environment such as a granuloma, these gene transfer events are significant in the adaption of MTB to the life of a pathogen.

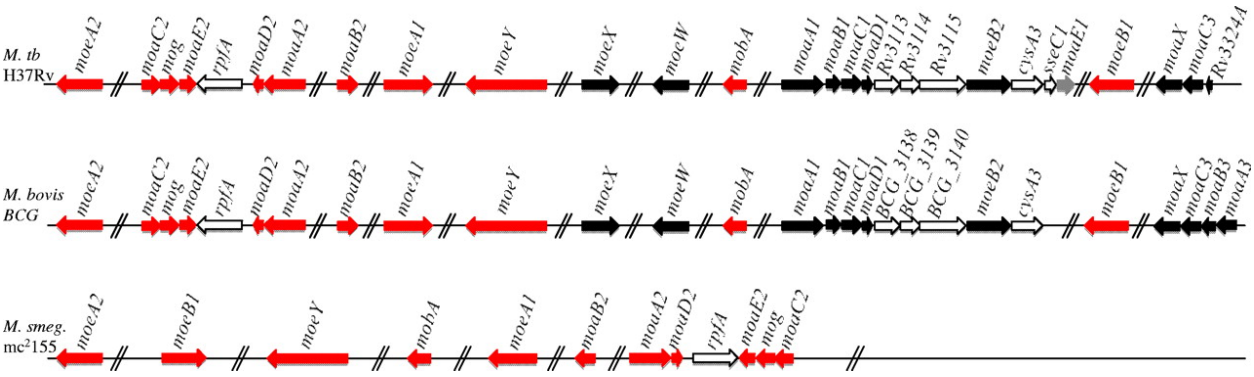


Figure 5.4: Structure of the Moa Genes in different MTB isolates. Figure obtained from a publication by Williams *et al.*[1].

The organisation of the MoCo biosynthetic genes in MTB genomes differs significantly both between and within species [1] with multiple clusters of MoCo genes found within the genome (Figure 5.4). Though these genes are annotated as having have the same function, they differ significantly at the sequence level (Table 5.7). It is this high level of sequence divergence that allowed the unique mapping of the reads to the correct operon.

Table 5.7: Similarity of the MoaA genes to one another in the CDC1551 genome based on sequence alignment.

	MoaA1	MoaA2	MoaA3
MoaA1	100	45.93	67.47
MoaA2	45.93	100	48.03
MoaA3	67.47	48.03	100

Table 5.8: Similarity of the MoaB genes to one another in the CDC1551 genome based on sequence alignment.

	MoaB1	MoaB2	MoaB3
MoaB1	100	46.30	58.93
MoaB2	46.30	100	46.78
MoaB3	58.93	46.78	100

The operon in question, the moaA3-moaB3-moaC3-moaX gene cluster (Refered to as the Moa3 operon hence-forth), is known to have been acquired from horizontal gene transfer

Table 5.9: Similarity of the MoaC genes to one another in the CDC1551 genome based on sequence alignment.

	MoaC1	MoaC2	MoaC3
MoaC1	100	54.45	66.47
MoaC2	54.45	100	55.75
MoaC3	66.47	55.75	100

[209] and studies have shown that the presence of the *moaA3* gene varies [211]. Unique to the operon in question is the *moaX* gene, a fusion of *moaD* and *moaE* which encode the two subunits of molybdopterin synthase which have maintained their catalytic ability [1]. The operon is present in this structure in *M. bovis*, BCG, *M. tuberculosis* CDC1551, W-148 as well as in isolates S507 and S5527 (Figure 5.5) and it thought to be the ancestral form. This structure was confirmed in the assembly of the two isolate genomes. No polymorphisms are present in this region, with the first variant seen 2,877bp upstream of the *moaA3* gene in a region at the end of the scaffold where differences may be as a result of incorrect assembly. No variants are seen in the 9,000bp downstream from the *moaX* gene. In H37Rv however, the *moaB3* gene is truncated and the *moaA3* gene is missing due to an IS6110-mediated deletion. The evolution of the MoCo genes is interesting, and highlights the importance of using the most common ancestor available for use as a reference when investigating gene expression, as well as the utility of having whole genome sequencing data for the strains in question. Use of the H37Rv genome would have resulted in the loss of information on the expression of *moaB3* and *moaA3* due to the truncation in this genome.

Regulation of the Moa3 operon

Though ChIP-Seq experiments have been conducted to identify the regulators of MTB genes [43], the isolate H37Rv was used as the reference strain which contains the IS6110-mediated deletion of the sequence upstream of the differentially expressed Moa3 operon. As a result there is no information available on regulatory interactions for the isolates in question. The known regulator of the MoCo biosynthesis pathway MoaR1 [212] showed lower expression in isolate S5527, but this decrease is not statistically significant making it likely that it does

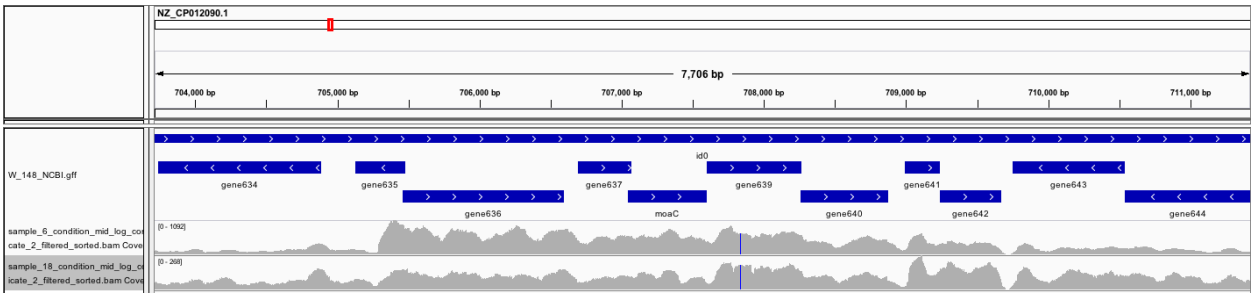


Figure 5.5: The structure of the Moa genes in the Moa3 operon as found in W-148 visualised in the IGV browser. Coverage of the regions from two of the samples included.

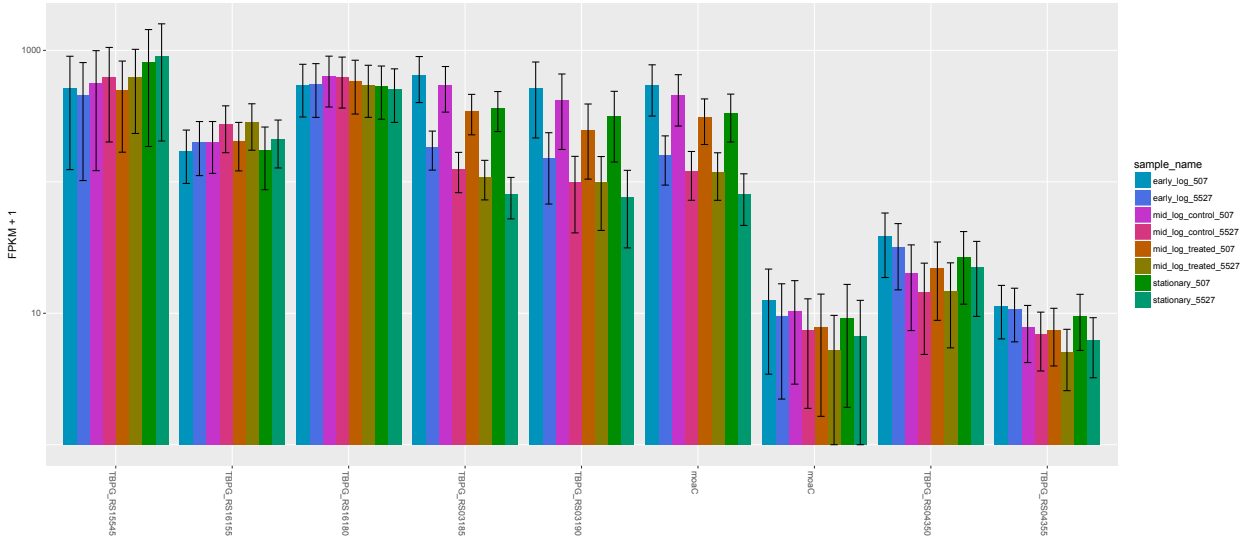


Figure 5.6: Expression levels of the different Moa Genes found in W-148.

not regulate the Moa3 operon. Six TF binding sites were found within the H37Rv Moa3 genes, though none of the transcription factors were found to be significantly differentially expressed, Rv0023 (TBPG_RS00145 in W-148) which is reported to bind to a region within Rv3323c showed an increase in expression in all S5527 samples relative to their S507 counterparts. As the whole Moa3 operon is down-regulated in S5527, it is unlikely that it is a result of regulation by a TF that binds within the operon, and further investigation into the upstream region was required.

We proceeded to use *in silico* methods to identify regulators of the Moa3 operon, which led to the identification of a binding site 96bp upstream from the start of the *moaA3* gene for a copper-sensitive operon repressor, CsoR (Rv0967 / gene3127). The *csoR* gene was

Gene	Function	Elog	ML(C)	ML(T)	Stat
moaA3	Cyclic pyranopterin monophosphate synthase 3	-1.75, **yes**	-2.15, **yes**	-1.68, **yes**	-2.13, **yes**
moaB3	Pterin-4-alpha-carbinolamine dehydratase	-1.65, **yes**	-2.13149, **yes**	-1.40, no	-2.16, **yes**
moaC3	Molybdenum cofactor biosynthesis protein	-1.80, **yes**	-1.92, **yes**	-1.41, **yes**	-2.14, **yes**
moaX	MoaD-MoaE fusion protein	-1.30, **yes**	-1.70, **yes**	-1.19, **yes**	-1.55, **yes**
moaD1	Molybdenum cofactor biosynthesis protein	no*	no*	no*	no*
moaC1	Cyclic pyranopterin monophosphate synthase accessory protein 1	no*	no*	no*	no*
moaA2	Cyclic pyranopterin monophosphate synthase 2	0.325, no	0.458, no	0.495, no	0.137, no
moaD2	Molybdenum cofactor biosynthesis protein	-0.0278, no	0.0690, no	-0.0179, no	0.0207, no
rpfA	Resuscitation-promoting factor	0.168, no	0.0968, no	0.237, no	0.0493, no
moaE2	Molybdopterin synthase catalytic subunit 2	0.0357, no	0.119, no	0.0477, no	-0.0903, no
mog	Molybdopterin biosynthesis protein	-0.0713, no	-0.147, no	-0.133, no	0.143, no
moaC2	Cyclic pyranopterin monophosphate synthase accessory protein 2	0.0804, no	-0.0309, no	-0.119, no	-0.265, no

Table 5.10: Comparison of the expression of the genes involved in the biosynthesis of MoCo in isolate S507 and S5527. The log2FC is such that negative values represent genes with a lower level of expression in isolate S5527. * These genes showed levels of expression too low for statistical analysis.

observed to be significantly down regulated in the Elog phase in the hyper-virulent S5527 strain (Figure 5.7) and showed lower expression in the other growth phases for S5527, though not significantly. Unfortunately, *in silico* methods such as this have low specificity and the probability of a false positive is high. Further tests would be required to confirm that CsoR is a regulator of the Moa3 operon. The other clusters of Moa genes did not show the CsoR binding site upstream from the TSS (Table 5.11), indicating they are likely not under the same regulatory control as the Moa3 operon. This is likely why they are not also found to be differentially expressed between the two isolates.

The involvement of a copper sensitive repressor is particularly interesting, as high levels of copper are known to interfere with proteins that have Fe-S clusters. As mentioned previously, proteins including WhiB3 and IlvD contain Fe-S clusters that are destabilised by high copper levels [213, 203]. MoaA is a [4Fe-4S] cluster protein [214], and these [4Fe-4S] clusters play a key role in the biosynthesis of molybdenum cofactor and the activity of molybdoenzymes in bacteria [215]. This could indicate that the expression of the Moa3 operon is linked to copper levels by CsoR as MoaA3, the first enzyme in the MoCo biosynthetic pathway, is inactivated by high levels of copper.

CsoR is a metalloregulatory repressor induced by copper that is known to regulate the copper sensitive operon (Cso). This operon contains three genes (Rv0968-Rv0970), and

includes includes *ctpV* (Rv0969), a metal cation-transporting ATPase / efflux pump. This operon was also found to have decreased expression in isolate S5527 in our datasets. An interesting note is that in a copper accumulating mutant of *E. coli*, copper sensitivity differed between anaerobic growth and aerobic growth [216], and in our results we likewise observed a difference in the expression of the Cso operon between the ML(C) and ML(T) samples, with the differences between the two samples shrinking during the ML(T) when compared to ML(C).

Investigating other genes that may be regulated by CsoR or influenced by copper, we find the copper sensing repressor encoding gene *ricR* (Rv0190/MT0200/TBPG_RS01010), a paralogue of *csoR* previously mentioned as a ML(C) gene of interest that regulates an operon containing *lpqS* (Rv0847/ maybe TBPG_RS16265) which encodes a probable lipoprotein induced by copper [201, 216]. In our results, we see that *lpqS* is significantly down regulated in S5527 during ML(C) and down regulated in all other conditions in S5527, though below the threshold of significance. Other genes found differentially expressed with a link to metals include a cation transporter (Rv2025c) that was differentially expressed during the Elog phase of growth, a cadmium-induced protein (CadI / Rv2641) [217] down regulated during Elog phase and ML(C) and more distantly *fmdB* (Rv0991c), a gene involved in the regulatory response to nitrogen limitation which is predicted to contain a zinc ribbon. Finally the metallothionein gene *mymT* was also found to be down regulated in S5527, and has previously been identified along with *lpqS* to be involved in the resistance mechanisms of MTB to phagosomal copper overload [216].

In the sRNA results, CsoR binding sites were identified close to three of the sRNA (B11, G2, MTS2975) found to have significantly lower expression in isolate S5527 (Table S7.3). The sRNA MTS2975 stands out in particular, as previously mentioned its targets included many genes that encoded proteins with cation binding function specifically iron including the ferredoxin FdxD and bacterioferritin BfrB. The other TF predicted to bind near MTS2975 was Lsr2, an essential MTB protein that protects the cell from reactive oxygen species (ROS) and linked to the expression of *bfrB* and an iron-responsive regulatory protein IdeR [173]. Although many of these iron-response genes are not significantly differentially expressed between the isolates, it is possible that the lowered level of CsoR in S5527 resulted

in lowered levels of MTS2975, resulting in more protein being produced by the MTS2975 target genes as a result of the decreased sRNA interference.

Metals like copper, zinc, iron and molybdenum are essential micronutrients in most forms of life including MTB with even small fluctuations having large effects on the ability of the pathogen to establish infection and large fluctuations being fatal to the organism. Biosynthesis of MoCo requires copper and iron in many organisms [218, 219, 220, 221] while in some including *Escherichia coli* and *Rhodobacter sphaeroides* copper is used when available in the biosythesis of MoCo, but is not essential [222].

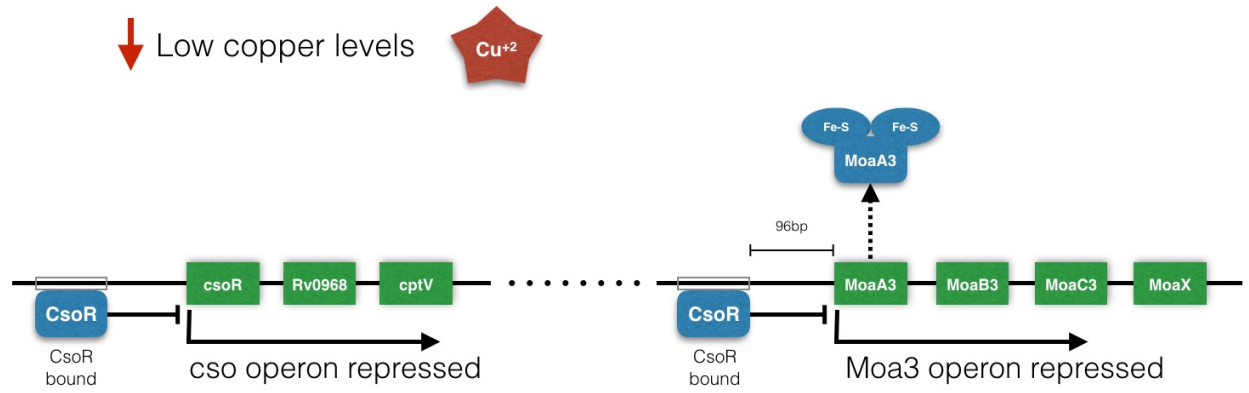
The phagosome attacks this delicate homeostasis, attempting to overload the bacteria with toxic levels of copper [216]. In turn, MTB has been shown to respond to copper levels during infection with copper sensitive transcription factors [199]. This is known as copper overload, and is one of the mechanisms used by macrophages to destroy MTB within their phagosomes [216]. The known mechanisms of this copper toxicity is iron-sulfur cluster degradation [Fe-S] [213, 216, 203] and metal cofactor replacement [217]. This is notable for the CadI protein that likely contains zinc, which is replaced by copper, resulting in an inactive enzyme [216].

Table 5.11: TF binding sites close to the start sites of Moa gene clusters in H37Rv as reported on TBDB.

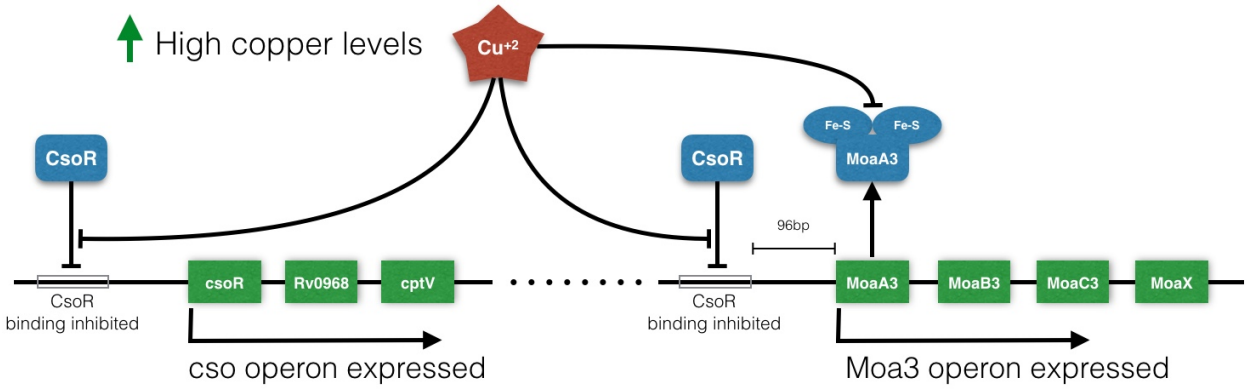
MoaA1 - MoaD1	Rv3597c (*lsr2*), Rv0081
MoaC2 - MoaE2	Rv0767c, Rv2034 and Rv1353c
MoaA2 - MoaD2	Rv1353c, Rv0691c, Rv1776c

The involvement of MoCo in pathogenesis

MoCo is an important cofactor for enzymes linked to virulence in MTB, both directly as a cofactor, and indirectly as a modulator of gene expression. It has been suggested that due to the different affinities of MoCo-dependent enzymes to the cofactors, altering the supply of the cofactor may impair the enzymes at different availability of the cofactor, thus acting as a concentration specific regulatory agent [207] . The disruption of *moaC1* and *moaD1* by



(a) When copper levels are low, the expression of the cso and Moa3 operons is repressed by CsoR.



(b) Under copper stress, the binding of CsoR is inhibited, and the cso and Moa3 operons are expressed. The increased levels of MoaA3 are probably to counteract the effect of copper disrupting Fe-S clusters.

Figure 5.7: The proposed link between the cos and Moa3 operon is the response to intracellular copper levels, and its effect on the binding of the CsoR repressor.

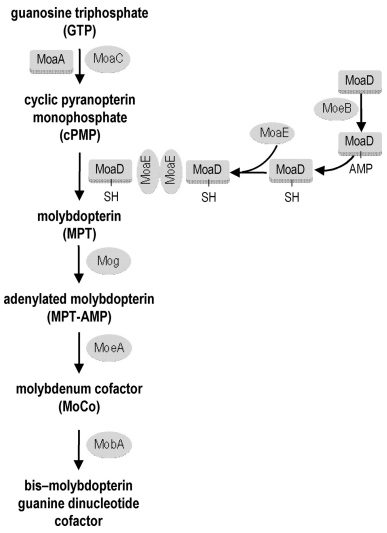


Figure 5.8: The molybdenum cofactor synthesis pathway as found in MTB. Figure obtained from a publication by Williams *et al.*[1].

transposon based mutagenesis was shown to impair the ability of MTB to block phagosome maturation, reducing the ability of the bacteria to parasitise macrophages [223].

Some of the MoCo-dependent enzymes include the narGHI-encoded nitrate reductase, an important protein for MTB to survive in the oxygen deprived environment of the granuloma [24]. MTB's ability to use nitrate from the host environment aids in its survival. As MTB has evolved, increased levels of nitrate reductase activity has been associated with increasing levels of virulence in the pathogen [224]. Another is carbon monoxide dehydrogenase (CODH) which is said to catalyse the conversion of $CO + H_2O \rightarrow CO_2 + 2H^+ + 2e^-$, with the presence of CO inducing the dormancy (Dos) regulon [225], and thus conversion of CO to CO_2 resulting in reactivation of the bacilli.

5.4 Conclusions

In this chapter we identified copper as one of the key drivers of the differences in gene expression between the two isolates. This manifested as the differential expression of a number of genes who's function is linked to cation levels, most notably the Moa3 operon. The link between copper and a number of the encoded proteins was the presence of a Fe-S cluster, that is destabilised by high levels of copper. This, together with the decreased

expression of the copper sensitive operon, indicated that the hyper-virulent S5527 strain is responding as though experiencing decreased levels of intracellular copper.

In detecting the differentially expressed Moa3 operon, the importance of selecting the correct reference genome was highlighted by the altered structure in the region when comparing the H37Rv and W-148 genomes. This added to the complexity of identifying regulatory elements for these genes, but *in silico* methods allowed for the identification of the copper sensitive CsoR TF binding site adjacent to the operon.

The cause of the different virulence phenotypes observed between the two strains is therefore more likely tied to their ability to resist phagosomal copper overload, a mechanism found in macrophages. The origin of the ability of S5527 to better survive the toxic copper levels remains unknown, as no clear link to intracellular copper regulation was found with any of the genes linked to the variants that distinguish the two isolates. It is possible that this missing link may be in one of the many hypothetical proteins, or in a uncharacterised pathway in which they are involved. Regardless, the are a starting point for further investigation of toxic copper resistance in MTB.

"'Nothing ever finishes. Nothing's ever really over.' It was Johnny who said that. He was surprised at himself. 'Correct! Are you a physicist?' 'Me?' said Johnny. 'I don't know anything about science!' 'Marvellous! Ideal qualification!' said Einstein. 'What?' 'Ignorance is very important! It is an absolutely essential step in the learning process!'"

Terry Pratchett, Johnny and the Dead

6

Conclusions

Identifying the mutations that lead to strains becoming more virulent is a vital step in understanding the epidemiology of tuberculosis, and stemming its dissemination. The availability of the closely related S507 and S5527 isolates provided the opportunity to study virulence in detail, and identify the genomic events that lead to their phenotypic divergence. We high-lighted the need for adoption of functional genomics methods that are able to utilise heterogenous data, and take advantage of new methods including the use of reference graph genomes. Thus, our aims were to identify the cause of the altered virulence phenotype, and identify the mechanisms through which isolate S5527 is more virulent.

6.1 The era of graph genomes: GenGraph

The shortcomings of using a single reference genome were exemplified in this project, where the structure of the differentially expressed Moa3 operon differed between the sequenced isolates and the H37Rv reference, obscuring the genes involved as well as the composition of the upstream regulatory region. We were able to use the less well annotated, but more closely related, Beijing family W-148 genome for read mapping and differential expression analysis by creating a genome graph with GenGraph, a tool developed during this project. GenGraph was developed to facilitate the adoption of genome graphs by making their creation simple, and providing functions that allow for their downstream manipulation and use in existing workflows. One such downstream function was the mapping of annotations of orthologous genes between isolates, allowing us to align reads to the W-148 genome, while taking advantage of the rich data available for the H37Rv annotations. We demonstrated that genome graphs provide scalability, and are able to represent a number of different genomes in a single reference graph. The tool was made available for use on GitHub and a paper was submitted for publication describing the tool and its use.

6.2 Adding to the regulation picture: sRNA

In order to have a more complete picture of the differences in gene regulation, we sequenced the sRNA component of the two isolates under different conditions. We then identified the targets of the sRNA using *in silico* prediction tools and created profiles of each sRNA, detailing when they are expressed, and the cellular functions they are regulating. We then identified sRNA that were differentially expressed between the two isolates, and found that isolate S5527 appeared to have a tempered dormancy response, existing in a generally more active state than isolate S507. We also find that the current catalogue of sRNA may be a underestimate of the total sRNA population in MTB, with over 150 potentially novel sRNA found in our dataset after mapping the sequences to the genome.

6.3 Putting together all the pieces

As the number of different data types increases, so does the complexity of the interactions. The development of Holmes allowed us to incorporate data into a traversable network, that could then be used to identify subnetworks of interest, and generate reports on the findings. This tool allowed us to sift through the heterogeneous datasets and identify the components of the network that were pertinent to our research questions.

6.4 The altered copper systems of S507 and S5527

The most striking results in the differential expression analysis was a set of MoCo biosynthesis genes found to have consistently lowered levels of expression in the more virulent isolate S5527. Because of the structural differences between the H37Rv and W-148 genomes at this location, *in silico* methods were used to identify TF binding sites upstream from this operon based on known motifs. A copper sensitive repressor was identified, and together with the lowered expression of an operon containing genes known to be involved in the response to copper levels in MTB, indicated that the two isolates were responding to copper in a significantly different manner. As high levels of copper have the ability to disrupt the iron-sulfur clusters found in the MoaA proteins, these two systems are interlinked. We thus described a novel regulatory link between copper levels and MoCo biosynthesis that is not found in isolates containing the disrupted Moa3 operon, as is the case with isolate H37Rv. We therefore hypothesized that the S5527 isolate showed increased levels of virulence due to its superior ability to survive in the phagosome, where the macrophage is reported to use high levels of copper to kill enveloped bacteria.

7

End matter

7.1 Ethics approval and consent to participate

Not applicable.

7.2 Availability of data and material

Project name: GenGraph

Project home page: https://github.com/jambler24/GenGraph

Operating system(s): Linux, Mac, Windows

Programming language: Python

Other requirements: Networkx, Mauve, and Mafft.

License: GNU LGPL

Datasets used during the current study are available from the NCBI https://www.ncbi.nlm.nih.gov.

7.3 Competing interests

The authors declare that they have no competing interests.

7.4 Funding

This work was funded by the National Research Foundation of South Africa, grant number 86934.

7.5 Appendix

7.5.1 Supplementary figures

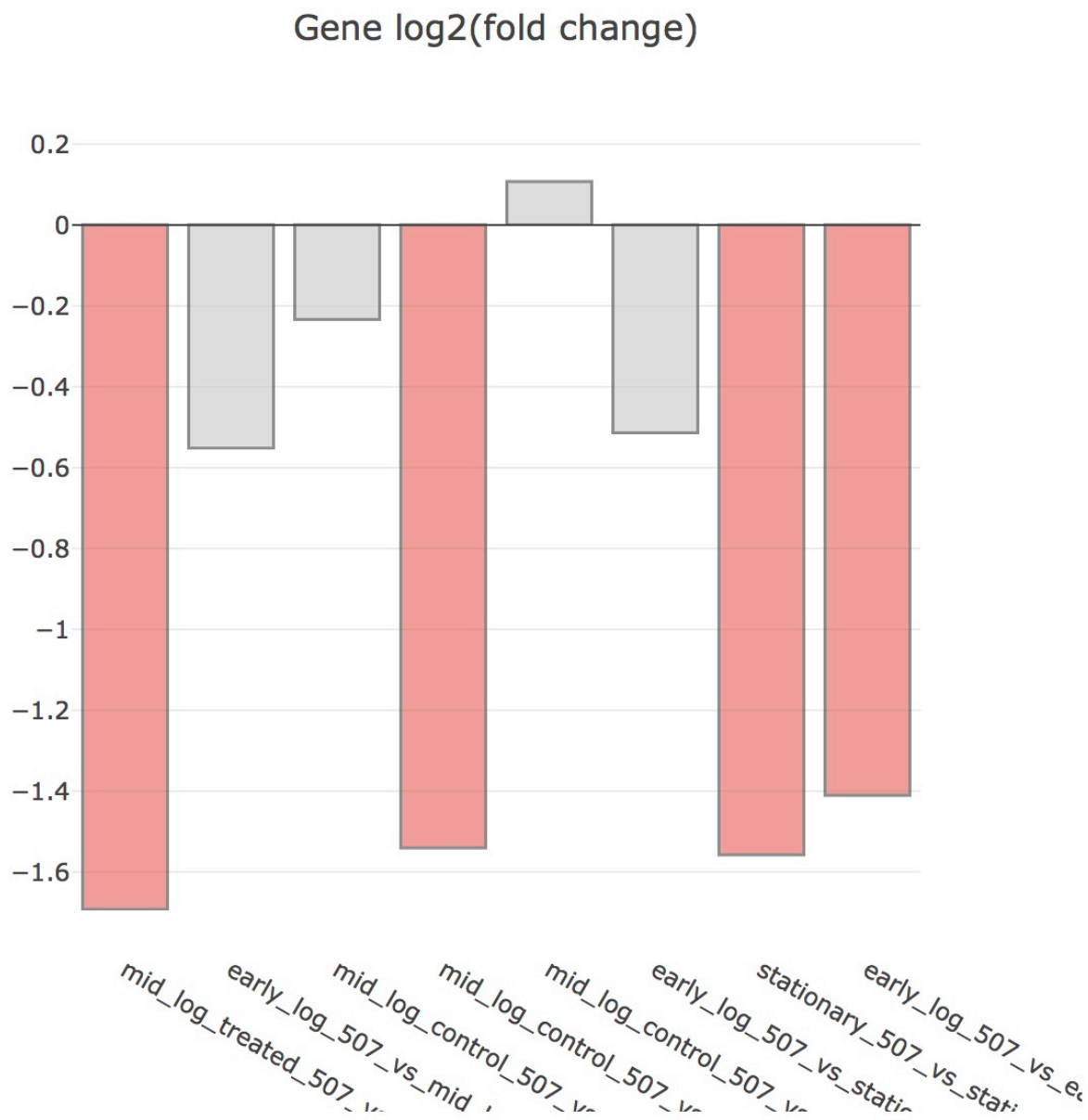


Figure 7.1: An example of a plot from a Holmes gene report. The expression of this gene under different conditions is shown, with significant differential expression (by q-value) represented by red high lighted bars in the bar chart. In theHTML report, this figure is interactive, and hovering the curser over a bar shows the related fold change and q-value.

String interactions

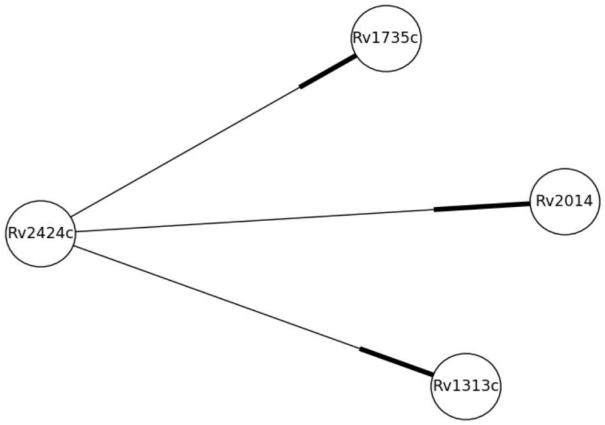


Figure 7.2: This plot shows protein-protein interactions for Rv2424c. These interactions are retrieved from the STRING database when the report is created via the API resulting in the most up-to-date information being used. The figure represents a subgraph relating to Rv2424c extracted from a larger interaction graph for the organism.

Gene	Condition	Log2 fold change	q value
Rv2338c	early_log_507_vs_mid_log_control_507	-0.893058	0.00214176
Rv2338c	early_log_507_vs_stationary_507	-1.07444	0.000898157
Rv3206c	stationary_507_vs_stationary_5527	-0.844124	0.0149923
Rv3206c	early_log_507_vs_stationary_507	0.775837	0.0134242
Rv3206c	mid_log_control_507_vs_stationary_507	0.861859	0.0321429
Rv3116	mid_log_treated_507_vs_mid_log_treated_5527	-0.519326	0.0256192
Rv3116	mid_log_control_507_vs_mid_log_control_5527	-0.486535	0.0445829

Figure 7.3: This table from the Holmes report shows proteins that interact with Rv3323c - MoaD-MoaE fusion protein MoaX. The genes in this table are differentially expressed with a q-value less than 0.05. Rows that are high lighted in red, are genes differentially expressed during conditions that the query gene (in this case Rv3323c) are also differentially expressed. In this way, genes that interact and have similar expression patterns are high lighted. In this example, the genes shaded in red are moeB1 (Rv3206c) and moeB2 (Rv3116), indicating they interact with MoaD-MoaE fusion protein MoaX (Rv3323c) and show similar expression patterns suggesting a regulatory link.

ncRNA regulation

ncRNA	Gene	Synonym	Energy	P value
F6	fadD11	Rv1550	-11.75	0.009

Figure 7.4: This section of the Holmes report shows any sRNA predicted by the RNA target prediction tool to interact with the RNA in question with a significant q-value for the interaction. The energy is the binding energy of the sRNA to the mRNA for the target gene.

Linked transcription factors

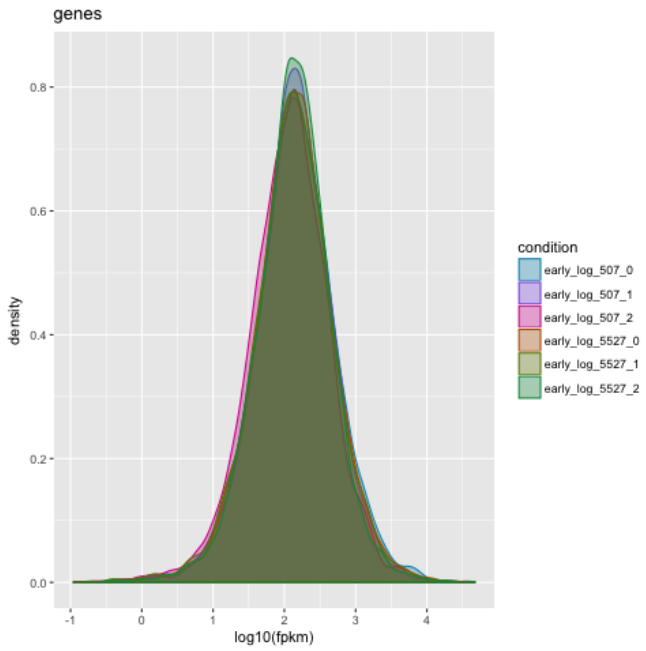
Regulator	Diff ex conditions	Log2(FC)	q value	Significant differential expression of target gene
Rv0967	mid_log_treated_507_vs_mid_log_treated_5527	-0.673829	0.00336379	no
Rv0967	mid_log_control_507_vs_mid_log_control_5527	-0.711332	0.00282681	no
Rv0967	stationary_507_vs_stationary_5527	-0.694968	0.101059	no
Rv0967	early_log_507_vs_early_log_5527	-1.37646	0.0102526	no
Rv0967	early_log_507_vs_mid_log_control_507	-0.748608	0.00214176	no
Rv0967	early_log_507_vs_stationary_507	-1.30546	0.000898157	no
Rv0967	mid_log_control_507_vs_mid_log_treated_507	-0.547636	0.07798	no
Rv0967	mid_log_control_507_vs_stationary_507	-0.417777	0.698426	no
Rv1776c	mid_log_treated_507_vs_mid_log_treated_5527	-0.0156182	0.996616	no
Rv1776c	mid_log_control_507_vs_mid_log_control_5527	0.0550637	0.976688	no
Rv1776c	stationary_507_vs_stationary_5527	-0.114482	0.998549	no
Rv1776c	early_log_507_vs_early_log_5527	0.299638	0.999657	no
Rv1776c	early_log_507_vs_mid_log_control_507	0.0787273	0.948217	no
Rv1776c	early_log_507_vs_stationary_507	-0.00657838	0.990289	no
Rv1776c	mid_log_control_507_vs_mid_log_treated_507	0.0517497	0.9993	no
Rv1776c	mid_log_control_507_vs_stationary_507	-0.154779	0.99893	no

Figure 7.5: This table shows the transcription factors predicted to regulate the target gene. These interactions are derived from ChIP seq analysis. Rows shaded red are conditions where the transcription factor is significantly differentially expressed based on a q-value less than 0.05. The last column shows if the target gene is also differentially expressed under that condition.

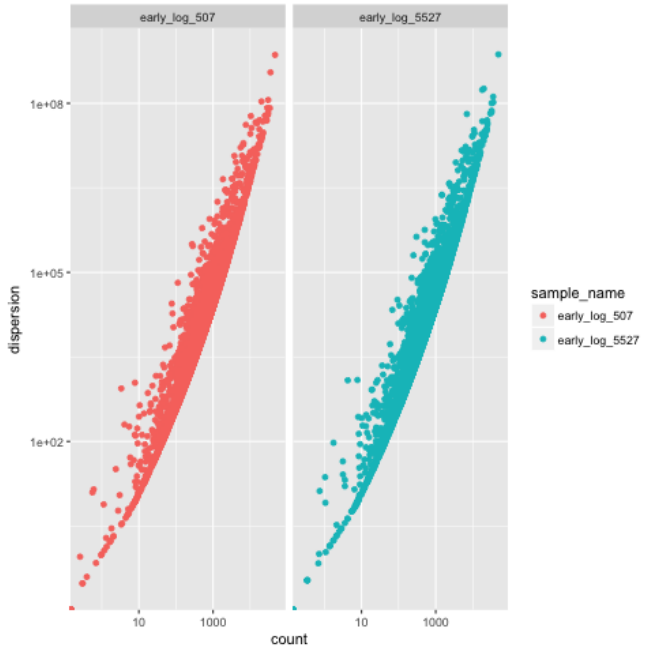
SNPs found within the query gene:

Position	Quality	Ref	Alt	Effect
2720952	205.0	T	C	missense_variant(MODERATE|MISSENSE|Acc/Gcc|p.Thr276Ala/c.826A>G|333|Rv2424c|protein_coding|CODING|NP_216940.1|1|1|1|WARNING_TRANSCRIPT_NO_START_CODON)
2720954	207.0	G	A	missense_variant(MODERATE|MISSENSE|aCc/aTc|p.Thr275Ile/c.824C>T|333|Rv2424c|protein_coding|CODING|NP_216940.1|1|1|1|WARNING_TRANSCRIPT_NO_START_CODON)

Figure 7.6: This section of the report shows any variants found within / near the target gene. These variants are annotated by snpEff, indicating the type of variant, and the effects of the mutation.



(a) Density plot.



(b) Dispersion plot.

Figure 7.7: Quality control plots for samples undergoing early log phase growth in isolates S507 and S5527.

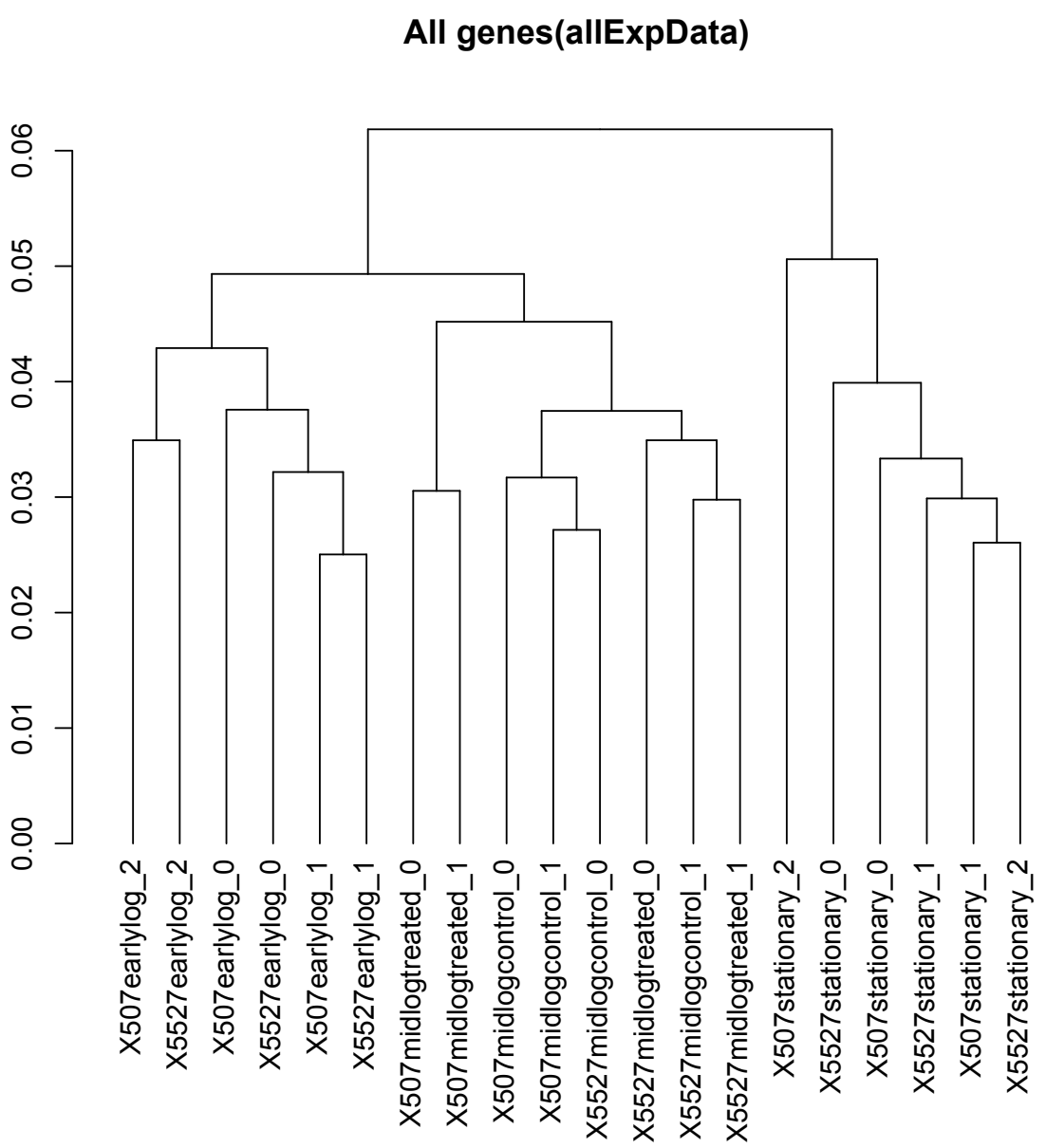
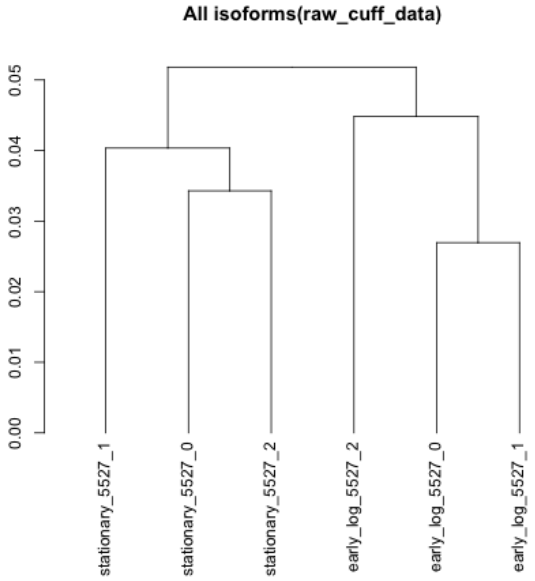
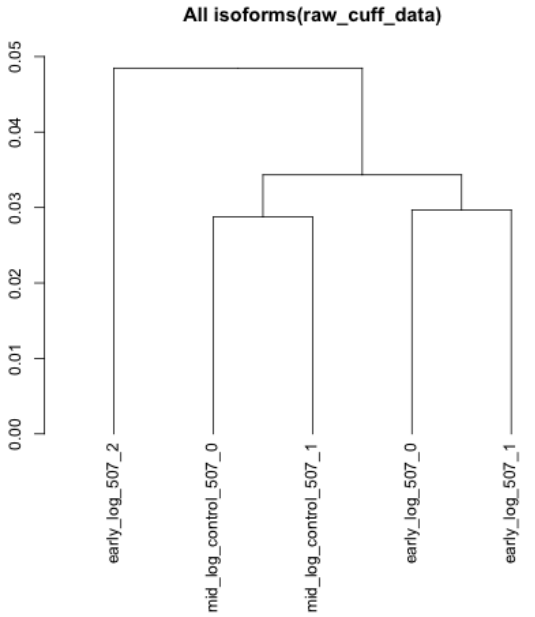


Figure 7.8: Dendrogram showing the clustering of all samples based on gene expression after normalisation.

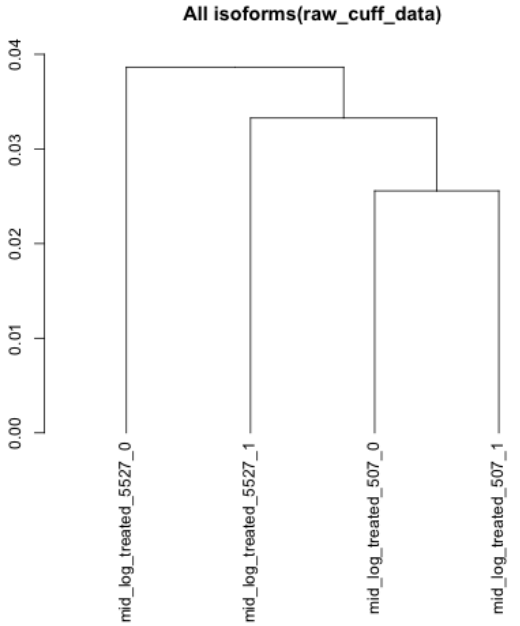


(a) Dendrogram showing the clustering of early log phase growth samples and stationary phase samples from isolate S5527.

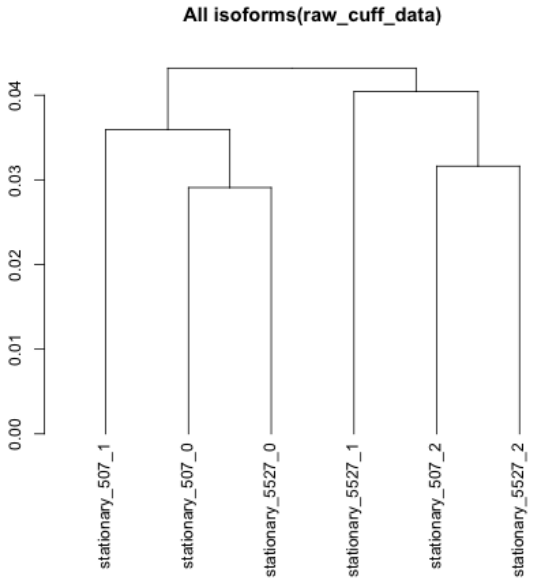


(b) Dendrogram showing the clustering of early log phase growth samples and middle log phase growth samples from isolate S507.

Figure 7.9: Dendrograms showing the clustering of samples in the condition versus condition comparisons. Figure (a) shows ideal clustering of the samples while in figure (b) one of the samples is incorrectly clustered.

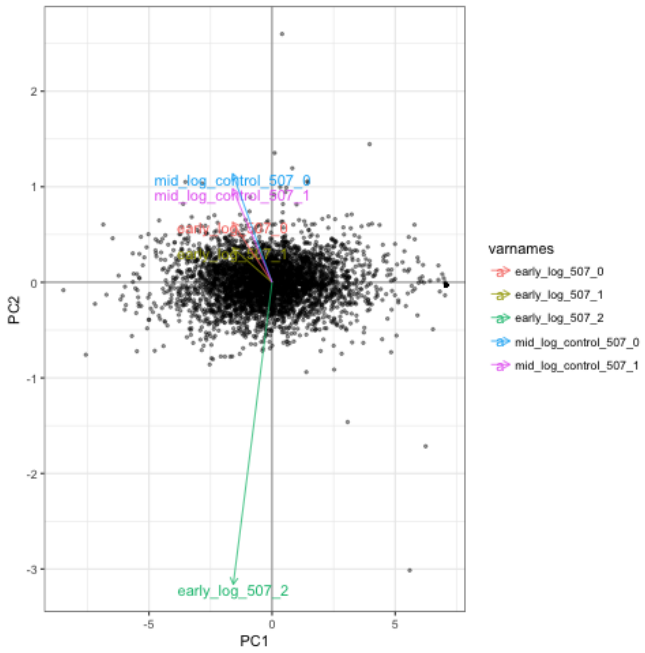


(a) Dendrogram showing the clustering of middle log phase growth samples from isolates S507 and S5527 where two problematic samples have been excluded.

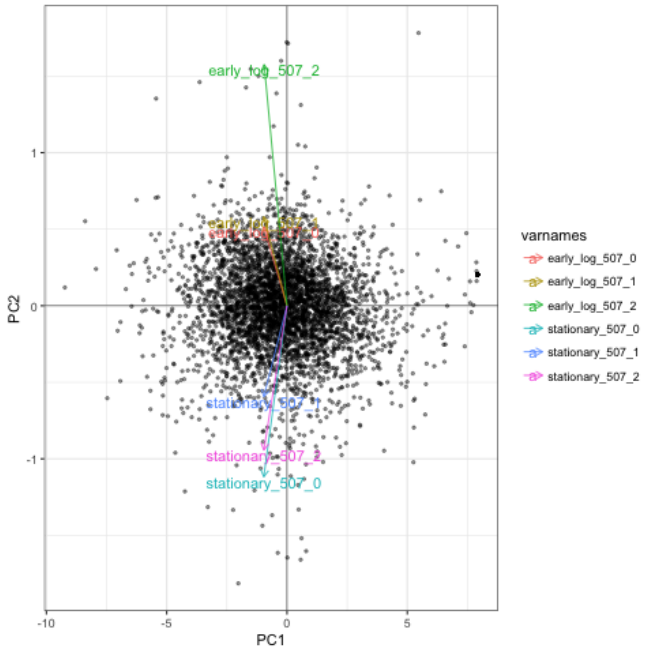


(b) Dendrogram showing the clustering of stationary phase growth samples from isolates S507 and S5527.

Figure 7.10: Dendrograms showing the clustering of samples in the isolate versus isolate comparisons.



(a) Comparing early log phase growth versus middle log phase growth samples where a sample from isolate S507 appears as an outlier.



(b) Comparing early log phase growth versus stationary phase samples where the sample from isolate S507 separates correctly by the second principle component.

Figure 7.11: PCA plots of condition versus condition experiments.

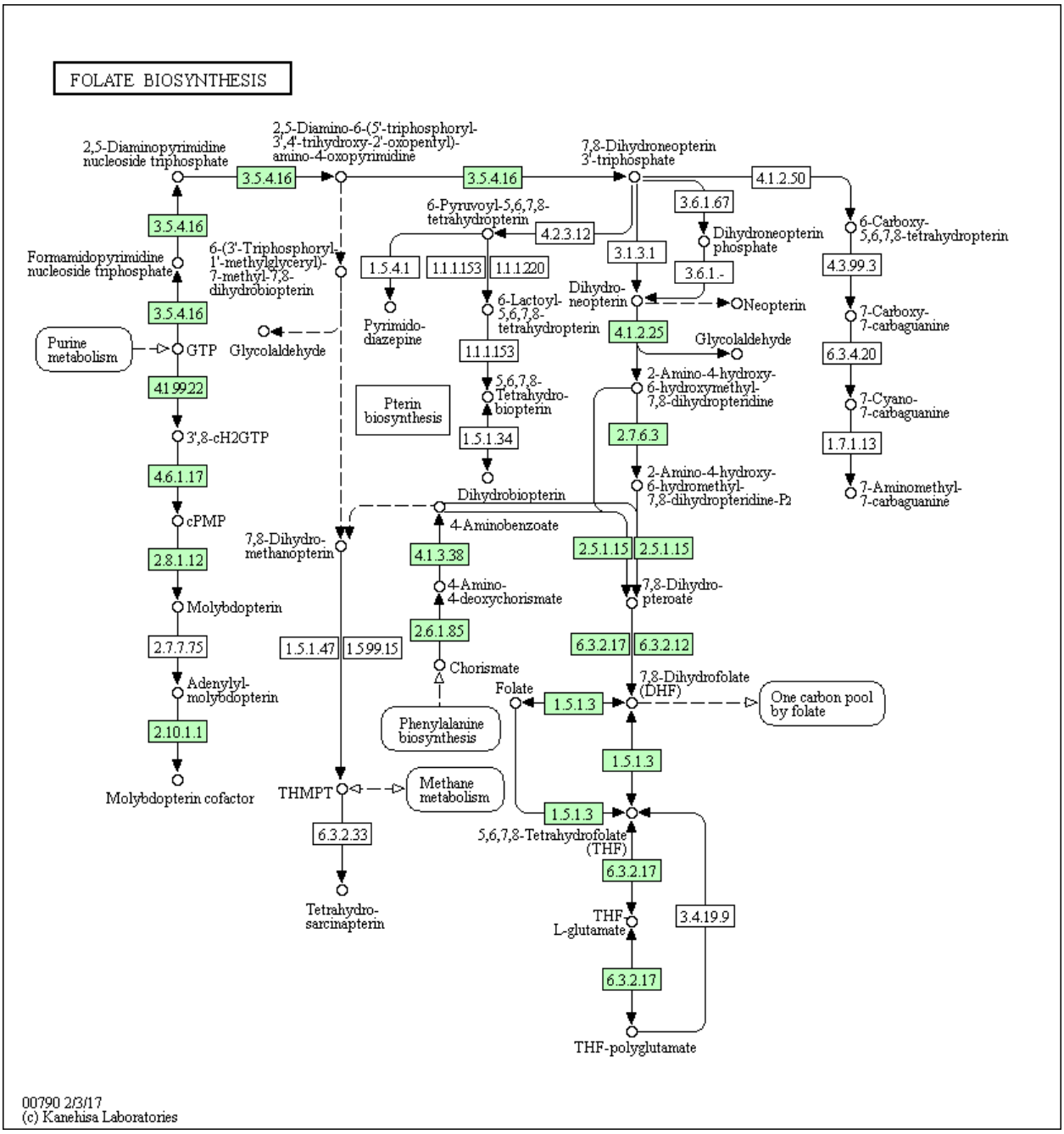
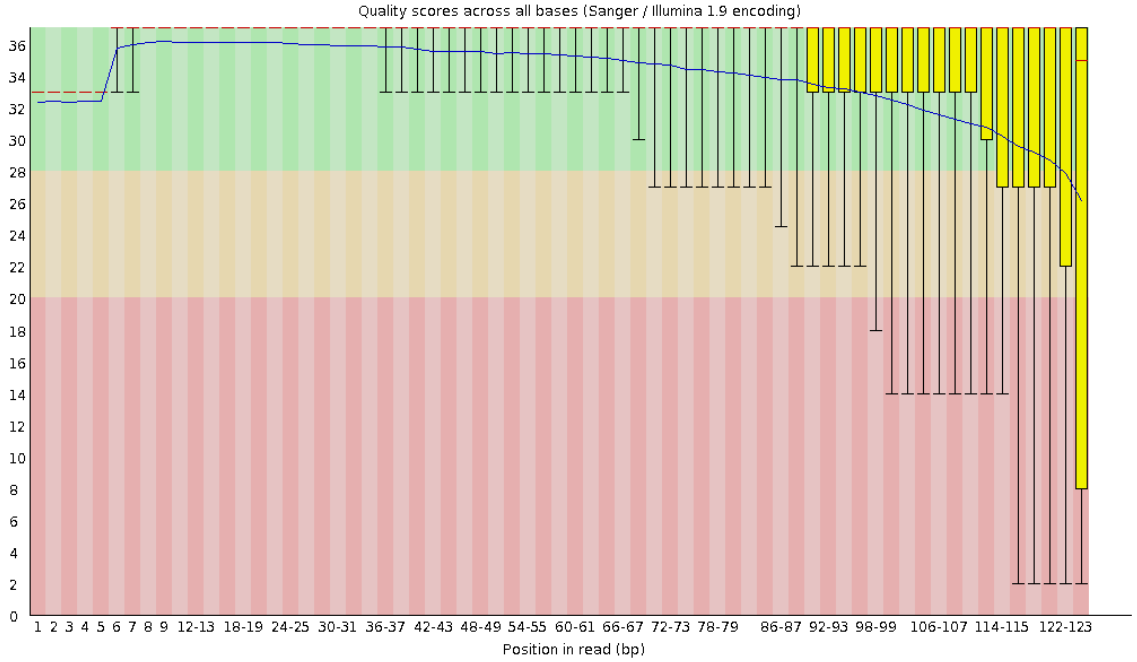
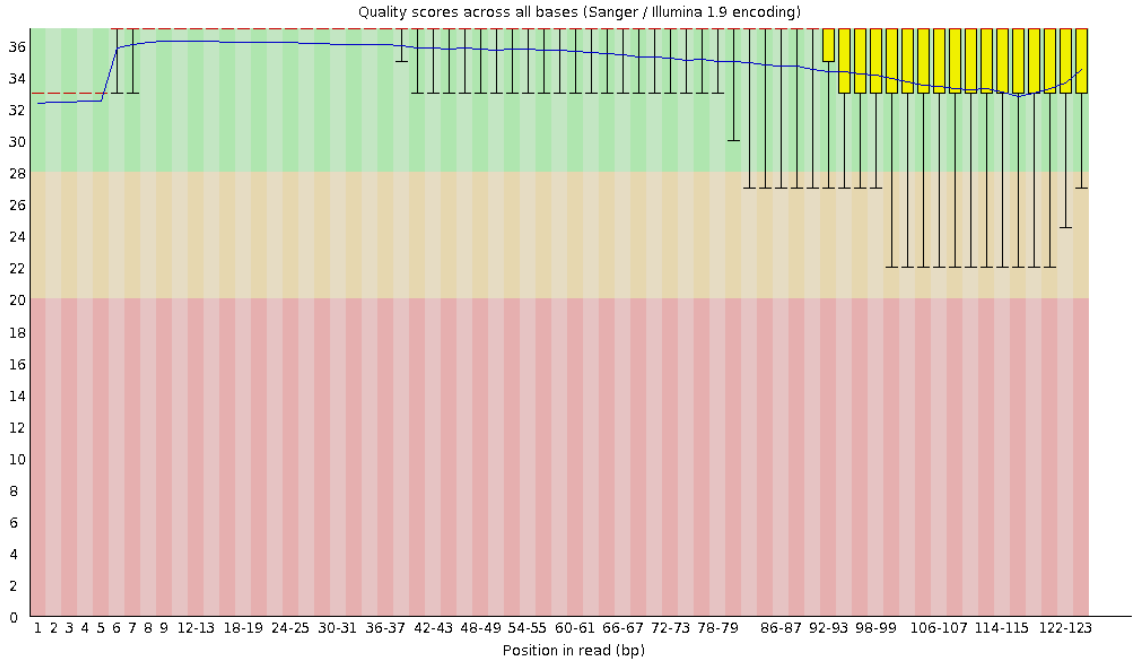


Figure 7.12: The folate biosynthesis pathway including the synthesis of molybdopterin and the link to purine metabolism. Figure generated by the KEGG webpage [2]



(a) Per base quality scores before filtration



(b) Per base quality scores after filtration

Figure 7.13: The per base quality scores for the sample 1 reads before (a) and after (b) read trimming and filtration.

7.5.2 Supplementary tables

Table 7.1: Summary of the genomes used for the alignment

Genome	H37Rv	W-148
Accession number	NC_000962.3	NZ_CP012090
length (in bp)	4,411,533	4,418,548
File size	4.3M	4.3M
Number of genes	3,999	4,133

Table 7.2: A comparison of the read mapping performance of two aligners BWA and tophat2 to the H37Rv genome.

Sample	Number of reads	BWA	Tophat2	Difference (BWA - Tophat2)
Sample 1	20 659 492	19 311 567	19 289 210	22 357
Sample 5	26 127 542	24 695 685	24 649 758	45 927
Sample 9	27 079 858	21 922 195	21 898 234	23 961
Sample 13	22 901 756	21 449 814	21 419 861	29 953
Sample 17	21 934 190	20 597 364	20 555 160	42 204
Sample 21	17 860 088	15 909 868	15 873 674	36 194

Table 7.3: Different sRNA considered in the analysis.

	earlylog	midlogcontrol	midlogtreated	stationary
AS1726	no	no	no	no
AS1890	no	no	no	no
ASdes	yes	yes	yes	yes
ASpks	no	no	no	no
B11	yes	yes	yes	yes
B55	yes	no	yes	yes
F6	no	no	no	yes
G2	yes	yes	no	yes
MTS0858	no	no	no	no
MTS1082	yes	no	no	no
MTS1338	yes	yes	yes	yes
MTS2823	yes	yes	yes	yes
MTS2975	yes	yes	no	yes
mcr10	no	no	no	no
mcr11	yes	yes	yes	no
mcr15	no	no	no	no
mcr16	no	no	no	no
mcr19	no	no	no	no
mcr3	yes	yes	yes	yes
mcr5	no	no	no	no
mcr7	no	yes	no	no
mpr11	yes	no	no	yes
mpr12	no	no	no	no
mpr17	no	no	no	no
mpr18	no	no	no	no
mpr5	no	no	no	no
mpr6	no	yes	yes	no
ncrMT1234	no	no	no	no
ncrMT3949	yes	yes	yes	yes

Table 7.4: Read mapping results: The number of reads aligned to the W-148 and H37Rv genomes respectively.

Sample	Total reads	W-148 mapped	W-148 percentage	H37Rv mapped	H37Rv percentage	Difference	Percentage difference
1	20,659,492	19,354,517	93,68%	19,311,567	93,48%	42,950	0,21%
3	14,680,064	13,958,217	95,08%	13,918,091	94,81%	40,126	0,27%
4	22,208,686	21,030,204	94,69%	20,984,532	94,49%	45,672	0,21%
5	26,127,542	24,762,993	94,78%	24,695,685	94,52%	67,308	0,26%
6	26,735,668	24,979,903	93,43%	24,907,022	93,16%	72,881	0,27%
7	23,389,294	21,637,207	92,51%	21,580,257	92,27%	56,950	0,24%
8	29,927,058	27,521,489	91,96%	27,456,879	91,75%	64,610	0,22%
9	27,079,858	21,994,792	81,22%	21,922,195	80,95%	72,597	0,27%
10	30,931,972	28,475,532	92,06%	28,391,370	91,79%	84,162	0,27%
12	21,864,758	20,696,416	94,66%	20,647,761	94,43%	48,655	0,22%
13	22,901,756	21,485,130	93,81%	21,449,814	93,66%	35,316	0,15%
15	28,080,056	27,111,743	96,55%	27,053,480	96,34%	58,263	0,21%
16	24,198,692	22,673,403	93,70%	22,635,718	93,54%	37,685	0,16%
17	21,934,190	20,634,002	94,07%	20,597,364	93,91%	36,638	0,17%
18	18,010,582	16,659,906	92,50%	16,623,332	92,30%	36,574	0,20%
19	19,040,746	18,328,934	96,26%	18,293,555	96,08%	35,379	0,19%
20	17,500,114	16,347,550	93,41%	16,317,265	93,24%	30,285	0,17%
21	17,860,088	15,950,149	89,31%	15,909,868	89,08%	40,281	0,23%
22	23,193,538	21,670,721	93,43%	21,626,259	93,24%	44,462	0,19%
24	24,492,936	22,858,510	93,33%	22,816,525	93,16%	41,985	0,17%

Table 7.5: Amalgamated results of the differential expression of genes during the early log phase of growth. This figure was generated as an output from Holmes. The log2FC is the difference in expression between S507 and S5527, with negative values representing decreased expression in S5527. The SNPs column refers to any variants found within the coding sequence of the gene. In the sRNA column, any sRNA predicted to interact with the gene are listed. If those sRNA are likewise differentially expressed, they will be followed by a greater than symbol identifying which isolate had the higher expression of the sRNA. The operon column high-lights clusters of genes that are all differentially expressed and probably part of an operon. Any gene proximal to the differentially expressed gene that is also differentially expressed is listed. The final column shows the homologue of the gene in H37Rv as predicted by GenGraph.

Gene	start	stop	Log2FC	SNPs	sRNA	Operon	Function	Homologue
gene196	214533	216608	-0.865271		Mcr11	gene197,gene199	beta-glucosidase	Rv0186
gene197	216654	216800	-1.04985			gene196,gene199	hypothetical protein	.
gene199	217750	218181	-0.901423		Mpr18, AS1890	gene197,gene196	hypothetical protein, possibly membrane bound	Rv0188
gene636	705461	706597	-1.74814			gene637,gene638,gene639	Cyclic pyranopterin monophosphate synthase 3	.
gene637	706694	707068	-1.64651			gene636,gene638,gene639	pterin-4-alpha-carbinolamine dehydratase	.
gene638	707044	707598	-1.80066			gene637,gene639,gene636	cyclic pyranopterin monophosphate synthase accessory protein	.
gene639	707599	708264	-1.30052			gene638,gene637,gene636	MoaD-MoaE fusion protein MoaX	.
gene669	741376	742185	-0.783855				hypothetical protein	.
gene694	766721	767002	-0.78453				hypothetical protein	.
gene744	819471	819686	-0.837649	SNP 819740			acetyl-CoA carboxylase biotin carboxyl carrier protein subunit	.
gene838	914748	915080	-0.68158			gene840	hypothetical protein	.
gene840	916719	917081	-0.90487			gene838	hypothetical protein	.
gene928	1003861	1004415	-1.18696				NAD(P)H-dependent oxidoreductase	.
gene1850	2007416	2007772	-0.849472				transcriptional regulator	Rv1994c
gene1887	2046894	2047892	0.78573		MTS2823 , G2: S507<S5527, Mpr6		cation transporter	Rv2025c
gene2481	2676932	2677954	0.892599		Mcr3, G2: S507<S5527		radical SAM protein	Rv2578c
gene2547	2740362	2740820	-1.34179				cadmium-induced protein CadI	Rv2641
gene2874	3082860	3084146	-0.740629			gene2875,gene2876	glycosyl transferase family 1	.
gene2875	3084404	3085543	-1.24528			gene2874,gene2876	membrane protein	.
gene2876	3085616	3086548	-0.691266		MTS2975: S507>S5527, Mcr19	gene2875,gene2874	formyltetrahydrofolate deformylase	Rv2964
gene3102	3305559	3305891	-0.730994	SNP 3307993			FmdB family transcriptional regulator	.
gene3127	3335742	3336101	-1.38523				transcriptional regulator	.
gene3252	3468030	3469244	-1.21104			gene3253,gene3254	pyridoxal-5'-phosphate-dependent protein subunit beta	.
gene3253	3469351	3469743	-0.871991			gene3252,gene3254	hypothetical protein	.
gene3254	3469892	3471406	-0.715796			gene3253,gene3252	oxidase	.
gene3268	3486476	3488844	0.862447				hypothetical protein	.
gene3697	3936799	3941659	0.860924	SNP 3935965		gene3700	PE-PGRS family protein	.
gene3700	3944566	3946275	0.965745			gene3701,gene3697	PE-PGRS family protein	.
gene3701	3946310	3951068	1.44011			gene3700	PE-PGRS family protein	.
gene3977	4228021	4228434	1.65136				hypothetical protein	.

Table 7.6: Amalgamated results of the differential expression of genes during the mid log phase of growth (control). This figure was generated as an output from Holmes.

Gene	start	stop	Log2FC	SNPs	sRNA	Operon	Function	Homologue
gene264	300441	300920	-1.34577				heat-shock protein Hsp20	Rv0251c
gene636	705461	706597	-2.15333			gene637,gene638,gene639	Cyclic pyranopterin monophosphate synthase 3	.
gene637	706694	707068	-2.13149			gene636,gene638,gene639	pterin-4-alpha-carbinolamine dehydratase	.
gene638	707044	707598	-1.92437			gene637,gene639,gene636,gene641	cyclic pyranopterin monophosphate synthase accessory protein	.
gene639	707599	708264	-1.70485			gene638,gene637,gene641,gene636	MoaD-MoaE fusion protein MoaX	.
gene641	708997	709239	-1.01712			gene639,gene638	antitoxin	.
gene840	916719	917081	-0.995408				hypothetical protein	.
gene847	924282	924602	-1.68372				hypothetical protein	.
gene857	931214	931684	-1.15749				hypothetical protein	.
gene928	1003861	1004415	-1.51905				NAD(P)H-dependent oxidoreductase	.
gene1030	1118212	1118511	1.05145				PE family protein	Rv1195
gene1533	1686738	1687505	1.11543		ASdes: S507<S5527		multidrug ABC transporter ATP-binding protein	Rv1687c
gene2547	2740362	2740820	-1.73384				cadmium-induced protein CadI	Rv2641
gene2875	3084404	3085543	-1.14462				membrane protein	.
gene3043	3247258	3248283	-0.948479				serine protease	.
gene3253	3469351	3469743	-1.29186			gene3254	hypothetical protein	.
gene3254	3469892	3471406	-0.979665			gene3253	oxidase	.
gene3563	3793607	3795112	-1.05701		MTS2823		type B diterpene cyclase	Rv3377c
gene3697	3936799	3941659	1.39419	SNP 3935965			PE-PGRS family protein	.
gene3701	3946310	3951068	1.49949				PE-PGRS family protein	.
gene4082	4359243	4359545	0.923179				ESAT-6-like protein EsxB	Rv3874

Table 7.7: Amalgamated results of the differential expression of genes during the middle log phase of growth when treated with H_2O_2. * Mpr12, F6, Mcr5, Mcr3, Mpr17, Mcr15, AS1890, and Mcr11. This figure was generated as an output from Holmes.

Gene	start	stop	Log2FC	SNPs	sRNA	Operon	Function	Homologue
gene44	43511	43858	0.90106		MTS1082, ncrMT1234, Mcr3		membrane protein	Rv0039c
gene150	165988	166899	1.15063				hypothetical protein	.
gene264	300441	300920	0.949428				heat-shock protein Hsp20	Rv0251c
gene311	359660	361435	1.07322		Mpr18, Mcr5, ncrMT1234, ASpks		PE family protein	Rv0297
gene636	705461	706597	-1.6803			gene638,gene639	Cyclic pyranopterin monophosphate synthase 3	.
gene638	707044	707598	-1.41051			gene639,gene636	cyclic pyranopterin monophosphate synthase accessory protein	.
gene639	707599	708264	-1.192			gene638,gene636	MoaD-MoaE fusion protein MoaX	.
gene788	863803	864147	-0.906326				hypothetical protein	.
gene840	916719	917081	-0.853546				hypothetical protein	.
gene847	924282	924602	-1.45256			gene850	hypothetical protein	.
gene850	925949	926173	-1.14935			gene847,gene853	hypothetical protein	.
gene853	927651	927965	-1.03783			gene850	hypothetical protein	.
gene928	1003861	1004415	0.806812				NAD(P)H-dependent oxidoreductase	.
gene993	1063387	1064070	0.982146		C8		hypothetical protein	Rv1158c
gene1169	1268885	1270660	0.936648		F6, AS1890, Mcr16, ASpks, ASdes: S507>S5527		PE family protein	Rv1325c
gene1175	1281356	1281661	-0.92429		Mpr18, Mcr7		ATP-dependent Clp protease adaptor ClpS	Rv1331
gene1251	1358304	1360271	1.0225				primosomal protein N'	Rv1402
gene1291	1398921	1400378	1.0262		B11: S507>S5527		PE family protein	Rv1441c
gene1333	1450863	1451588	1.0007				peptidoglycan endopeptidase RipB	Rv1478
gene1394	1518419	1518655	0.957961				hypothetical protein	Rv1535
gene1411	1540124	1541875	0.855798		Mcr5		fumarate reductase flavoprotein subunit	Rv1552
gene1533	1686738	1687505	1.14298		ASdes: S507>S5527		multidrug ABC transporter ATP-binding protein	Rv1687c
gene1580	1730636	1731892	-0.925171				penicillin-binding protein	.
gene1649	1812793	1814004	0.979079				PPE family protein	.
gene1660	1825585	1827090	0.946392		ncrMT3949: S507>S5527 and 8 others*		PE family protein	Rv1818c
gene1682	1852331	1853878	1.07726		Mpr17		PE family protein	Rv1840c
gene1773	1948310	1948954	-0.952689		Mcr3, Mcr11		TIGR03085 family protein	Rv1929c
gene1864	2026403	2027728	-0.861491		ncrMT1234		hypothetical protein	Rv2008c
gene2326	2503970	2504269	-0.888074		MTS1338: S507>S5527		PE family protein	Rv2431c
gene2521	2719753	2720253	1.02195	SNP 2724330	MTS1338: S507>S5527	gene2522	hypothetical protein	Rv2616
gene2522	2720270	2720710	1.05579	SNP 2724330	ASdes: S507>S5527, MTS0858	gene2521	hypothetical protein	Rv2617c
gene2544	2738828	2739274	0.904132		MTS1338: S507>S5527, Mpr17		hypothetical protein	Rv2638
gene2651	2828587	2830164	0.812223		Mpr12, Mpr17, F6, MTS2823		PE family protein	Rv2741
gene2789	2956426	2957289	0.948404		Mpr17, ncrMT3949: S507>S5527, ASpks		integral membrane protein	Rv2877c
gene2879	3087988	3088554	1.01677				16S rRNA (guanine(966)-N(2))-methyltransferase RsmD	Rv2966c
gene2988	3194873	3197434	0.980537				PE family protein	.
gene3043	3247258	3248283	-0.842084				serine protease	.
gene3050	3253237	3253521	-0.894638				ESAT-6-like protein EsxI	.
gene3266	3483601	3486249	0.834782			gene3268	PE family protein	.
gene3268	3486476	3488844	1.05998			gene3266	.	.
gene3563	3793607	3795112	-0.977656		MTS2823	gene3564	type B diterpene cyclase	Rv3377c
gene3564	3795117	3796007	-0.864669		Mpr12	gene3563	diterpene synthase	Rv3378c
gene3626	3858950	3860371	0.847202				hypothetical protein	Rv3433c
gene3697	3936799	3941659	1.89106	SNP 3935965			.	.
gene3701	3946310	3951068	1.84834			gene3704	.	.
gene3704	3954011	3956581	1.71063			gene3701	hypothetical protein	.
gene3853	4107820	4108590	1.0417		C8, Mcr10, Mpr12, ASpks		hypothetical protein	Rv3662c
gene3949	4206256	4206714	1.14712		Mcr11		tRNA-specific adenosine deaminase	Rv3752c
gene4082	4359243	4359545	1.32258				ESAT-6-like protein EsxB	Rv3874

Table 7.8: Amalgamated results of the differential expression of genes during the stationary phase of growth. * Mcr3, MTS0858, Mcr19. This figure was generated as an output from Holmes.

Gene	start	stop	Log2FC	SNPs	sRNA	Operon	Function	Homologue
gene302	350174	350464	-0.67179				ESAT-6-like protein EsxH	Rv0288
gene636	705461	706597	-2.13411			gene636-gene639	Cyclic pyranopterin monophosphate synthase 3	.
gene637	706694	707068	-2.16319			gene636-gene640	pterin-4-alpha-carbinolamine dehydratase	.
gene638	707044	707598	-2.14281			gene636-gene640	cyclic pyranopterin monophosphate synthase accessory protein	.
gene639	707599	708264	-1.54679			gene637-gene640,gene636	MoaD-MoaE fusion protein MoaX	.
gene640	708261	708875	-1.56857			gene637-gene640	SAM-dependent methyltransferase	.
gene665	734347	735261	-0.692078				esterase	.
gene669	741376	742185	-0.825914				hypothetical protein	.
gene696	767855	769348	-0.748944				LytR family transcriptional regulator	.
gene705	777934	778425	-0.688388				hypothetical protein	.
gene720	794145	794987	-0.701392				hypothetical protein	.
gene741	817569	818336	-1.11008	SNP 819740			RNA polymerase sigma factor RpoE	.
gene760	833528	834706	-0.965007				adenylyltransferase/sulfurtransferase MoeZ	.
gene782	859953	860138	-0.937859			gene784	hypothetical protein	.
gene784	860550	861815	-0.779702			gene782	hypothetical protein	.
gene814	886716	888617	-0.71999				PPE family protein	.
gene837	913850	914632	-0.734041			gene838,gene839	histidinol-phosphatase	.
gene838	914748	915080	-0.828911			gene837,gene839	hypothetical protein	.
gene839	915089	916231	-0.973209			gene838,gene837	PPE family protein	.
gene857	931214	931684	-0.871894				hypothetical protein	.
gene928	1003861	1004415	-1.04054				NAD(P)H-dependent oxidoreductase	.
gene934	1010458	1011432	-0.639842				ribonucleoside-diphosphate reductase subunit beta nrdF2	.
gene956	1031805	1032761	-0.701902				electron transfer flavoprotein subunit alpha	.
gene1325	1441814	1442185	-0.781206		Mpr18		thiol reductase thioredoxin	Rv1471
gene2754	2921457	2923049	0.767641		Mpr17, Mpr6		MFS-type transporter EfpA	Rv2846c
gene2947	3156591	3157724	-0.624226	SNP 3147782			2-methylcitrate synthase	.
gene3516	3714875	3716092	-0.737959		ncrMT1234		D-alanyl-D-alanine carboxypeptidase	Rv3330
gene3525	3725225	3726574	-0.686146		MTS2975: S507>S5527 and 3 others*		o-acetylhomoserine/o-acetylserine sulfhydrylase	Rv3340
gene3558	3789073	3790248	-0.698534		Mpr5, AS1890	gene3559,gene3561	trehalose-phosphate phosphatase	Rv3372
gene3559	3790485	3791126	-0.79195		ncrMT3949: S507<S5527	gene3558,gene3561,gene3562	enoyl-CoA hydratase	Rv3373
gene3561	3791380	3792807	-0.772888			gene3558,gene3559,gene3561-gene3564	amidase	Rv3375
gene3562	3792915	3793568	-0.973479		Mcr15	gene3559,gene3561-gene3565	haloacid dehalogenase	Rv3376
gene3563	3793607	3795112	-0.994172		MTS2823	gene3561-gene3565	type B diterpene cyclase	Rv3377c
gene3564	3795117	3796007	-0.833298		Mpr12	gene3562-gene3565,gene3561	diterpene synthase	Rv3378c
gene3565	3796016	3797626	-0.960355		Mpr17	gene3562-gene3565	1-deoxy-D-xylulose-5-phosphate synthase	Rv3379c
gene3594	3828710	3829597	-0.797705				taurine catabolism dioxygenase	Rv3406
gene3606	3840366	3840668	-0.848008		Mcr15		molecular chaperone GroES	Rv3418c
gene3613	3845619	3846148	-0.832893		ASpks		.	Rv3425

7.5.3 Supplementary data

Table 7.9: The sequences and positions of the novel sRNA after they were filtered. These positions and sequences are relative to the W-148 genome (NZ_CP012090.1)

sRNA number	Start	Stop	Strand	Sequence
sRNA_3	53232	53287	+	TCGAGTCACCTCCTTTTGTATGGCTTTTGAATGGCCGTTACGACGGTTCGACGCCT
sRNA_4	70069	70134	-	CCGCGCCTCCTAACTCCACAGCCGTATCGCGACGAATCGGCTACCGTCGCAACGG TGATGTGGCCG
sRNA_5	155863	155916	+	TAGCTACTACCAATCCCAACTCTCATCTGCCGCACGACGCGGTCGAATCTGTTC
sRNA_8	216800	216874	+	TTCGTATCACCCTCATCCGTGTCGGGGATCCCCGAGGAATCCCAGGTGGTCAGC TGTCGGTAATCCAGAA
sRNA_22	540265	540319	+	GTAGCCTAAGTAAAACATGGTTTTAGGCCCGAGCTCTCGACTCCTTACCTCGTTC
sRNA_23	559972	560038	+	GCCTTTCTGTGTTCCGAGGCCCGCATCCGCTTGCCTCGACGCACCCCTGATCTA TTTCCGATGCATC
sRNA_27	577541	577605	-	AGTCACAAGGAAAATCCTACAAATCCGGTGAACGTCGCCCTAGCGCGGCAAGGC CAAAATCGGAC
sRNA_30	611135	611186	-	CAGTACTGGCGCCGCCGGGCAACTCCGTGCCCCCGGCTGTCACGCAGTCCGT
sRNA_31	635783	635852	+	ATTCTGCTGGTCGGGATATTGCGTTGTCGATCAAACGAGTACGCGAAATGCGGG TGTATCTCGACTCGTC
sRNA_32	652136	652189	-	CAGTGCGTGTCCTCCTAGAACACCTCAACTTGGGAGATTACTGCTGGTTCAACG
sRNA_39	795795	795846	-	AAGATGTCGATTGCTCACCTCCTATCGCGGGATGCTGATTCAACCTGGGAAG
sRNA_40	810136	810202	-	CGTCTTCTCCCTGCGTCATACGGCCGATGACCTACGCTATCGTAACTTACGATTCC GTAGGTTACCT
sRNA_42	821524	821664	+	TTAGTCGTTACTCCTCACTATGTGCGCAGCGGTACGCACAGGCGTTTCTTCTTGG CTGGTAACGGGTGCACAAGAAAGGCTTCTGTACCCCTACATTTCTCTACATGCAAC CTTTTGATCGTTTCACAGGCTCAACAGATG
sRNA_45	831064	831141	+	GCTCTCGTCGTCCGGGTAGTCCTGGACTTCCCGGACGTTCCGAACGCACTCGTA GAGGTCGTTAACTGTGTTACCGAT
sRNA_47	951400	951842	-	GTCGCATCTGCAAATGGGCCACCGACCTGGCCCTTCGGTGGAGCTGCCGGGAAT CGAACCCGGGTCCTACGGCATTCCCTCAAGGCTTCTCCGTGCGCAGTTCGCTATG CCTCTGCTCGGATCTCCCGGTCACGCGAACTAGCCGAGATGACGATCCCAGTCGC TGTGGTTGTCCCGAGGAGTCCCGCGACCGGACTCATCGGTGGATCCCTCTAGCTG ATGCCAGGGTCCGGGCCGAGGGCGTTCCCGGTCTGACAGACTAGCCGTCGCTTAG GCAGCGAGAGCGTAGTCGCGCTGATGTGAATCGGCGCTTATTTGGTCGCAACGACG CTTACGGTGGTCTCTTGCCTGCACCGGCACGCTTCCCTTGATTCGATGCGCGAAGT CGAAACCGTTCAGCCCCTCGCATCCCTGCCGACCTTCGGCAGGACCATCAATCCTAC
sRNA_56	1118099	1118178	-	CACGGGTACGATCAGGAGCGGTTAGCAGTCAAGCGAACTGGTCGCCTACGTCCT TGGCCCCGATTCGTCGGTGACGAACT
sRNA_62	1252445	1252557	-	GACCCAACAGTGTGTTGGTGGCCAACTTTGTTGTCATGCACCCGGCTCTCGCCC ACTACAGACAAGAACCCCTCACGGCCTACGCCCCACCAGTTGGGGCGTTTTCGT GGTGC
sRNA_64	1255855	1255941	-	GTTCTGGTGGGGGGTGTGGACGCACGGCTAGCGCCGTGAACGGATGTGGTTG CGAGTTTGTTTTTGCCTCCCTTTTCCAAAAGGGAG
sRNA_71	1434664	1434756	-	CGGGTTTCGCTCGCACCGCCCGCCGAACGTGAACTCACGGCGGTATTTTGCCG GATTCTCCGCCCTCAGTTCACGTTCGGCGACGCGCCGGTT
sRNA_72	1451657	1451726	-	CAAAATCCTCCACAGCTCAATCGGACACGACTGCCGACATGACCAACGTCGCGG GGCAGCGACGCGCCCG
sRNA_77	1557084	1557165	+	ACTCTCGGGTGGTGTGTGTCTCAGCACGTGACTTCACCGTCTGCCATTCCAGCC GGAAGTCACTTTATTCACACCAATCACT
sRNA_78	1567019	1567088	+	CAGCGCCGGCACTCAAGGTCAGCGTCGGCACTCGAATGGCGCCAGCGGCTCTT ATCCGGCTCTTAAAGTC
sRNA_85	1736863	1736950	+	TCGAAATCGACGCCAGCGCGGACTTGTTCGACGAGTAGACGTGTCGCTAACGTC GATCTCGATGGGCAGTCCTGTCCGCTCGCCGAAG
sRNA_92	1812001	1812079	-	TCCTGACCCTGGGCGTCTTTCGACGTTCGAGGTCAGTGGCCTATATCCGCGCAGA CGCCTCACCTAGCGAGGTGCTTGC
sRNA_95	1820519	1820600	+	TGTCCTGCCCCCTTTCTGCGGTCGGTAATCCAGCGGTTTGAAAGGGTTGAGCCGA CTTACGCGCAGTGGATGCGTCGAAGGG
sRNA_98	1838257	1838327	-	TCCGACTCACGCTCGGTGCGACGACGCGGTGGGCGCCACGTCTCTACCGTG ACGTCGAACACCAAGCTGTC
sRNA_99	1839109	1839245	-	GGTGCTTCCTCTCGCCGATCGCGCTCTTGTCACTAGCTGCGTCACTGGCTCC CCCGACAGCCATTACTGAGGGCCCGAGTGTTAGAGAGCATACGCTGTTCGA TGGGCGCGATGCTCCCGGTTAACTATCAAGTCGT

Table 7.10: The sequences and positions of the novel sRNA after they were filtered. These positions and sequences are relative to the W-148 genome (NZ_CP012090.1)

sRNA number	Start	Stop	Strand	Sequence
sRNA_109	2166177	2166250	-	TTCAGGAGTCTCGGGCGGCTTCGTAATGGCGGTCCTATCGTTGTCCTACCGG CGCGGAATTTGCTCTTGCGATC
sRNA_112	2183786	2183867	+	CTCGGGTGGCGTGCCCACATCTCATGGCGGGCCACGCCCCGCCCAGCGTG GATGCCCAATGGGTCTACAGGCGACCGTCGCG
sRNA_117	2245626	2245722	+	TGACTTCTCCTAGATGTCTCATCGTTGGGTGGGCCCCGCCCACTAGCGTTTC AGCCTGCGGAATCCAGTCTGGGGTCTGCTTGGGGAAAATCCCACT
sRNA_119	2286595	2286680	-	GGCTCTAGATCGCCGAGCGTGAACCTGGCGACGCGACACACGCCGCCGGT GTCGGGCTGTCAGGCTCACACTCGGCCGGGCTCTAT
sRNA_126	2429988	2430044	+	CGCTTGATTCTCCTATGCCGCGTCTTTATGCCGCTTCTCAAGCGGCTATCCACAAAC
sRNA_127	2456390	2456458	-	CGTCGGGTCCTCTCTGTTGACTGGCTAACGATCGGGCGATGCCTGGGCAGA CCCAGCGGACATACCGAG
sRNA_131	2520915	2520972	+	ATGACGGCCCGACATGCCGCGCCGCATCGAAGGATGGCGATGGTCGCGGT TGCCTAAG
sRNA_135	2539544	2539603	+	TGTCAGTCACTTGGCTCACAGTGGGGCACCTGCTTTCCTCGAGTTCTTCTATGCTCCGAC
sRNA_138	2625440	2625502	+	CGTAGGCAGCCCCGTGCGCTTGCCGGGCAGGTGTCCTCAAAGGTCCAACT AGACACACATATC
sRNA_141	2723960	2724015	-	CGCGGTCCTCCTGTCGTTGGCTGATCGCCCGCTCGGCCGATGTGGCTTGTCCCTAC
sRNA_144	2729993	2730117	-	TCAACCGGTGATCTCTTCGCTGTCGAGACTGACCAGTACGACGCGGAA GGGTATGTGCCCACGGGTTTCATAACTCAAACCCACTGTTGCCCATAA GGATCCTGGTAGAGCAGTACATGTAGCTT
sRNA_151	2753974	2754027	-	TGGGCCGGTGGGATACTGTCTGACCTGTACAGAAAGCCTCTGACCAGGCGACAT
sRNA_153	2776606	2776663	-	GGCCATCCTCCGAATCTTCTCGTAGGCCGCTCATCGCGGCGTCGTCTGGT GCTACAGG
sRNA_158	2918409	2918486	+	CTTGAGTTGTCCGGTCATCTAGCGGAGGCGCCGCCAGGGCGGCTCCCAG TGTCCCGCCGGCACGCAGCAGCCGGCGTA
sRNA_163	3051796	3051849	+	ACTCTTCCAACCTCGTCTCAGCTCAACCGGTGTTACCCGACGACATCAGCGAAT
sRNA_164	3051872	3051970	+	CCGCGGTGCCGCTCTCCCAGCTCTTAAGTAATCCGAGCCAACCCGGATCC CGACACCAAAGACAAGTGTTACACGACGCCAAGACCCCCCGCGGGTAGC
sRNA_170	3132230	3132353	+	AAGAAAAAACCCTCGCCAGCTCAGCTGCTGCACGAGGGTCGCGTTGGTGC TCGCTTGGGCTAGTCAGGCACCAACGCGCCGACCAATTACTACGAGCATCC CGGGCTTTCCGGCCTTGTCCATA
sRNA_172	3172270	3172321	+	TTGCCGCTCAGGGTACCCGGGCCGGGCACGTACACTCCATACTCCCACTTTG
sRNA_174	3210629	3210689	-	TGGCTCATCTCACCGCCGGGCGTTCCGGTGAATCCGGTCCTCAGGTACTGTA GTCCCGCCT
sRNA_175	3216696	3216783	+	CCCACCGACACAGCCGTTGATCCTGCGTCTACCACGCAAAAGTGCGCGTGGT CAGCCTGGTGGACGCAGAATCAACGGCCAACGAGCG
sRNA_186	3504289	3504341	-	TCCTAGGCTGCTCTAAGTGCTGCGCGGACGTGCGCGGCTACTCAGCAGCTACA
sRNA_187	3508621	3508704	-	GGGGGAACCCCCTCGCGAAATAACGGAGCGGCCTAACGAGTAGGCGGCTCC GATCTCTGGTGTCGTTTATTGTCCTGCCGACAG
sRNA_190	3628995	3629099	+	TTGATGTTTTCAGTTATGCCGCGGTCGTAACCGGGCCAGCCTACTCGTATGGT TGTATCTACTCGCCAGTCGCCTTCGCGCTGCTCATTGCTCGTGAGCTTCGCT
sRNA_202	3913251	3913315	+	CCCCCATCAATCATTCGGTGGCGCGAAGTTCACCAGAGTCCCGGACACGCTC ACGCGAACTACCT
sRNA_211	4107239	4107538	-	CAAAAAGCGGGCGGCGACCCGAAGAACTCGAATCGCCGCCCACCAACACGG TTCTCGGTTACCAAGCGTGCACTCCTGGGGTTGCGTGGGTGGCCTCGGCGA TCTTGCGACGCTTCTAGCTGTAGCCCCACCCAAAGGGCCGTCGATGCCATCT GCTGTTCGCAATTACGCAGACCCGCAACACTTCTGCCGTTATCGTGGCTATGA CTCCGCGTGGGTGCCGGGAACCATGCTGGGCGCCCGGGTCGGGAGACCGA ACCTTCTCTTCTGGATCGAGCCTTGGCCTCACCGGGCTTCCGT
sRNA_213	4160629	4160725	-	AAGGAACCTCAGACCGGCGCATCGCGAACGTCCCGCGACGGGAAGCCGGTCT GGATCAGACCCCGTCGCGGCGTCCGAGGAGGAGCACCCGCTGCAC
sRNA_214	4175201	4175290	+	AAGGGGACCCCGCGCACCCGACAGAGCCCGTTGACCCTTGCTGCCTTCCAGC CCTGGGGGAGTTCACAGGATAGACGCCGCGCGGGGTCC
sRNA_220	4280467	4280532	+	GGTGCGCATGGCCGACCAGTGTGGTTGGCCGGAGGTCGTTTGGTCGCGATTG CCTCACGATTCGAT
sRNA_223	4357581	4357638	-	GTCAGCATTGCCCGAACTCGATTACCGGGCTGAGCTCGGGTCCTGCTACACCGGCAAT

Bibliography

[1] Williams, M.J., Kana, B.D., Mizrahi, V.: Functional analysis of molybdopterin biosynthesis in mycobacteria identifies a fused molybdopterin synthase in Mycobacterium tuberculosis. J. Bacteriol. **193**(1), 98–106 (2011). doi:10.1128/JB.00774-10

[2] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M.: KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **27**(1), 29–34 (1999). doi:10.1093/nar/27.1.29

[3] M Cristina, G., Brisse, S., Brosch, R., Fabre, M., Omaïs, B., Marmiesse, M., Supply, P., Vincent, V.: Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis. PLoS Pathog. **1**(1), 0055–0061 (2005). doi:10.1371/journal.ppat.0010005

[4] Wirth, T., Hildebrand, F., Allix-Béguec, C., Wölbeling, F., Kubica, T., Kremer, K., Van Soolingen, D., Rüsch-Gerdes, S., Locht, C., Brisse, S., Meyer, A., Supply, P., Niemann, S.: Origin, spread and demography of the Mycobacterium tuberculosis complex. PLoS Pathog. **4**(9) (2008). doi:10.1371/journal.ppat.1000160

[5] Morse, D., Brothwell, D.R., Ucko, P.J.: Tuberculosis in ancient Egypt. Am. Rev. Respir. Dis. **90**(4), 524–41 (1964). doi:10.1164/arrd.1964.90.4.524

[6] Herzog, Basel, H.: History of Tuberculosis. Respiration **65**(1), 5–15 (1998). doi:10.1159/000029220

[7] Brontë, E.: Wuthering Heights, (1847)

[8] Ford, C., Yusim, K., Ioerger, T., Feng, S., Chase, M., Greene, M., Korber, B., Fortune, S.: Mycobacterium tuberculosis–heterogeneity revealed through whole genome sequencing. Tuberculosis (Edinb). **92**(3), 194–201 (2012). doi:10.1016/j.tube.2011.11.003

[9] Bifani, P.J., Mathema, B., Kurepina, N.E., Kreiswirth, B.N.: Global dissemination of the Mycobacterium tuberculosis W-Beijing family strains. Trends Microbiol. **10**(1), 45–52 (2002)

[10] Filliol, I., Motiwala, A., Cavatore, M.: Global phylogeny of Mycobacterium tuberculosis based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recomendations for a minimal standard SNP set. J. Bacteriol. **188**(2), 759–772 (2006). doi:10.1128/JB.188.2.759

[11] de Souza, G.a., Fortuin, S., Aguilar, D., Pando, R.H., McEvoy, C.R.E., van Helden, P.D., Koehler, C.J., Thiede, B., Warren, R.M., Wiker, H.G.: Using a label-free proteomics method to identify differentially abundant proteins in closely related hypo- and hypervirulent clinical Mycobacterium tuberculosis Beijing isolates. Mol. Cell. Proteomics **9**(11), 2414–23 (2010). doi:10.1074/mcp.M900422-MCP200

[12] Meissner-Roloff, R.J., Koekemoer, G., Warren, R.M., Loots, D.T.: A metabolomics investigation of a hyper- and hypo-virulent phenotype of Beijing lineage M. tuberculosis. Metabolomics **8**(6), 1194–1203 (2012). doi:10.1007/s11306-012-0424-6

[13] Lewandowski, C.M., Co-investigator, N., Lewandowski, C.M.: WHO Glocal tuberculosis report 2015. Eff. Br. mindfulness Interv. acute pain Exp. An Exam. Individ. Differ. **1**, 1689–1699 (2015). doi:10.1017/CBO9781107415324.004. arXiv:1011.1669v3

[14] Corbett, E.L., Watt, C.J., Walker, N., Maher, D., Williams, B.G., Raviglione, M.C., Dye, C.: The Growing Burden of Tuberculosis. Arch. Intern. Med. **163**(9), 1009–1021 (2003). doi:10.1001/archinte.163.9.1009

[15] Getahun, H., Gunneberg, C., Granich, R., Nunn, P.: HIV infection-associated tuberculosis: the epidemiology and the response. Clin. Infect. Dis. **50 Suppl 3**, 201–207 (2010). doi:10.1086/651492

[16] Singh, J.A., Upshur, R., Padayatchi, N.: XDR-TB in South Africa: no time for denial or complacency. PLoS Med. **4**(1), 50 (2007). doi:10.1371/journal.pmed.0040050

[17] Knechel, N.a.: Tuberculosis: pathophysiology, clinical features, and diagnosis. Crit. Care Nurse **29**(2), 34–4344 (2009). doi:10.4037/ccn2009968

[18] Frieden, T.R., Sterling, T.R., Munsiff, S.S., Watt, C.J., Dye, C.: Tuberculosis. Lancet **362**(9387), 887–99 (2003). doi:10.1016/S0140-6736(03)14333-4

[19] Aderem, a., Underhill, D.M.: Mechanisms of phagocytosis in macrophages. Annu. Rev. Immunol. **17**, 593–623 (1999). doi:10.1146/annurev.immunol.17.1.593

[20] Skaar, E.P.: The battle for iron between bacterial pathogens and their vertebrate hosts. PLoS Pathog. **6**(8), 1000949 (2010). doi:10.1371/journal.ppat.1000949

[21] Schluger, N., Rom, W.: The Host Immune Response to Tuberculosis. Am. J. Respir. Crit. Care Med. **157**(19), 679–691 (1998)

[22] Clemens, D., Horwitz, M.: Characterization of the Mycobacterium tuberculosis Phagosome and Evidence that Phagosomal Maturation is Inhibited. J. Exp. Med. **181**(January), 257–270 (1995)

[23] Sakamoto, K.: The pathology of Mycobacterium tuberculosis infection. Vet. Pathol. **49**(3), 423–39 (2012). doi:10.1177/0300985811429313

[24] Aly, S., Wagner, K., Keller, C.: Oxygen status of lung granulomas in Mycobacterium tuberculosis-infected mice. J. Pathol. **210**(September), 298–305 (2006). doi:10.1002/path

[25] Schnappinger, D., Ehrt, S., Voskuil, M.I., Liu, Y., Mangan, J.a., Monahan, I.M., Dolganov, G., Efron, B., Butcher, P.D., Nathan, C., Schoolnik, G.K.: Transcriptional Adaptation of Mycobacterium tuberculosis within Macrophages: In-

sights into the Phagosomal Environment. J. Exp. Med. **198**(5), 693–704 (2003). doi:10.1084/jem.20030846

[26] Cox, J.S., Chen, B., McNeil, M., Jacobs, W.R.: Complex lipid determines tissue-specific replication of Mycobacterium tuberculosis in mice. Nature **402**(6757), 79–83 (1999). doi:10.1038/47042

[27] Reed, M.B., Domenech, P., Manca, C., Su, H., Barczak, A.K., Kreiswirth, B.N., Kaplan, G., Barry, C.E.: A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. Nature **431**(7004), 84–7 (2004). doi:10.1038/nature02837

[28] Muñoz-Elías, E.J., Timm, J., Botha, T., Chan, W.-t., Gomez, J.E., Mckinney, J.D., Mun, E.J.: Replication Dynamics of Mycobacterium tuberculosis in Chronically Infected Mice Replication Dynamics of Mycobacterium tuberculosis in Chronically Infected Mice. Infect. Immun. **73**(1), 546–551 (2005). doi:10.1128/IAI.73.1.546

[29] Gill, W.P., Harik, N.S., Whiddon, M.R., Liao, R.P., Mittler, J.E., Sherman, D.R.: A replication clock for Mycobacterium tuberculosis. Nat. Med. **15**(2), 211–4 (2009). doi:10.1038/nm.1915

[30] Gillespie, S.: Evolution of drug resistance in Mycobacterium tuberculosis: clinical and molecular perspective. Antimicrob. Agents Chemother. **46**(2), 267–274 (2002). doi:10.1128/AAC.46.2.267

[31] Milano, A., Pasca, M.R., Provvedi, R., Lucarelli, A.P., Manina, G., Luisa de Jesus Lopes Ribeiro, A., Manganelli, R., Riccardi, G.: Azole resistance in Mycobacterium tuberculosis is mediated by the MmpS5-MmpL5 efflux system. Tuberculosis **89**(1), 84–90 (2009). doi:10.1016/j.tube.2008.08.003

[32] Nicol, M., Sola, C., February, B.: Distribution of Strain Families of Mycobacterium tuberculosis Causing Pulmonary and Extrapulmonary Disease in Hospitalized Children in Cape Town, South Africa. J. Clin. Microbiol. **43**(11), 5779–5781 (2005). doi:10.1128/JCM.43.11.5779

[33] Cowley, D., Govender, D., February, B., Wolfe, M., Steyn, L., Evans, J., Wilkinson, R.J., Nicol, M.P.: Recent and rapid emergence of W-Beijing strains of Mycobacterium tuberculosis in Cape Town, South Africa. Clin. Infect. Dis. **47**(10), 1252–9 (2008). doi:10.1086/592575

[34] Hanekom, M., van der Spuy, G.D., Streicher, E., Ndabambi, S.L., McEvoy, C.R.E., Kidd, M., Beyers, N., Victor, T.C., van Helden, P.D., Warren, R.M.: A recently evolved sublineage of the Mycobacterium tuberculosis Beijing strain family is associated with an increased ability to spread and cause disease. J. Clin. Microbiol. **45**(5), 1483–90 (2007). doi:10.1128/JCM.02191-06

[35] Buu, T.N., van Soolingen, D., Huyen, M.N.T., Lan, N.T.N., Quy, H.T., Tiemersma, E.W., Kremer, K., Borgdorff, M.W., Cobelens, F.G.J.: Increased transmission of Mycobacterium tuberculosis Beijing genotype strains associated with resistance to streptomycin: a population-based study. PLoS One **7**(8), 42323 (2012). doi:10.1371/journal.pone.0042323

[36] Glynn, J., Kremer, K.: Beijing/W genotype Mycobacterium tuberculosis and drug resistance. Emerg. Infect. Dis. **12**(5), 736–743 (2006)

[37] Merker, M., Kohl, T.a., Roetzer, A., Truebe, L., Richter, E., Rüsch-Gerdes, S., Fattorini, L., Oggioni, M.R., Cox, H., Varaine, F., Niemann, S.: Whole genome sequencing reveals complex evolution patterns of multidrug-resistant Mycobacterium tuberculosis Beijing strains in patients. PLoS One **8**(12), 82551 (2013). doi:10.1371/journal.pone.0082551

[38] Ford, C.B., Shah, R.R., Maeda, M.K., Gagneux, S., Murray, M.B., Cohen, T., Johnston, J.C., Gardy, J., Lipsitch, M., Fortune, S.M.: Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat. Genet. **45**(7), 784–790 (2013). doi:10.1038/ng.2656

[39] Ebrahimi-Rad, M., Bifani, P., Martin, C., Kremer, K., Samper, S., Rauzier, J., Kreiswirth, B., Blazquez, J., Jouan, M., van Soolingen, D., Gicquel, B.: Mutations

in putative mutator genes of Mycobacterium tuberculosis strains of the W-Beijing family. Emerg. Infect. Dis. **9**(7), 838–45 (2003). doi:10.3201/eid0907.020589

[40] Ford, C.B., Shah, R.R., Maeda, M.K., Gagneux, S., Murray, B., Cohen, T., Johnston, J.C., Gardy, J., Lipsitch, M.: Emergence of Drug Resistant Tuberculosis **45**(7), 784–790 (2014). doi:10.1038/ng.2656.Mycobacterium

[41] Shell, S.S., Prestwich, E.G., Baek, S.-H., Shah, R.R., Sassetti, C.M., Dedon, P.C., Fortune, S.M.: DNA methylation impacts gene expression and ensures hypoxic survival of Mycobacterium tuberculosis. PLoS Pathog. **9**(7), 1003419 (2013). doi:10.1371/journal.ppat.1003419

[42] Cole, S., Brosch, R., Parkhill, J., Garnier, T.: Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature **396**(NOVEMBER), 537–544 (1998)

[43] Galagan, J.E., Sisk, P., Stolte, C., Weiner, B., Koehrsen, M., Wymore, F., Reddy, T.B.K., Zucker, J.D., Engels, R., Gellesch, M., Hubble, J., Jin, H., Larson, L., Mao, M., Nitzberg, M., White, J., Zachariah, Z.K., Sherlock, G., Ball, C.a., Schoolnik, G.K.: TB database 2010: overview and update. Tuberculosis (Edinb). **90**(4), 225–35 (2010). doi:10.1016/j.tube.2010.03.010

[44] and MIT, B.I.o.H.: Mycobacterium tuberculosis Comparative Sequencing Project. http://www.broadinstitute.org/

[45] Galperin, M.Y., Koonin, E.V.: Searching for drug targets in microbial genomes. Curr. Opin. Biotechnol. **10**(6), 571–8 (1999)

[46] Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S., Musser, J.M.: Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. Proc. Natl. Acad. Sci. U. S. A. **94**(18), 9869–74 (1997)

[47] Borrell, S., Gagneux, S.: Strain diversity, epistatis and the evolution of drug resistance in Mycobacterium tuberculosis. Clin. Microbiol. Infect. **17**(6), 815–820 (2011). doi:10.1111/j.1469-0691.2011.03556.x.Strain

[48] Fleischmann, R.D., Alland, D., Eisen, J.a., Carpenter, L., White, O., Peterson, J., Deboy, R., Dodson, R., Gwinn, M., Haft, D., Hickey, E., Kolonay, J.F., Nelson, W.C., Umayam, L.a., Ermolaeva, M., Salzberg, S.L., Delcher, a., Utterback, T., Weidman, J., Khouri, H., Gill, J., Mikula, a., Bishai, W., Venter, J.C., Fraser, C.M., Jacobs, W.R.: Whole-Genome Comparison of Mycobacterium tuberculosis Clinical and Laboratory Strains Whole-Genome Comparison of Mycobacterium tuberculosis Clinical and Laboratory Strains. J. Bacteriol. **184**(19), 5479–5490 (2002). doi:10.1128/JB.184.19.5479

[49] Hughes, A.L., Friedman, R., Murray, M.: Genomewide pattern of synonymous nucleotide substitution in two complete genomes of Mycobacterium tuberculosis. Emerg. Infect. Dis. **8**(11), 1342–1346 (2002). doi:10.3201/eid0811.020064

[50] Shitikov, E.A., Bespyatykh, J.A., Ischenko, D.S., Alexeev, D.G., Karpova, I.Y., Kostryukova, E.S., Isaeva, Y.D., Nosova, E.Y., Mokrousov, I.V., Vyazovaya, A.a., Narvskaya, O.V., Vishnevsky, B.I., Otten, T.F., Zhuravlev, V.I., Zhuravlev, V.Y., Yablonsky, P.K., Ilina, E.N., Govorun, V.M.: Unusual large-scale chromosomal rearrangements in Mycobacterium tuberculosis Beijing B0/W148 cluster isolates. PLoS One **9**(1), 84971 (2014). doi:10.1371/journal.pone.0084971

[51] Alland, D., Lacher, D.W., Hazbón, M.H., Motiwala, A.S., Qi, W., Fleischmann, R.D., Whittam, T.S.: Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of Mycobacterium tuberculosis and the utility of LSPs in phylogenetic analysis. J. Clin. Microbiol. **45**(1), 39–46 (2007). doi:10.1128/JCM.02483-05

[52] Tsolaki, A.G., Hirsh, A.E., Deriemer, K., Enciso, J.A., Wong, M.Z., Hannan, M., Salmoniere, Y.-o.L.G.D., Aman, K., Kato-maeda, M., Small, P.M.: Functional and evolutionary genomics of Mycobacterium tuberculosis: insights from genomic deletions in 100 strains. Proc. Natl. Acad. Sci. **101**(14), 4865–4870 (2004)

[53] Merhej, V., Raoult, D.: Rickettsial evolution in the light of comparative genomics. Biol. Rev. **86**(2), 379–405 (2011). doi:10.1111/j.1469-185X.2010.00151.x

[54] Moran, N.A.: Microbial minimalism: Genome reduction in bacterial pathogens. Cell **108**(5), 583–586 (2002). doi:10.1016/S0092-8674(02)00665-7

[55] Reed, M.B., Gagneux, S., DeRiemer, K., Small, P.M., Barry, C.E.: The W-Beijing Lineage of Mycobacterium tuberculosis Overproduces Triglycerides and Has the DosR Dormancy Regulon Constitutively Upregulated. J. Bacteriol. **189**(7), 2583–2589 (2007). doi:10.1128/JB.01670-06

[56] Domenech, P., Kolly, G.S., Leon-Solis, L., Fallow, A., Reed, M.B.: Massive Gene Duplication Event among Clinical Isolates of the Mycobacterium tuberculosis W/Beijing Family. J. Bacteriol. **192**(18), 4562–4570 (2010). doi:10.1128/JB.00536-10

[57] Weiner, B., Gomez, J., Victor, T.C., Warren, R.M., Sloutsky, A., Plikaytis, B.B., Posey, J.E., van Helden, P.D., Gey van Pittius, N.C., Koehrsen, M., Sisk, P., Stolte, C., White, J., Gagneux, S., Birren, B., Hung, D., Murray, M., Galagan, J.: Independent large scale duplications in multiple M. Tuberculosis lineages overlapping the same genomic region. PLoS One **7**(2) (2012). doi:10.1371/journal.pone.0026038

[58] Becq, J., Gutierrez, M.C., Rosas-Magallanes, V., Rauzier, J., Gicquel, B., Neyrolles, O., Deschavanne, P.: Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli. Mol. Biol. Evol. **24**(8), 1861–71 (2007). doi:10.1093/molbev/msm111

[59] Hacker, J., Blum-Oehler, G., Muhldorfer, I., Tschape, H.: Pathogenicity islands of virulent bacteria: Structure, function and impact on microbial evolution. Mol. Microbiol. **23**(6), 1089–1097 (1997). doi:10.1046/j.1365-2958.1997.3101672.x

[60] Thierry, D., Cave, M.D., Eisenach, K.D., Crawford, J.T., Bates, J.H., Gicquel, B., Guesdon, J.L.: IS6110, an IS-like element of Mycobacterium tuberculosis complex. Nucleic Acids Res. **18**(1), 188 (1990)

[61] Coros, A., DeConno, E., Derbyshire, K.M.: IS6110, a Mycobacterium tuberculosis complex-specific insertion sequence, is also present in the genome of Mycobacterium smegmatis, suggestive of lateral gene transfer among mycobacterial species. J. Bacteriol. **190**(9), 3408–10 (2008). doi:10.1128/JB.00009-08

[62] Van Embden, J.D.A., Cave, M.D., Crawford, J.T., Dale, J.W., Eisenach, K.D., Gicquel, B., Hermans, P., Martin, C., McAdam, R., Shinnick, T.M., Small, P.M.: Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: Recommendations for a standardized methodology. J. Clin. Microbiol. **31**(2), 406–409 (1993). doi:10.1128/JCM.39.4.1683.2001

[63] Sekine, Y., Eisaki, N., Ohtsubo, E.: Translational Control in Production of Transposase and in Transposition of Insertion Sequence IS3. J. Mol. Biol. **235**(5), 1406–1420 (1994). doi:10.1006/jmbi.1994.1097

[64] McEvoy, C.R.E., Falmer, A.a., Gey van Pittius, N.C., Victor, T.C., van Helden, P.D., Warren, R.M.: The role of IS6110 in the evolution of Mycobacterium tuberculosis. Tuberculosis (Edinb). **87**(5), 393–404 (2007). doi:10.1016/j.tube.2007.05.010

[65] Kidwell, M.G., Lisch, D.R.: Transposable elements and host genome evolution. TREE **15**(March), 95–99 (2000)

[66] Le Rouzic, A., Dupas, S., Capy, P.: Genome ecosystem and transposable elements species. Gene **390**, 214–220 (2007). doi:10.1016/j.gene.2006.09.023

[67] Le Rouzic, A., Dupas, S., Capy, P.: Genome ecosystem and transposable elements species. Gene **390**(1-2), 214–220 (2007). doi:10.1016/j.gene.2006.09.023

[68] Beggs, M.L., Eisenach, K.D., Cave, M.D.: Mapping of IS6110 insertion sites in two epidemic strains of Mycobacterium tuberculosis. J. Clin. Microbiol. **38**(8), 2923–2928 (2000)

[69] Hermans, P.W.M., Van Soolingen, D., Bik, E.M., De Haas, P.E.W., Dale, J.W., Van Embden, J.D.A.: Insertion element IS987 from Mycobacterium bovis BCG is located

in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains. Infect. Immun. **59**(8), 2695–2705 (1991)

[70] Banu, S., Honoré, N., Saint-Joanis, B., Philpott, D., Prévost, M.C., Cole, S.T.: Are the PE-PGRS proteins of Mycobacterium tuberculosis variable surface antigens? Mol. Microbiol. **44**(1), 9–19 (2002). doi:10.1046/j.1365-2958.2002.02813.x

[71] Soto, C.Y., Menéndez, M.C., Pérez, E., Samper, S., Gómez, A.B., García, M.J., Martín, C.: IS6110 Mediates Increased Transcription of the phoP Virulence Gene in a Multidrug-Resistant Clinical Isolate Responsible for Tuberculosis Outbreaks. J. Clin. Microbiol. **42**(1), 212–219 (2004). doi:10.1128/JCM.42.1.212-219.2004

[72] Lemaitre, N., Sougakoff, W., Truffot-Pernot, C., Jarlier, V.: Characterization of New Mutations in Pyrazinamide-Resistant Strains of Mycobacterium tuberculosisand Identification of Conserved Regions Important for the Catalytic Activity of the Pyrazinamidase {PncA}. Antimicrob. Agents Chemother. **43**(7), 1761–1763 (1999)

[73] Arnvig, K.B., 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.: Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of Mycobacterium tuberculosis. PLoS Pathog. **7**(11), 1002342 (2011). doi:10.1371/journal.ppat.1002342

[74] Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D., Betel, D.: Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. **14**(9), 95 (2013). doi:10.1186/gb-2013-14-9-r95

[75] Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome Biol. **11**(R106), 1–12 (2010). doi:10.1186/gb-2010-11-10-r106. 1310.0424

[76] Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–40 (2010). doi:10.1093/bioinformatics/btp616

[77] Hardcastle, T.J.: baySeq: Empirical Bayesian analysis of patterns of differential expression in count data. R package version 2.0.50, 1–13 (2012)

[78] Washburn, M.P., Koller, A., Oshiro, G., Ulaszek, R.R., Plouffe, D., Deciu, C., Winzeler, E., Yates, J.R.: Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S. A. **100**(6), 3107–3112 (2003). doi:10.1073/pnas.0634629100

[79] Nie, L., Wu, G., Zhang, W.: Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in Desulfovibrio vulgaris: A quantitative analysis. Genetics **174**(4), 2229–2243 (2006). doi:10.1534/genetics.106.065862

[80] Cortes, T., Schubert, O.T., Banaei-Esfahani, A., Collins, B.C., Aebersold, R., Young, D.B.: Delayed effects of transcriptional responses in Mycobacterium tuberculosis exposed to nitric oxide suggest other mechanisms involved in survival. Sci. Rep. **7**(1), 8208 (2017). doi:10.1038/s41598-017-08306-1

[81] Breaker, R.R.: Riboswitches and the RNA world. Cold Spring Harb. Perspect. Biol. **4**(2), 1–15 (2012). doi:10.1101/cshperspect.a003566

[82] Beiter, T., Reich, E., Williams, R.W., Simon, P.: Antisense transcription: a critical look in both directions. Cell. Mol. Life Sci. **66**(1), 94–112 (2009). doi:10.1007/s00018-008-8381-y

[83] Arnvig, K., Young, D.: Non-coding RNA and its potential role in Mycobacterium tuberculosis pathogenesis. RNA Biol. **9**(4), 427–36 (2012). doi:10.4161/rna.20105

[84] Georg, J., Hess, W.R.: cis-Antisense RNA, Another Level of Gene Regulation in Bacteria. Microbiol. Mol. Biol. Rev. **75**(2), 286–300 (2011). doi:10.1128/MMBR.00032-10

[85] Kawano, M., Aravind, L., Storz, G.: An antisense RNA controls synthesis of an SOS-induced toxin evolved from an antitoxin. Mol. Microbiol. **64**(3), 738–754 (2007). doi:10.1111/j.1365-2958.2007.05688.x

[86] Nahvi, A., Sudarsan, N., Ebert, M.S., Zou, X., Brown, K.L., Breaker, R.R.: Genetic control by a metabolite binding mRNA. Chem. Biol. **9**(9), 1043–1049 (2002). doi:10.1016/S1074-5521(02)00224-7

[87] Babitzke, P., Romeo, T.: CsrB sRNA family: sequestration of RNA-binding regulatory proteins. Curr. Opin. Microbiol. **10**(2), 156–63 (2007). doi:10.1016/j.mib.2007.03.007

[88] Gripenland, J., Netterling, S., Loh, E., Tiensuu, T., Toledo-Arana, A., Johansson, J.: RNAs: regulators of bacterial virulence. Nat. Rev. Microbiol. **8**(12), 857–66 (2010). doi:10.1038/nrmicro2457

[89] Waters, L.S., Storz, G.: Regulatory RNAs in Bacteria. Cell **136**(4), 615–628 (2009). doi:10.1016/j.cell.2009.01.043. NIHMS150003

[90] Komorowski, M., Mikisz, J., Kierzek, A.M.: Translational repression contributes greater noise to gene expression than transcriptional repression. Biophys. J. **96**(2), 372–384 (2009). doi:10.1016/j.bpj.2008.09.052

[91] Smollett, K.L., Smith, K.M., Kahramanoglou, C., Arnvig, K.B., Buxton, R.S., Davis, E.O.: Global analysis of the regulon of the transcriptional repressor LexA, a key component of SOS response in Mycobacterium tuberculosis. J. Biol. Chem. **287**(26), 22004–22014 (2012). doi:10.1074/jbc.M112.357715

[92] Papenfort, K., Pfeiffer, V., Mika, F., Lucchini, S., Hinton, J.C.D., Vogel, J.: σE-dependent small RNAs of Salmonella respond to membrane stress by accelerating global omp mRNA decay. Mol. Microbiol. **62**(6), 1674–1688 (2006). doi:10.1111/j.1365-2958.2006.05524.x

[93] Arnvig, K.B., Young, D.B.: Identification of small RNAs in Mycobacterium tuberculosis. Mol. Microbiol. **73**(3), 397–408 (2009). doi:10.1111/j.1365-2958.2009.06777.x

[94] Svenningsen, S.L., Tu, K.C., Bassler, B.L.: Gene dosage compensation calibrates four regulatory RNAs to control Vibrio cholerae quorum sensing. EMBO J. **28**(4), 429–439 (2009). doi:10.1038/emboj.2008.300

[95] Mandin, P., Gottesman, S.: Integrating anaerobic/aerobic sensing and the general stress response through the ArcZ small RNA. EMBO J. **29**(18), 3094–107 (2010). doi:10.1038/emboj.2010.179

[96] Monteiro, C., Papenfort, K., Hentrich, K., Ahmad, I., Le Guyon, S., Reimann, R., Grantcharova, N., Römling, U.: Hfq and Hfq-dependent small RNAs are major contributors to multicellular development in Salmonella enterica serovar Typhimurium. RNA Biol. **9**(4), 489–502 (2012). doi:10.4161/rna.19682

[97] Gottesman, S.: Micros for microbes: Non-coding regulatory RNAs in bacteria. Trends Genet. **21**(7), 399–404 (2005). doi:10.1016/j.tig.2005.05.008

[98] Sun, X., Zhulin, I., Wartell, R.: Predicted structure and phyletic distribution of the RNAbinding protein Hfq. Nucleic Acids Res. **30**(17), 3662–3671 (2002)

[99] Sledjeski, D.D., Whitman, C., Zhang, A.: Hfq is necessary for regulation by the untranslated RNA DsrA. J. Bacteriol. **183**(6), 1997–2005 (2001). doi:10.1128/JB.183.6.1997-2005.2001

[100] Chao, Y., Vogel, J.: The role of Hfq in bacterial pathogens. Curr. Opin. Microbiol. **13**(1), 24–33 (2010). doi:10.1016/j.mib.2010.01.001

[101] Lasa, I., Toledo-Arana, A., Dobin, A., Villanueva, M., de los Mozos, I.R., Vergara-Irigaray, M., Segura, V., Fagegaltier, D., Penades, J., Valle, J., Solano, C., Gingeras, T.R.: Genome-wide antisense transcription drives mRNA processing in bacteria. Proc. Natl. Acad. Sci. **108**, 20172–20177 (2011). doi:10.1073/pnas.1113521108. arXiv:1408.1149

[102] Akey, D.L., Berger, J.M.: Structure of the nuclease domain of ribonuclease III from M. tuberculosis at 2.1 A. Protein Sci. **14**(10), 2744–2750 (2005). doi:10.1110/ps.051665905

[103] Pichon, C., Felden, B.: Small RNA gene identification and mRNA target predictions in bacteria. Bioinformatics **24**(24), 2807–2813 (2008). doi:10.1093/bioinformatics/btn560

[104] Rivas, E., Eddy, S.R.: Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics **2**, 8 (2001). doi:10.1186/1471-2105-2-8

[105] Washietl, S., Hofacker, I.L., Stadler, P.F.: Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. U. S. A. **102**(7), 2454–9 (2005). doi:10.1073/pnas.0409169102

[106] Wang, C., Ding, C., Meraz, R.F., Holbrook, S.R.: PSoL: A positive sample only learning algorithm for finding non-coding RNA genes. Bioinformatics **22**(21), 2590–2596 (2006). doi:10.1093/bioinformatics/btl441

[107] Carter, R.J., Dubchak, I., Holbrook, S.R.: A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Res. **29**(19), 3928–3938 (2001). doi:10.1093/nar/29.19.3928

[108] Pichon, C., Felden, B.: Intergenic sequence inspector: Searching and identifying bacterial RNAs. Bioinformatics **19**(13), 1707–1709 (2003). doi:10.1093/bioinformatics/btg235

[109] Rivas, E., Eddy, S.R.: Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics **16**(7), 583–605 (2000)

[110] Livny, J., Fogel, M.A., Davis, B.M., Waldor, M.K.: sRNAPredict: An integrative computational approach to identify sRNAs in bacterial genomes. Nucleic Acids Res. **33**(13), 4096–4105 (2005). doi:10.1093/nar/gki715

[111] DiChiara, J.M., Contreras-Martinez, L.M., Livny, J., Smith, D., McDonough, K.a., Belfort, M.: Multiple small RNAs identified in Mycobacterium bovis BCG are also expressed in Mycobacterium tuberculosis and Mycobacterium smegmatis. Nucleic Acids Res. **38**(12), 4067–4078 (2010). doi:10.1093/nar/gkq101

[112] Arnvig, K.B., Young, D.B.: Non-coding RNA and its potential role in Mycobacterium tuberculosis pathogenesis. RNA Biol. **9**(April), 427–436 (2012)

[113] Vogel, J., Bartels, V., Tang, T.H., Churakov, G., Slagter-Jäger, J.G., Hüttenhofer, A., Wagner, E.G.H.: RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. Nucleic Acids Res. **31**(22), 6435–6443 (2003). doi:10.1093/nar/gkg867

[114] Jeeves, R.E., Marriott, A.A.N., Pullan, S.T., Hatch, K.A., Allnutt, J.C., Freire-Martin, I., Hendon-Dunn, C.L., Watson, R., Witney, A.A., Tyler, R.H., Arnold, C., Marsh, P.D., McHugh, T.D., Bacon, J.: Mycobacterium tuberculosis is resistant to isoniazid at a slow growth rate by single nucleotide polymorphisms in katg codon ser315. PLoS One **10**(9), 1–21 (2015). doi:10.1371/journal.pone.0138253

[115] Abramovitch, R.B., Rohde, K.H., Hsu, F.-F., Russell, D.G.: aprABC: a Mycobacterium tuberculosis complex-specific locus that modulates pH-driven adaptation to the macrophage phagosome. Mol. Microbiol. **80**(3), 678–694 (2011). doi:10.1111/j.1365-2958.2011.07601.x. NIHMS150003

[116] Wang, J., Rennie, W., Liu, C., Carmack, C.S., Prévost, K., Caron, M.P., Massé, E., Ding, Y., Wade, J.T.: Identification of bacterial sRNA regulatory targets using ribosome profiling. Nucleic Acids Res. **43**(21), 10308–10320 (2015). doi:10.1093/nar/gkv1158

[117] Antal, M., Bordeau, V., Douchin, V., Felden, B.: A small bacterial RNA regulates a putative ABC transporter. J. Biol. Chem. **280**(9), 7901–7908 (2005). doi:10.1074/jbc.M413071200

[118] Pain, A., Ott, A., Amine, H., Rochat, T., Bouloc, P., Gautheret, D.: An assessment of bacterial small RNA target prediction programs. RNA Biol. **12**(5), 509–13 (2015). doi:10.1080/15476286.2015.1020269

[119] Wright, P.R., Georg, J., Mann, M., Sorescu, D.A., Richter, A.S., Lott, S., Kleinkauf, R., Hess, W.R., Backofen, R.: CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains. Nucleic Acids Res. **42**(Web Server issue), 119–23 (2014). doi:10.1093/nar/gku359

[120] Tafer, H., Hofacker, I.L.: RNAplex: a fast tool for RNA-RNA interaction search. Bioinformatics **24**(22), 2657–63 (2008). doi:10.1093/bioinformatics/btn193

[121] Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S.H., Stadler, P.F., Hofacker, I.L.: Thermodynamics of RNA-RNA binding. Bioinformatics **22**(10), 1177–82 (2006). doi:10.1093/bioinformatics/btl024

[122] Mazandu, G., Mulder, N.: Using the underlying biological organization of the Mycobacterium tuberculosis functional network for protein function prediction. Infect. Genet. Evol. **12**(5), 922–932 (2012). doi:10.1016/j.meegid.2011.10.027

[123] Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., von Mering, C.: STRING 8 - A global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. **37**(SUPPL. 1), 412–416 (2009). doi:10.1093/nar/gkn760

[124] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M.: KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. **44**(D1), 457–462 (2016). doi:10.1093/nar/gkv1070

[125] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K.: KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. **45**(D1), 353–361 (2017). doi:10.1093/nar/gkw1092. 1611.06654

[126] Vailaya, A., Bluvas, P., Kincaid, R., Kuchinsky, A., Creech, M., Adler, A.: An architecture for biological information extraction and representation. Bioinformatics **21**(4), 430–438 (2005). doi:10.1093/bioinformatics/bti187

[127] Mazandu, G.K., Mulder, N.J.: Generation and Analysis of Large-Scale Data-Driven Mycobacterium tuberculosis Functional Networks for Drug Target Identification. Adv. Bioinformatics **2011**, 801478 (2011). doi:10.1155/2011/801478

[128] Galagan, J.E., Minch, K., Peterson, M., Lyubetskaya, A., Azizi, E., Sweet, L., Gomes, A., Rustad, T., Dolganov, G., Glotova, I., Abeel, T., Mahwinney, C., Kennedy, A.D.,

Allard, R., Brabant, W., Krueger, A., Jaini, S., Honda, B., Yu, W.-H., Hickey, M.J., Zucker, J., Garay, C., Weiner, B., Sisk, P., Stolte, C., Winkler, J.K., Van de Peer, Y., Iazzetti, P., Camacho, D., Dreyfuss, J., Liu, Y., Dorhoi, A., Mollenkopf, H.-J., Drogaris, P., Lamontagne, J., Zhou, Y., Piquenot, J., Park, S.T., Raman, S., Kaufmann, S.H.E., Mohney, R.P., Chelsky, D., Moody, D.B., Sherman, D.R., Schoolnik, G.K.: The Mycobacterium tuberculosis regulatory network and hypoxia. Nature **499**(7457), 178–83 (2013). doi:10.1038/nature12337

[129] Kim, Y., Kim, T.-K., Kim, Y., Yoo, J., You, S., Lee, I., Carlson, G., Hood, L., Choi, S., Hwang, D.: Principal network analysis: identification of subnetworks representing major dynamics using gene expression data. Bioinformatics **27**(3), 391–8 (2011). doi:10.1093/bioinformatics/btq670

[130] Smolen, P., Baxter, D.A., Byrne, J.H.: Modeling Transcriptional Control in Gene Networks-Methods, Recent Results, and Future Directions. Bull. Math. Biol. **62**(2), 247–292 (2000). doi:10.1006/bulm.1999.0155

[131] Hagberg, A.A., Schult, D.A., Swart, P.J.: Exploring network structure, dynamics, and function using NetworkX. Proc. 7th Python Sci. Conf. (SciPy 2008) (SciPy), 11–15 (2008)

[132] Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A.R., Vailaya, A., Wang, P.-L., Adler, A., Conklin, B.R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G.J., Ideker, T., Bader, G.D.: Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc. **2**(10), 2366–2382 (2007). doi:10.1038/nprot.2007.324

[133] Beisser, D., Klau, G.W., Dandekar, T., Müller, T., Dittrich, M.T.: BioNet: An R-Package for the functional analysis of biological networks. Bioinformatics **26**(8), 1129–1130 (2010). doi:10.1093/bioinformatics/btq089

[134] Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-liggett, C., Knight, R., Gordon, J.I.: The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature **449**(7164), 804–810 (2007). doi:10.1038/nature06244.The

[135] Resistance, D.: crossm FIND Tuberculosis Strain Bank : a Resource for Researchers and Developers Working on Tests To Detect Mycobacterium tuberculosis and Related Drug Resistance, 1066–1073 (2017). doi:10.1128/JCM.01662-16

[136] Simpson, J., Wong, K., Jackman, S.: ABySS: a parallel assembler for short read sequence data. Genome **19**, 1117–1123 (2009). doi:10.1101/gr.089532.108.

[137] Zerbino, D.R., Birney, E.: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**(5), 821–9 (2008). doi:10.1101/gr.074492.107

[138] Marcus, S., Lee, H., Schatz, M.C.: SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips. Bioinformatics **30**(24), 3476–3483 (2014). doi:10.1093/bioinformatics/btu756

[139] Beller, T., Ohlebusch, E.: A representation of a compressed de Bruijn graph for pan-genome analysis that enables search. Algorithms Mol. Biol. **11**, 20 (2016). doi:10.1186/s13015-016-0083-7. 1602.03333

[140] Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., Haussler, D.: Cactus: Algorithms for genome multiple sequence alignment. Genome Res. **21**(9), 1512–1528 (2011). doi:10.1101/gr.123356.111

[141] VG Team: Variant Graph. https://github.com/vgteam/vg/

[142] Sheikhizadeh, S., Schranz, M.E., Akdel, M., de Ridder, D., Smit, S.: PanTools: representation, storage and exploration of pan-genomic data. Bioinformatics **32**(17), 487–493 (2016). doi:10.1093/bioinformatics/btw455

[143] Novak, A.M., Hickey, G., McVean Li Ka Shing, G.: Genome Graphs. Curoverse IncEuropean Bioinforma. InstituteUCSC Genomics Institute) Nancy Ouyang (Curoverse Inc Goran Rakocevic (94040), 1–26 (2017). doi:10.1101/101378

[144] Periwal, V., Patowary, A., Vellarikkal, S.K., Gupta, A., Singh, M., Mittal, A., Jeyapaul, S., Chauhan, R.K., Singh, A.V., Singh, P.K., Garg, P., Katoch, V.M., Katoch, K., Chauhan, D.S., Sivasubbu, S., Scaria, V.: Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of Mycobacterium tuberculosis pangenome. PLoS One **10**(4), 1–26 (2015). doi:10.1371/journal.pone.0122979

[145] Friedena, T.R., Sbarbarob, J.A.: Promoting adherence to treatment for tuberculosis: The importance of direct observation (2007). ISSN 0042-9686. doi:10.2471/BLT.06.038927

[146] Tezera, L.B., Bielecka, M.K., Chancellor, A., Reichmann, M.T., Shammari, B.A., Brace, P., Batty, A., Tocheva, A., Jogai, S., Marshall, B.G., Tebruegge, M., Jayasinghe, S.N., Mansour, S., Elkington, P.T.: Dissection of the host-pathogen interaction in human tuberculosis using a bioengineered 3-dimensional model. Elife **6**, 1–19 (2017). doi:10.7554/eLife.21283

[147] Darling, A.E., Mau, B., Perna, N.T.: Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. PLoS One **5**(6) (2010). doi:10.1371/journal.pone.0011147

[148] Edgar, R.C.: MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**(5), 1792–1797 (2004). doi:10.1093/nar/gkh340

[149] Katoh, K., Kuma, K.I., Toh, H., Miyata, T.: MAFFT version 5: Improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. **33**(2), 511–518 (2005). doi:10.1093/nar/gki198

[150] Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G.: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. **7**(1), 539 (2011). doi:10.1038/msb.2011.75

[151] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: A Software Environment for Integrated

Models of Biomolecular Interaction Networks. Genome Res. **13**(11), 2498–2504 (2003). doi:10.1101/gr.1239303

[152] Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**(14), 1754–1760 (2009). doi:10.1093/bioinformatics/btp324

[153] Andrews, S.: FastQC (2012). http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[154] Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**(1), 10 (2011). doi:10.14806/ej.17.1.200. ISSN 2226-6089

[155] Lew, J.M., Kapopoulou, A., Jones, L.M., Cole, S.T.: TubercuList - 10 years after. Tuberculosis **91**(1), 1–7 (2011). doi:10.1016/j.tube.2010.09.008

[156] Voskuil, M.I., Bartek, I.L., Visconti, K., Schoolnik, G.K.: The response of Mycobacterium tuberculosis to reactive oxygen and nitrogen species. Front. Microbiol. **2**(MAY), 1–12 (2011). doi:10.3389/fmicb.2011.00105

[157] Hampshire, T., Soneji, S., Bacon, J., James, B.W., Hinds, J., Laing, K., Stabler, R.A., Marsh, P.D., Butcher, P.D.: Stationary phase gene expression of Mycobacterium tuberculosis following a progressive nutrient depletion: a model for persistent organisms? Tuberculosis **84**(3-4), 228–238 (2004). doi:10.1016/j.tube.2003.12.010

[158] Moores, A., Riesco, A.B., Schwenk, S., Arnvig, K.B.: Expression, maturation and turnover of DrrS, an unusually stable, DosR regulated small RNA in Mycobacterium tuberculosis. PLoS One **12**(3), 1–27 (2017). doi:10.1371/journal.pone.0174079

[159] Hartkoorn, R.C., Sala, C., Uplekar, S., Busso, P., Rougemont, J., Cole, S.T.: Genome-Wide definition of the SigF regulon in Mycobacterium tuberculosis. J. Bacteriol. **194**(8), 2001–2009 (2012). doi:10.1128/JB.06692-11

[160] Rolfe, M.D., Rice, C.J., Lucchini, S., Pin, C., Thompson, A., Cameron, A.D.S., Alston, M., Stringer, M.F., Betts, R.P., Baranyi, J., Peck, M.W., Hinton, J.C.D.: Lag

Phase Is a Distinct Growth Phase That Prepares Bacteria for Exponential Growth and Involves Transient Metal Accumulation. J. Bacteriol. **194**(3), 686–701 (2012). doi:10.1128/JB.06112-11

[161] Dukan, S., Nyström, T.: Bacterial senescence: Stasis results in increased and differential oxidation of cytoplasmic proteins leading to developmental induction of the heat shock regulon. Genes Dev. **12**(21), 3431–3441 (1998). doi:10.1101/gad.12.21.3431

[162] Solans, L., Gonzalo-Asensio, J., Sala, C., Benjak, A., Uplekar, S., Rougemont, J., Guilhot, C., Malaga, W., Martin, C., Cole, S.T.: The PhoP-Dependent ncRNA Mcr7 Modulates the TAT Secretion System in Mycobacterium tuberculosis. PLoS Pathog. **10**(5) (2014). doi:10.1371/journal.ppat.1004183

[163] Lee, J.S., Krause, R., Schreiber, J., Mollenkopf, H.J., Kowall, J., Stein, R., Jeon, B.Y., Kwak, J.Y., Song, M.K., Patron, J.P., Jorg, S., Roh, K., Cho, S.N., Kaufmann, S.H.E.: Mutation in the Transcriptional Regulator PhoP Contributes to Avirulence of Mycobacterium tuberculosis H37Ra Strain. Cell Host Microbe **3**(2), 97–103 (2008). doi:10.1016/j.chom.2008.01.002

[164] Voskuil, M.I., Visconti, K.C., Schoolnik, G.K.: Mycobacterium tuberculosis gene expression during adaptation to stationary phase and low-oxygen dormancy. Tuberculosis **84**(3-4), 218–227 (2004). doi:10.1016/j.tube.2004.02.003

[165] Gaballa, A., Antelmann, H., Aguilar, C., Khakh, S.K., Song, K.-B., Smaldone, G.T., Helmann, J.D.: The Bacillus subtilis iron-sparing response is mediated by a Fur-regulated small RNA and three small, basic proteins. Proc. Natl. Acad. Sci. **105**(33), 11927–11932 (2008). doi:10.1073/pnas.0711752105

[166] Sherman, D.R., Voskuil, M., Schnappinger, D., Liao, R., Harrell, M.I., Schoolnik, G.K.: Regulation of the Mycobacterium tuberculosis hypoxic response gene encoding alpha -crystallin. Proc. Natl. Acad. Sci. U. S. A. **98**(13), 7534–9 (2001). doi:10.1073/pnas.121172498

[167] Sun, R., Converse, P.J., Ko, C., Tyagi, S., Morrison, N.E., Bishai, W.R.: Mycobacterium tuberculosis ECF sigma factor sigC is required for lethality in mice and for the conditional expression of a defined gene set. Mol. Microbiol. **52**(1), 25–38 (2004). doi:10.1111/j.1365-2958.2003.03958.x

[168] Cowley, S., Ko, M., Pick, N., Chow, R., Downing, K.J., Gordhan, B.G., Betts, J.C., Mizrahi, V., Smith, D.A., Stokes, R.W., Av-Gay, Y.: The Mycobacterium tuberculosis protein serine/threonine kinase PknG is linked to cellular glutamate/glutamine levels and is important for growth in vivo. Mol. Microbiol. **52**(6), 1691–1702 (2004). doi:10.1111/j.1365-2958.2004.04085.x

[169] England, K., Crew, R., Slayden, R.A.: Mycobacterium tuberculosis septum site determining protein, Ssd encoded by rv3660c, promotes filamentation and elicits an alternative metabolic and dormancy stress response. BMC Microbiol. **11**(1), 79 (2011). doi:10.1186/1471-2180-11-79

[170] Singh, A., Varela, C., Bhatt, K., Veerapen, N., Lee, O.Y.C., Wu, H.H.T., Besra, G.S., Minnikin, D.E., Fujiwara, N., Teramoto, K., Bhatt, A.: Identification of a desaturase involved in mycolic acid biosynthesis in Mycobacterium Smegmatis. PLoS One **11**(10), 1–19 (2016). doi:10.1371/journal.pone.0164253

[171] Fernandes, N.D., Wu, Q.-l., Kong, D., Garg, S., Husson, R.N.: A Mycobacterial Extracytoplasmic Sigma Factor Involved in Survival following Heat Shock and Oxidative Stress A Mycobacterial Extracytoplasmic Sigma Factor Involved in Survival following Heat Shock and Oxidative Stress. J. Bacteriol. **181**(14), 4266–4274 (1999)

[172] Raman, S., Song, T., Puyang, X., Bardarov, S., Jacobs, J., Husson, R.N.: The alternative sigma factor sigh regulates major components of oxidative and heat stress responses in Mycobacterium tuberculosis. J. Bacteriol. **183**(20), 6119–6125 (2001). doi:10.1128/JB.183.20.6119-6125.2001

[173] Bartek, I.L., Woolhiser, L.K., Baughn, a.D., Basaraba, R.J., Jacobs, W.R., Lenaerts, a.J., Voskuil, M.I.: Mycobacterium tuberculosis Lsr2 Is a Global Transcriptional Regulator. MBio **5**(3), 01106–14 (2014). doi:10.1128/mBio.01106-14.Editor

[174] Singh, A., Guidry, L., Narasimhulu, K.V., Mai, D., Trombley, J., Redding, K.E., Giles, G.I., Lancaster, J.R., Steyn, A.J.C.: Mycobacterium tuberculosis WhiB3 responds to O2 and nitric oxide via its [4Fe-4S] cluster and is essential for nutrient starvation survival. Proc. Natl. Acad. Sci. U. S. A. **104**(28), 11562–11567 (2007). doi:10.1073/pnas.0700490104

[175] Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R.: The variant call format and VCFtools. Bioinformatics **27**(15), 2156–2158 (2011). doi:10.1093/bioinformatics/btr330. NIHMS150003

[176] Grant, C.E., Bailey, T.L., Noble, W.S.: FIMO: Scanning for occurrences of a given motif. Bioinformatics **27**(7), 1017–1018 (2011). doi:10.1093/bioinformatics/btr064. PMC3065696

[177] Minch, K.J., Rustad, T.R., Peterson, E.J.R., Winkler, J., Reiss, D.J., Ma, S., Hickey, M., Brabant, W., Morrison, B., Turkarslan, S., Mawhinney, C., Galagan, J.E., Price, N.D., Baliga, N.S., Sherman, D.R.: The DNA-binding network of Mycobacterium tuberculosis. Nat. Commun. **6**, 5829 (2015). doi:10.1038/ncomms6829. 9809069v1

[178] Sassetti, C.M., Boyd, D.H., Rubin, E.J.: Genes required for mycobacterial growth defined by high density mutagenesis. Mol. Microbiol. **48**(1), 77–84 (2003)

[179] Masiewicz, P., Brzostek, A., Wolański, M., Dziadek, J., Zakrzewska-Czerwińska, J.: A novel role of the PrpR as a transcription factor involved in the regulation of methylcitrate pathway in Mycobacterium tuberculosis. PLoS One **7**(8) (2012). doi:10.1371/journal.pone.0043651

[180] Stapleton, M.R., Smith, L.J., Hunt, D.M., Buxton, R.S., Green, J.: Mycobacterium tuberculosis WhiB1 represses transcription of the essential chaperonin GroEL2. Tuberculosis **92**(4), 328–332 (2012). doi:10.1016/j.tube.2012.03.001

[181] Smith, L.J., Stapleton, M.R., Fullstone, G.J.M., Crack, J.C., Thomson, A.J., Le Brun, N.E., Hunt, D.M., Harvey, E., Adinolfi, S., Buxton, R.S., Green, J.: Mycobacterium

tuberculosis WhiB1 is an essential DNA-binding protein with a nitric oxide-sensitive ironsulfur cluster. Biochem. J. **432**(3), 417–427 (2010). doi:10.1042/BJ20101440

[182] Lewthwaite, J.C., Coates, A.R.M., Tormay, P., Singh, M., Mascagni, P., Poole, S., Sharp, L., Henderson, B., Lewthwaite, J.O.C.: Mycobacterium tuberculosis Chaperonin 60 . 1 Is a More Potent Cytokine Stimulator a CD14-Binding Domain Mycobacterium tuberculosis Chaperonin 60 . 1 Is a More Potent Cytokine Stimulator than Chaperonin 60 . 2 (Hsp 65) and Contains a CD14-Binding Domain **2**(Hsp 65), 7349–7355 (2001). doi:10.1128/IAI.69.12.7349

[183] Betts, J.C., Lukey, P.T., Robb, L.C., McAdam, R.A., Duncan, K.: Evaluation of a nutrient starvation model of Mycobacterium tuberculosis persistence by gene and protein expression profiling. Mol. Microbiol. **43**(3), 717–731 (2002). doi:10.1046/j.1365-2958.2002.02779.x

[184] Salamon, H., Bruiners, N., Lakehal, K., Shi, L., Ravi, J., Yamaguchi, K.D., Pine, R., Gennaro, M.L.: Cutting edge: Vitamin D regulates lipid metabolism in Mycobacterium tuberculosis infection. J. Immunol. **193**(1), 30–4 (2014). doi:10.4049/jimmunol.1400736

[185] Tabira, Y., Ohara, N., Yamada, T.: Identification and characterization of the ribosome-associated protein, HrpA, of Bacillus Calmette-Guérin. Microb. Pathog. **29**, 213–222 (2000). doi:10.1006/mpat.2000.0384

[186] Rohde, K.H., Abramovitch, R.B., Russell, D.G.: Mycobacterium tuberculosis Invasion of Macrophages: Linking Bacterial Gene Expression to Environmental Cues. Cell Host Microbe **2**(5), 352–364 (2007). doi:10.1016/j.chom.2007.09.006

[187] Walters, S.B., Dubnau, E., Kolesnikova, I., Laval, F., Daffe, M., Smith, I.: The Mycobacterium tuberculosis PhoPR two-component system regulates genes essential for virulence and complex lipid biosynthesis. Mol. Microbiol. **60**(2), 312–330 (2006). doi:10.1111/j.1365-2958.2006.05102.x

[188] Duan, X., Li, Y., Du, Q., Huang, Q., Guo, S., Xu, M., Lin, Y., Liu, Z., Xie, J.: Mycobacterium Lysine ϵ-aminotransferase is a novel alarmone metabolism related per-

sister gene via dysregulating the intracellular amino acid level. Sci. Rep. **6**(December 2015), 19695 (2016). doi:10.1038/srep19695

[189] Golby, P., Nunez, J., Cockle, P.J., Ewer, K., Logan, K., Hogarth, P.: Europe PMC Funders Group Characterization of two in vivo-expressed methyltransferases of the Mycobacterium tuberculosis complex: antigenicity and genetic regulation **154**(Pt 4), 1059–1067 (2011). doi:10.1099/mic.0.2007/014548-0.Characterization

[190] Healy, C., Golby, P., MacHugh, D.E., Gordon, S.V.: The MarR family transcription factor Rv1404 coordinates adaptation of Mycobacterium tuberculosis to acid stress via controlled expression of Rv1405c, a virulence-associated methyltransferase. Tuberculosis **97**(November), 154–162 (2016). doi:10.1016/j.tube.2015.10.003

[191] Braibant, M., Gilot, P., Content, J.: The ATP binding cassette (ABC) transport systems of Mycobacterium tuberculosis. FEMS Microbiol. Rev. **24**(4), 449–467 (2000). doi:10.1111/j.1574-6976.2000.tb00550.x

[192] Zumft, W.G., ViebrockSambale, A., Braun, C.: Nitrous oxide reductase from denitrifying Pseudomonas stutzeri Genes for copperprocessing and properties of the deduced products, including a new member of the family of ATP/GTPbinding proteins. Eur. J. Biochem. **192**(3), 591–599 (1990). doi:10.1111/j.1432-1033.1990.tb19265.x

[193] Voskuil, M.I., Schnappinger, D., Visconti, K.C., Harrell, M.I., Dolganov, G.M., Sherman, D.R., Schoolnik, G.K.: Inhibition of respiration by nitric oxide induces a Mycobacterium tuberculosis dormancy program. J. Exp. Med. **198**(5), 705–713 (2003). doi:10.1084/jem.20030205

[194] He, H., Bretl, D.J., Penoske, R.M., Anderson, D.M., Zahrt, T.C.: Components of the Rv0081-Rv0088 Locus, which encodes a predicted formate hydrogenlyase complex, are coregulated by Rv0081, MprA, and DosR in Mycobacterium tuberculosis. J. Bacteriol. **193**(19), 5105–5118 (2011). doi:10.1128/JB.05562-11

[195] Petridis, M., Benjak, A., Cook, G.M.: Defining the nitrogen regulated transcriptome of Mycobacterium smegmatis using continuous culture. BMC Genomics **16**(1), 821 (2015). doi:10.1186/s12864-015-2051-x

[196] Hümpel, A., Gebhard, S., Cook, G.M., Berney, M.: The SigF regulon in Mycobacterium smegmatis reveals roles in adaptation to stationary phase, heat, and oxidative stress. J. Bacteriol. **192**(10), 2491–2502 (2010). doi:10.1128/JB.00035-10

[197] Cavet, J.S., Graham, A.I., Meng, W., Robinson, N.J.: A cadmium-lead-sensing ArsR-SmtB repressor with novel sensory sites. Complementary metal discrimination by NmtR and CmtR in a common cytosol. J. Biol. Chem. **278**(45), 44560–44566 (2003). doi:10.1074/jbc.M307877200

[198] Burts, M.L., Williams, W.A., DeBord, K., Missiakas, D.M.: EsxA and EsxB are secreted by an ESAT-6-like system that is required for the pathogenesis of Staphylococcus aureus infections. Proc. Natl. Acad. Sci. U. S. A. **102**(4), 1169–74 (2005). doi:10.1073/pnas.0405620102

[199] Festa, R.A., Jones, M.B., Butler-Wu, S., Sinsimer, D., Gerads, R., Bishai, W.R., Peterson, S.N., Darwin, K.H.: A novel copper-responsive regulon in Mycobacterium tuberculosis. Mol. Microbiol. **79**(1), 133–148 (2011). doi:10.1111/j.1365-2958.2010.07431.x

[200] Sakthi, S., Narayanan, S.: The lpqS knockout mutant of Mycobacterium tuberculosis is attenuated in Macrophages. Microbiol. Res. **168**(7), 407–414 (2013). doi:10.1016/j.micres.2013.02.007

[201] Festa, R.A., Jones, M.B., Butler-wu, S., Sinsimer, D., Bishai, W.R., Peterson, S.N., Darwin, K.H.: A Novel Copper-Responsive Regulon in Mycobacterium tuberculosis. Russell J. Bertrand Russell Arch. **79**(1), 133–148 (2012). doi:10.1111/j.1365-2958.2010.07431.x.A

[202] Gold, B., Deng, H., Bryk, R., Vargas, D., Eliezer, D., Roberts, J., Jiang, X., Nathan, C.: Identification of a copper-binding metallothionein in pathogenic mycobacteria. Nat. Chem. Biol. **4**(10), 609–616 (2008). doi:10.1038/nchembio.109

[203] Tan, G., Cheng, Z., Pang, Y., Landry, A.P., Li, J., Lu, J., Ding, H.: Copper binding in IscA inhibits iron-sulphur cluster assembly in E scherichia coli. Mol. Microbiol. **93**(4), 629–644 (2014). doi:10.1111/mmi.12676. NIHMS150003

[204] Prach, L., Kirby, J., Keasling, J.D., Alber, T.: Diterpene production in Mycobacterium tuberculosis. FEBS J. **277**(17), 3588–3595 (2010). doi:10.1111/j.1742-4658.2010.07767.x. NIHMS150003

[205] Zhang, M., Yang, Y., Xu, Y., Qie, Y., Wang, J., Zhu, B., Wang, Q., Jin, R., Xu, S., Wang, H.: Trehalose-6-phosphate Phosphatase from Mycobacterium tuberculosis induces humoral and cellular immune responses. FEMS Immunol. Med. Microbiol. **49**(1), 68–74 (2007). doi:10.1111/j.1574-695X.2006.00174.x

[206] Shi, T., Xie, J.: Molybdenum enzymes and molybdenum cofactor in mycobacteria. J. Cell. Biochem. **112**(10), 2721–2728 (2011). doi:10.1002/jcb.23233

[207] Williams, M., Mizrahi, V., Kana, B.D.: Molybdenum cofactor: A key component of Mycobacterium tuberculosis pathogenesis? Crit. Rev. Microbiol. **40**(1), 18–29 (2014). doi:10.3109/1040841X.2012.749211

[208] Leimkuhler, S., Iobbi-Nivol, C.: Bacterial molybdoenzymes: Old enzymes for new purposes (2015). doi:10.1093/femsre/fuv043

[209] Stinear, T.P., Seemann, T., Harrison, P.F., Jenkin, G.a., Davies, J.K., Johnson, P.D.R., Abdellah, Z., Arrowsmith, C., Chillingworth, T., Churcher, C., Clarke, K., Cronin, A., Davis, P., Goodhead, I., Holroyd, N., Jagels, K., Lord, A., Moule, S., Mungall, K., Norbertczak, H., Quail, M.a., Rabbinowitsch, E., Walker, D., White, B., Whitehead, S., Small, P.L.C., Brosch, R., Ramakrishnan, L., Fischbach, M.a., Parkhill, J., Cole, S.T.: Insights from the complete genome sequence of Mycobacterium marinum on the evolution of Mycobacterium tuberculosis. Genome Res. **18**(5), 729–741 (2008). doi:10.1101/gr.075069.107

[210] McGuire, A.M., Weiner, B., Park, S.T., Wapinski, I., Raman, S., Dolganov, G., Peterson, M., Riley, R., Zucker, J., Abeel, T., White, J., Sisk, P., Stolte, C., Koehrsen, M.,

Yamamoto, R.T., Iacobelli-Martinez, M., Kidd, M.J., Maer, A.M., Schoolnik, G.K., Regev, A., Galagan, J.: Comparative analysis of Mycobacterium and related Actinomycetes yields insight into the evolution of Mycobacterium tuberculosis pathogenesis. BMC Genomics **13**(1), 120 (2012). doi:10.1186/1471-2164-13-120

[211] Sekar, B., Arunagiri, K., Selvakumar, N., Preethi, K.S., Menaka, K.: Low frequency of moaA3 gene among the clinical isolates of Mycobacterium tuberculosisfrom Tamil Nadu and Pondicherry south eastern coastal states of India. BMC Infect. Dis. **9**(1), 114 (2009). doi:10.1186/1471-2334-9-114

[212] Lopez, P.M., Golby, P., Wooff, E., Garcia, J.N., Garcia Pelayo, M.C., Conlon, K., Camacho, A.G., Hewinson, R.G., Polaina, J., Garcia, A.S., Gordon, S.V.: Characterization of the transcriptional regulator Rv3124 of Mycobacterium tuberculosis identifies it as a positive regulator of molybdopterin biosynthesis and defines the functional consequences of a non-synonymous SNP in the Mycobacterium bovis BCG ortho. Microbiology **156**(7), 2112–2123 (2010). doi:10.1099/mic.0.037200-0

[213] Chillappagari, S., Seubert, A., Trip, H., Kuipers, O.P., Marahiel, M.A., Miethke, M.: Copper stress affects iron homeostasis by destabilizing iron-sulfur cluster formation in Bacillus subtilis. J. Bacteriol. **192**(10), 2512–2524 (2010). doi:10.1128/JB.00058-10

[214] Hänzelmann, P., Hernández, H.L., Menzel, C., García-Serres, R., Huynh, B.H., Johnson, M.K., Mendel, R.R., Schindelin, H.: Characterization of MOCS1A, an oxygen-sensitive iron-sulfur protein involved in human molybdenum cofactor biosynthesis. J. Biol. Chem. **279**(33), 34721–34732 (2004). doi:10.1074/jbc.M313398200

[215] Yokoyama, K., Leimkühler, S.: The role of FeS clusters for molybdenum cofactor biosynthesis and molybdoenzymes in bacteria. Biochim. Biophys. Acta - Mol. Cell Res. **1853**(6), 1335–1349 (2015). doi:10.1016/j.bbamcr.2014.09.021

[216] Rowland, J.L., Niederweis, M.: Resistance mechanisms of Mycobacterium tuberculosis against phagosomal copper overload. Tuberculosis **92**(3), 202–210 (2012). doi:10.1016/j.tube.2011.12.006

[217] Ward, S.K., Hoye, E.A., Talaat, A.M.: The global responses of Mycobacterium tuberculosis to physiological levels of copper. J. Bacteriol. **190**(8), 2939–2946 (2008). doi:10.1128/JB.01847-07

[218] Kuper, J., Llamas, A., Hecht, H.-J., Mendel, R.R., Schwarz, G.: Structure of the molybdopterin-bound Cnx1G domain links molybdenum and copper metabolism. Nature **430**(7001), 803–806 (2004). doi:10.1038/nature02681

[219] Burns, K.E., Baumgart, S., Dorrestein, P.C., Zhai, H., McLafferty, F.W., Begley, T.P.: Reconstitution of a new cysteine biosynthetic pathway in mycobacterium tuberculosis. J. Am. Chem. Soc. **127**(33), 11602–11603 (2005). doi:10.1021/ja053476x

[220] Voss, M., Nimtz, M., Leimkühler, S.: Elucidation of the dual role of mycobacterial MoeZR in molybdenum cofactor biosynthesis and cysteine biosynthesis. PLoS One **6**(11) (2011). doi:10.1371/journal.pone.0028170

[221] Mendel, R.R.: The molybdenum cofactor. J. Biol. Chem. **288**(19), 13165–13172 (2013). doi:10.1074/jbc.R113.455311

[222] Morrison, M.S., Cobine, P.A., Hegg, E.L.: Probing the role of copper in the biosynthesis of the molybdenum cofactor in Escherichia coli and Rhodobacter sphaeroides. JBIC J. Biol. Inorg. Chem. **12**(8), 1129–1139 (2007). doi:10.1007/s00775-007-0279-x

[223] Brodin, P., Poquet, Y., Levillain, F., Peguillet, I., Larrouy-Maumus, G., Gilleron, M., Ewann, F., Christophe, T., Fenistein, D., Jang, J., Jang, M.S., Park, S.J., Rauzier, J., Carralot, J.P., Shrimpton, R., Genovesio, A., Gonzalo-Asensio, J.A., Puzo, G., Martin, C., Brosch, R., Stewart, G.R., Gicquel, B., Neyrolles, O.: High content phenotypic cell-based visual screen identifies Mycobacterium tuberculosis acyltrehalose-containing glycolipids involved in phagosome remodeling. PLoS Pathog. **6**(9) (2010). doi:10.1371/journal.ppat.1001100

[224] Goh, K.S., Rastogi, N., Berchel, M., Huard, R.C., Sola, C.: Molecular evolutionary history of tubercle bacilli assessed by study of the polymorphic nucleotide within the

nitrate reductase (narGHJI) operon promoter. J. Clin. Microbiol. **43**(8), 4010–4014 (2005). doi:10.1128/JCM.43.8.4010-4014.2005

[225] Kumar, A., Deshane, J.S., Crossman, D.K., Bolisetty, S., Yan, B.S., Kramnik, I., Agarwal, A., Steyn, A.J.C.: Heme oxygenase-1-derived carbon monoxide induces the Mycobacterium tuberculosis dormancy regulon. J. Biol. Chem. **283**(26), 18032–18039 (2008). doi:10.1074/jbc.M802274200