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

Authors

Jennifer Stiens, Yen Yi Tan, Rosanna Joyce, Kristine B. Arnvig, Sharon L. Kendall, Irene Nobeli

ABSTRACT

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

Abbreviations

CDS, coding sequence

ME, module eigengene

MM, module membership

Mtb, *Mycobacterium tuberculosis*

MTBC, *Mycobacterium tuberculosis* complex

ncRNA, non-coding RNA

ORF, open reading frame

RNA-seq, RNA sequencing

RNAP, RNA polymerase

41
42 sORF, short open reading frame

43 sRNA, short non-coding RNA

44 TSS, transcription start site

45 UTR, untranslated region

46 WGCNA, weighted gene co-expression analysis

47

INTRODUCTION

Tuberculosis continues to be a leading cause of death worldwide, causing over 1.5 million deaths, and infecting over 10 million people in 2020 (World Health Organization, 2021). The human-adapted pathogen causing tuberculosis, *Mycobacterium tuberculosis* (Mtb), has a complex lifestyle that requires rapid adaptation to host defences and immune pressure, including nutritional immunity, hypoxia and lipid-rich environments. In order to eradicate the disease, it is crucial to understand how the pathogen survives attacks from host immune cells and persists in an extended latent state inside the host. To adapt to these environmental challenges, bacterial cells must make complex transcriptomic adjustments, and these are thought to be complemented and fine-tuned by post-transcriptional regulation.

The mycobacterial genome produces a range of conditionally expressed transcripts, many of which are poorly annotated and understood. In this paper, 'non-coding' RNA (ncRNA) refers to non-ribosomal RNA transcripts not known to be translated into peptides, such as short RNAs (sRNAs) acting on either distant or antisense mRNA targets and the expressed untranslated regions (UTRs) flanking coding genes (which may also contain short open reading frames (sORFs), upstream from coding regions). Non-coding RNA can alter the abundance of gene products by controlling mRNA stability and processing, access to ribosome binding sites and the translation of overlapping open reading frames (ORFs). Discovering the contribution of the non-coding genome to specific adaptation-response pathways may improve our ability to design therapeutics and prevent the evolution of persistent phenotypes.

Uncovering the role of non-coding RNA in adaptation and transcriptomic remodelling

The proportion of non-ribosomal, ncRNA in the Mtb transcriptome has been shown to increase in stationary and hypoxic conditions, indicating a potential role in adjusting to environmental cues (Aguilar-Ayala et al., 2017; Arnvig et al., 2011; Gerrick et al., 2018; Ignatov et al., 2015). Several mycobacterial ncRNA transcripts (particularly, sRNA) have been extensively studied and found to be associated with regulatory systems controlling adaptation to stress conditions or growth phase, linked to virulence pathways and to access to lipid media (Arnvig et al., 2011; Gerrick et al., 2018; Girardin & McDonough, 2020; Mai et al., 2019; Moores et al., 2017; Solans et al., 2014). Non-coding regulation in Mtb appears to function quite differently compared to model organisms, eschewing the use of any known chaperone proteins for RNA-RNA interactions and with few sRNA homologs found outside the phyla (Gerrick et al., 2018; Mai et al., 2019; Schwenk & Arnvig, 2018). The discovery and characterisation of ncRNA in Mtb, especially sRNAs, has progressed using both molecular biology methods and high-throughput sequence-based approaches (reviewed in Schwenk & Arnvig, 2018) but characterising the gene interactions of a particular sRNA is an experimentally-expensive process, and the number of fully-characterised ncRNAs remains limited. Annotation of identified transcripts remains incomplete, as well, with only 30 listed in the Mtb H37Rv reference sequence (GenBank AL123456.3). Efforts to compile a comprehensive list of annotated ncRNAs for Mtb faces challenges of non-standardised nomenclature, different standards of experimental validation, incomplete reference annotations (especially for the animal-adapted species of the Mycobacterium tuberculosis complex (MTBC)) and the variable expression of non-coding transcripts in response to different experimental conditions (Stiens et al., 2022).

Prediction of ncRNA from RNA-sequencing (RNA-seq) data in the compact Mtb genome is challenging. Paradoxically, more sensitive, high-depth sequencing can make it more difficult to identify the small, low-abundance, functional transcripts above stochastic gene expression and technical noise. Parameters of detection must therefore be carefully considered for each dataset to account for variation in expression levels. Though RNA-seq-based ncRNA prediction algorithms are often assumed to overpredict putative ncRNAs, especially at the 5' and 3' ends of coding genes, there are biological and technical reasons for detecting abundant signal in the unannotated regions of the genome. Ribosome profiling (Ribo-seq) methods that sequence the ribosome-protected fragments of mRNA have identified actively translated RNA in the 5' UTRs of annotated protein-coding mRNA transcripts (Canestrari et al., 2020; D'Halluin et al., 2022; Sawyer et al., 2021; Shell et al., 2015; C. Smith et al., 2022). These unannotated sORFs may represent functional peptides or function to regulate the translation of the downstream transcript; however, it is impossible to tell the difference between a putative ncRNA and a sORF from RNA-seq signal alone. Additionally, post-transcriptional processing may be the norm for prokaryotes at both the 5' and 3' ends of coding transcripts, with 3' ends in mycobacteria often lacking clear signal termination (Dar & Sorek, 2018; D'Halluin et al., 2022; Wang et al., 2019). Finally, polycistronic transcripts often include non-coding sequence between the genes of an operon, and this may contain functional elements and/or processing sites (Martini et al., 2019).

The location of a transcription start site (TSS) in the 5' end of a predicted transcript supports the biological relevance of a predicted ncRNA. However, the available lists of Mtb TSS sites (Cortes et al., 2013; Shell et al., 2015) have been mapped only in starvation and exponential growth and may not include TSSs that are expressed under different experimental conditions. New TSS maps, published subsequent to this analysis may

increase the number of predicted transcripts with a TSS (D'Halluin et al., 2022). Furthermore, functional ncRNA elements generated from the 3' UTRs of coding genes through RNase processing would presumably lack a TSS. 3' UTRs that are functionally independent from their cognate coding sequence (CDS) have been identified in other bacteria (Desgranges et al., 2021; Menendez-Gil et al., 2020; Ponath et al., 2022). Therefore, it is important to consider predicted UTRs as separate annotated elements from protein-coding transcripts when quantifying differential expression.

To include a complete picture of the interaction of the non-coding genome with coding genes involved in adaptation pathways, we have generated a novel set of ncRNA sequence-based predictions (sRNAs and UTRs) from the same datasets using our in-house software package, *baerhunter* (Ozuna et al., 2019). Some of these predicted transcripts overlap with predictions from previous studies, but many represent novel predictions. The expression of these transcripts is quantified along with the protein-coding genes and used in network analysis to provide a more complete picture of the functional groupings involved in adaptation to environmental changes. Including a variety of culture conditions that replicate aspects of the host environment improves the chances that the expression of any ncRNA that is restricted to one or more conditions is included in the network (Ami et al., 2020).

Using WGCNA to implicate functional associations of non-coding RNA

Weighted gene co-expression network analysis (WGCNA) (B. Zhang & Horvath, 2005) has been widely used to identify functional groups of genes, called 'modules', through the application of hierarchical clustering to differential expression levels of RNA transcripts in microarray or RNA-seq experiments. Recent studies have focussed entirely on the protein-coding portion of the transcriptome, using WGCNA with RNA-seq to cluster the

differentially expressed genes of *Mycobacterium marinum* in response to resuscitation after hypoxia (Jiang et al., 2020) and *Mycobacterium aurum* infected macrophages (Lu et al., 2021). Mtb microarray data have been used to cluster protein-coding genes that show differential expression among species-specific strains (Puniya et al., 2013) and in response to two different hypoxic models to identify potential transcription factors (Jiang et al., 2016). Another recent network analysis, using a matrix deconvolution method followed by module clustering, uses a large number of RNA-seq samples including deletion mutants, infection models and antibiotic-treated samples as well as restricted media and culture conditions (Yoo, et al., 2022). They identify 80 modules of protein-coding genes that each approximate an isolated source of variance, together estimated to account for 61% of the total variance seen in in the dataset. This proportion is reportedly lower than results from similar analyses in other organisms, potentially due to the bias in the types of conditions available in the database and/or the complex nature of regulation in Mtb (Yoo, et al., 2022). However, the contribution of regulatory ncRNA elements may be a considerable unexplored source of variance in this complex system. Here we use an alternative, complementary approach by including ncRNA, as well as annotated protein-coding genes, in the modules.

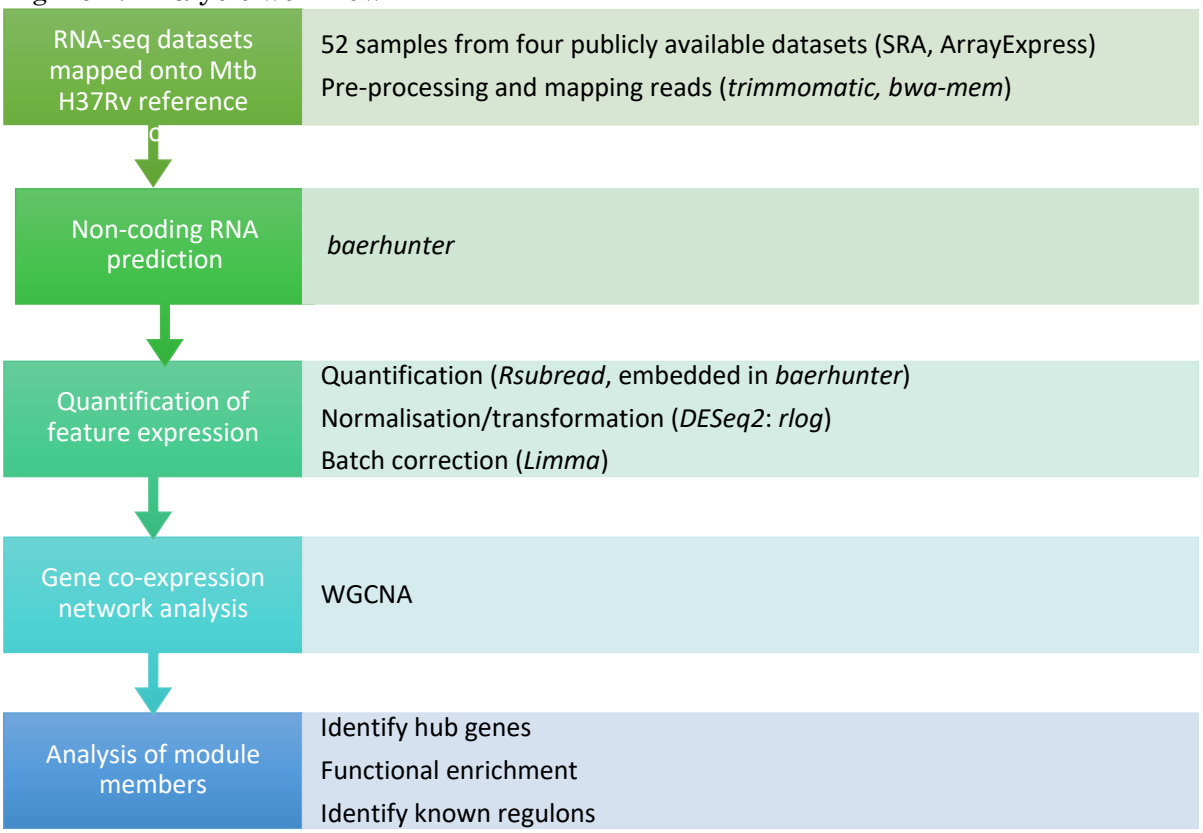
In this study, WGCNA was applied to multiple Mtb H37Rv datasets covering 15 different culture conditions replicating various growth conditions, nutrient sources and stressors encountered in the host environment. We present a global view of the non-coding genome across an extensive WGCNA network and interrogate selected modules to identify functional groupings between protein-coding and non-coding transcripts, as well as between well-characterised genes and those with little functional annotation. The correlation of the modules with the various conditions can identify participants in large-

scale transcriptomic remodelling programs in response to changes in environmental conditions.

MATERIALS AND METHODS

The overall workflow for this analysis is presented in Figure 1. All scripts for *baerhunter*, WGCNA and subsequent analysis are available at: https://github.com/jenjane118/mtb_modules.

Figure 1. Analysis workflow



Data Acquisition and Mapping

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

Table 1. Datasets used in analysis. Accession numbers from SRA and Array Express.

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

Non-coding RNA prediction

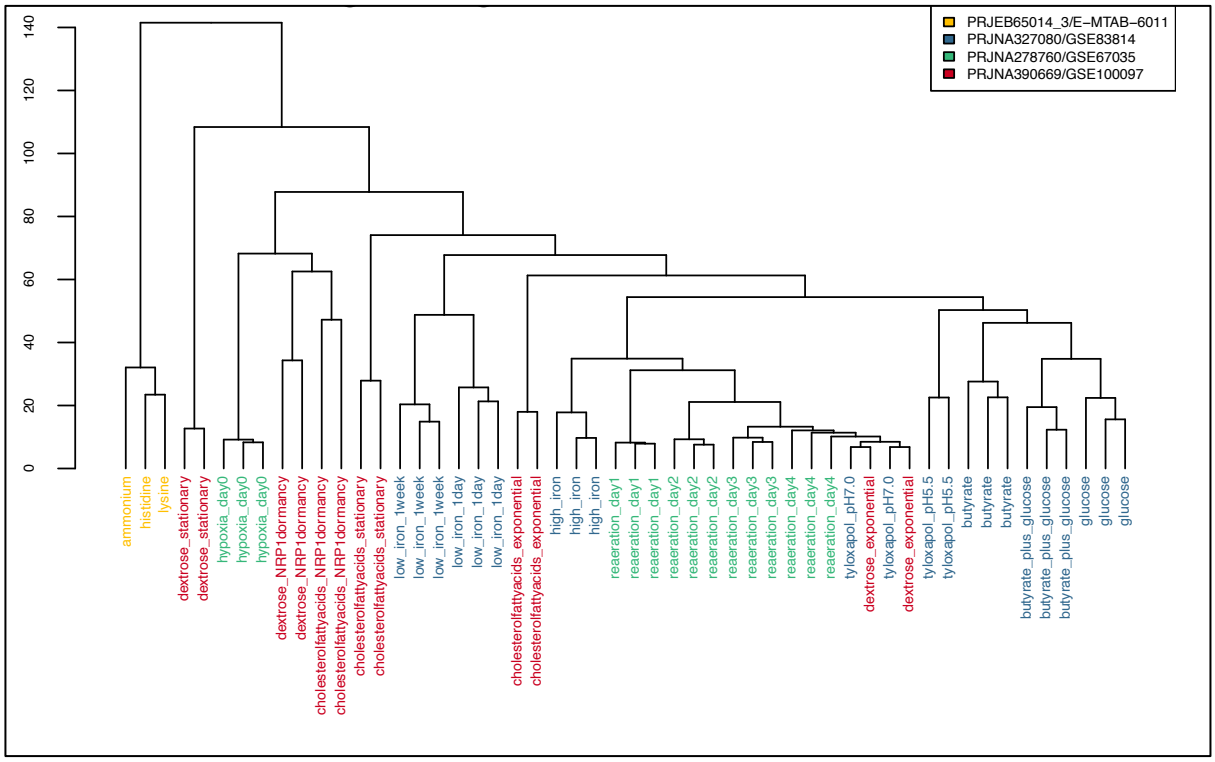
Each dataset was run through the R-package, *baerhunter* (Ozuna et al., 2019), using the '*feature_file_editor*' function optimised to the most appropriate parameters for the sequencing depth (https://github.com/jenjane118/mtb_modules). '*Count_features*' and '*tpm_norm_flagging*' functions were used for transcript quantification and to identify low

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

DESeq2 v1.30.1 (Love et al., 2014) was used on the complete counts matrix including the filtered *baerhunter* predictions to calculate size factors, estimate dispersion and normalise the data with the regularised log transformation function (Supp figures, S1 and S2). The normalised data was checked for potential batch effects using PCA plots and hierarchical dendrograms. *Limma* v3.46.0 (Ritchie et al., 2015) *'removeBatchEffect'* was applied with a single batch argument to remove batch effects associated with the first component (batching the data according to dataset due to technical differences) while preserving differences between samples. The final hierarchical dendrogram, post-batch

correction, indicates successful application as samples cluster by similar experimental conditions, rather than by dataset alone (Figure 2 compared to Supp figure S3). Samples from experiment PRJEB65014 continue to group together, but as they represent single replicates in unique conditions, it is difficult to estimate the influence of confounding batch effects for these samples.

Figure 2. Hierarchical dendrogram of *rlog* transformed and *limma* batch corrected expression data by sample. The sample labels are coloured by dataset, demonstrating that they are clustering by condition, rather than experiment.



Creation of the WGCNA network

The normalised and batch-corrected expression matrix was used to create a signed co-expression network using the R package, *WGCNA* v1.69 (Langfelder & Horvath, 2008), with the following parameters: corType = "bicor", networkType = "signed", power = 14, TOMType = "signed", minModuleSize = 20, reassignThreshold = 0, mergeCutHeight =

0.15, deepSplit = 2. In this type of network, the 'nodes' are the genes, and the 'edges', or links, are created when gene expression patterns correlate. In contrast to unweighted binary networks where links are assigned 0 or 1 to indicate whether or not the genes are linked, in a weighted network, the links are given a numeric weight based on how closely correlated the expression is. *WGCNA* first calculates the signed co-expression similarity for each gene pair. The absolute value of this correlation is raised to a power (determined by the user, based on a scale-free topology model that mimics biological systems (Supp figure S4) in order to weight the strong connections more highly than the weaker connections. The resulting similarity matrix is used to cluster groups of genes with strong connections to each other in a non-supervised manner (i.e., it doesn't use any previous information about gene groups or connected regulons). A cluster dendrogram is created (Supp figure, S8) and closely connected branches of the dendrogram are merged into modules based on a cut-off value (also a parameter controlled by the user). The modules are defined by a 'module eigengene' (ME), which explains most of the variance in the expression values in the module. The connectivity of the MEs define the shape of the overall network (Supp figure, S9). The modules can then be tested for potential correlations with experimental conditions without incurring the same punitive penalties for multiple testing as individual gene correlations would (Supp figure, S10). In signed networks, correlation of the module with a condition can be in either the positive or negative direction, as modules include transcripts that are similar in both the degree and direction of correlation, allowing for a more fine-grained analysis than with unsigned networks.

To test correlations of modules with experimental conditions, the individual RNA-seq samples were assigned to a condition based on the experimental description in the project metadata. Some of these conditions were shared among the different projects, so when

appropriate, samples from different datasets were assigned the same condition, resulting in 15 tested conditions. For example, late-stage reaeration samples were tested along with exponential growth samples, and samples that tested hypoxia and cholesterol utilisation together were included in multiple conditions. Models of hypoxia differed between the RNA-seq projects, and these samples were assigned to different conditions: 'hypoxia' versus 'extended hypoxia' (Supp Table 1, 'Condition summary' tab). All correlations were made using robust biweight midcorrelation tests and all p-values were corrected for multiple testing with the BH-fdr method (Benjamini & Hochberg, 1995). Significance was evaluated as an adjusted p-value (p_{adj}) of < 0.05.

Module Enrichment

Modules were interrogated for enrichment for Gene Onotology (GO) terms (Ashburner et al., 2000; The Gene Ontology Consortium, 2021), Clusters of Orthologous Groups (COG) (Galperin et al., 2021), KEGG pathway genes (Kanehisa et al., 2022), functional categories and literature searches for known regulons. GO terms, COG term and KEGG pathway enrichment were accessed programmatically using the David web service (Huang et al., 2009b, 2009a; Jiao et al., 2012) to query the list of protein-coding genes from each module for enrichment. Enrichment was determined using a modified one-sided Fisher's Exact Test ('EASE' score) with fdr correction for multiple testing, with $p_{adj} < 0.01$ considered significantly enriched for a particular term or pathway, and $p_{adj} < 0.05$ for COG term. Enrichment for the 11 functional categories from Mycobrowser annotation (Kapopoulou et al., 2011) was determined using a one-sided Fisher's Exact Test with fdr correction for multiple testing. Modules were enriched for a particular functional category if $p_{adj} < 0.01$. Lists of genes associated with known regulons were mined from literature and enrichment was tested using the same one-sided Fisher's Exact Test as above with a $p_{adj} < 0.01$ cut-off for enrichment.

312

313 Non-coding RNA prediction, network analysis and subsequent data manipulation was
314 performed with R (v4.0.5, 2021-03-31). All plots were made in R with the following
315 packages: *WGCNA* (v1.69), *dendextend* (v1.15.2), *ggplot2* (v3.3.5). Scripts and expression
316 data are available at https://github.com/jenjane118/mtb_modules .

317

318

RESULTS AND DISCUSSION

Mtb expresses an extensive range of ncRNA transcripts over a wide variety of experimental conditions

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

nucleotides of the sequence (Table 2 and Supp Table 2 'putative_sRNAs', 'putative_UTRs' tabs).

Table 2. Tally of predicted expressed elements in the *baerhunter*-generated combined annotation file. 4015 protein-coding genes were included in the annotation. *TSS predictions from (Cortes et al., 2013; Shell et al., 2015).

Predicted element	Number predicted	With predicted TSS* (exponential and starvation)
Total sRNA	1283	553
sRNA 'intergenic'	91	23
sRNA 'antisense'	1192	530
Total UTRs	1715	273
5' UTRs	475	200
3' UTRs	602	16
'Between' UTRs	638	57

The predicted sRNAs were further annotated using the accepted nomenclature (Lamichhane et al., 2013) which identifies the putative ncRNA relative to annotated gene loci and differently signifies truly intergenic sRNAs and those that overlap any part of a protein-coding region on the opposite strand. Most of the putative sRNAs are antisense to the protein-coding region of one or more genes, but 91 putative sRNAs have predicted boundaries that do not overlap an annotated transcript on either strand (or overlap an annotated transcript on the opposite strand by fewer than 10 nucleotides). This number is most probably an underestimate of the truly 'intergenic' sRNAs in the genome, as many of the sRNA predictions appear over-estimated at the 3' end, effectively classifying them as an antisense RNA even though the 5' half of the transcript does not overlap any genes on the opposite strand. Isoforms of annotated sRNAs can be subject to post-transcriptional processing to create an active transcript (Moores et al., 2017) and post-transcriptional

366 processing of 3' ends *in vivo* is more likely the norm for most prokaryotic transcripts
367 (Wang et al., 2019). However, for our purposes, any RNA-seq transcripts that extend to
368 overlap a protein-coding gene on the other strand in any dataset will be labelled as
369 antisense RNA.
370
371 The generated combined annotation file was used to quantify the expression of all 7043
372 expressed elements, including every annotated CDS, annotated ncRNA and predicted
373 ncRNA, in each sample. Raw counts of expression varied greatly among the datasets due
374 to different sequencing depth, as well as between some samples within datasets (as would
375 be expected with different environmental conditions), and only three protein-coding genes
376 showed no expression in any sample. The raw expression counts were transformed using
377 DESeq2's rlog function (Love et al., 2014), and plots of the dispersion of count data show
378 that the median expression level between samples and between datasets has been
379 normalised (Supp figures S1, S2). The distribution of the normalised expression levels of
380 protein-coding regions alone shows consistent median expression levels across the entire
381 dataset, however distribution of the normalised data restricted to putative sRNAs shows
382 more variability, with certain conditions showing increased or decreased expression of
383 these transcripts (Supp figures S5-S7). This is not unexpected, given that several studies
384 have identified pervasive transcription in hypoxic infection models, stationary phase and
385 dormancy. This is accompanied by a concomitant increase in non-rRNA expression
386 (especially antisense RNA transcripts) and in the number of predicted TSSs in Mtb and
387 *M. smegmatis* (a fast-growing, non-pathogenic strain) (Arnvig et al., 2011; Ignatov et al.,
388 2015; Martini et al., 2019).
389

390 **Module networks represent groups of co-expressed genes and predicted non-**
391 **coding RNA**

392

393 ***Creation of the WGCNA network***

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

403

404 The signed co-expression network presented in this paper consists of 56 different modules,
405 assigning 97.6% of the expressed elements (CDS, putative UTRs and putative sRNAs)
406 into 55 modules, with 168 unassigned elements clustered in the 'grey' module (Supp Table
407 2, 'Module_Overview' tab). Module size ranged from 1086 to 25 expressed elements. The
408 modules (using the ME) were tested for correlations with the various conditions used in
409 the RNA-seq experiments (see Materials and Methods). The RNA-seq data was
410 categorised into 15 different experimental conditions in total with varying numbers of
411 replicates (Supp Table S1, 'Condition Summary' tab), therefore, a statistically significant
412 correlation of modules with every condition was not expected. However, some modules do
413 show significant correlations with conditions such as iron restriction, cholesterol media,
414 hypoxia and growth phase and this can be informative when considering the association
415 of the gene groups with biological processes (Figure 3).

416

417 ***Well-established regulons cluster together in single modules***

In many cases, the gene membership of the modules includes well-established regulons or groups of functionally related genes, establishing the biological relevance of the module sub-networks and proof of concept for the application of WGCNA on such a heterogenous dataset. For example, the DosR regulon is a well-studied regulon associated with hypoxia and stress responses (Du et al., 2016; Rustad et al., 2008; Voskuil et al., 2004). 40 of the 48 DosR-regulated genes are found in a single module, '*greenyellow*', which is negatively correlated with reaerated culture and exponential growth (Figure 3) and enriched for the GO term, 'response to hypoxia'. Unsurprisingly, this represents statistically significant enrichment of DosR-regulated genes in the module (one-sided Fisher's exact test, p_{adj}= 6.6e-50). The '*greenyellow'* module is also enriched for genes from the PhoP regulon (one-sided Fisher's exact test, p_{adj}=0.021) which is associated with hypoxic response and coordination with the DosR regulon (Gonzalo-Asensio et al., 2008; Singh et al., 2020). The KstR regulon includes 74 genes under control of the TetR-type transcriptional repressor, KstR, known to be involved in lipid catabolism and upregulated during infection (Kendall et al., 2007, 2010). The '*turquoise*' module is significantly enriched for known KstR-regulated genes (one-sided Fisher's exact test, p_{adj} = 0.0026) with 35 of 74 KstR-regulated genes clustering together in the module. This module showed significant positive correlation with hypoxia, extended hypoxia and stationary growth phase, and a negative correlation with exponential growth (Figure 3).

Other examples include genes involved in mycobactin synthesis which are nearly all found in the *'steelblue'* module (positively correlated with the low iron condition), and the genes of the DIM locus which are significantly enriched (one-sided Fisher's Exact test, p_{adj}=4.95e-5) in the *'paleturquoise'* module (positively correlated for exponential growth and reaerated culture and negatively correlated to slow growth conditions) (Figure 3). As these examples show, known associated genes are co-located in modules which represent

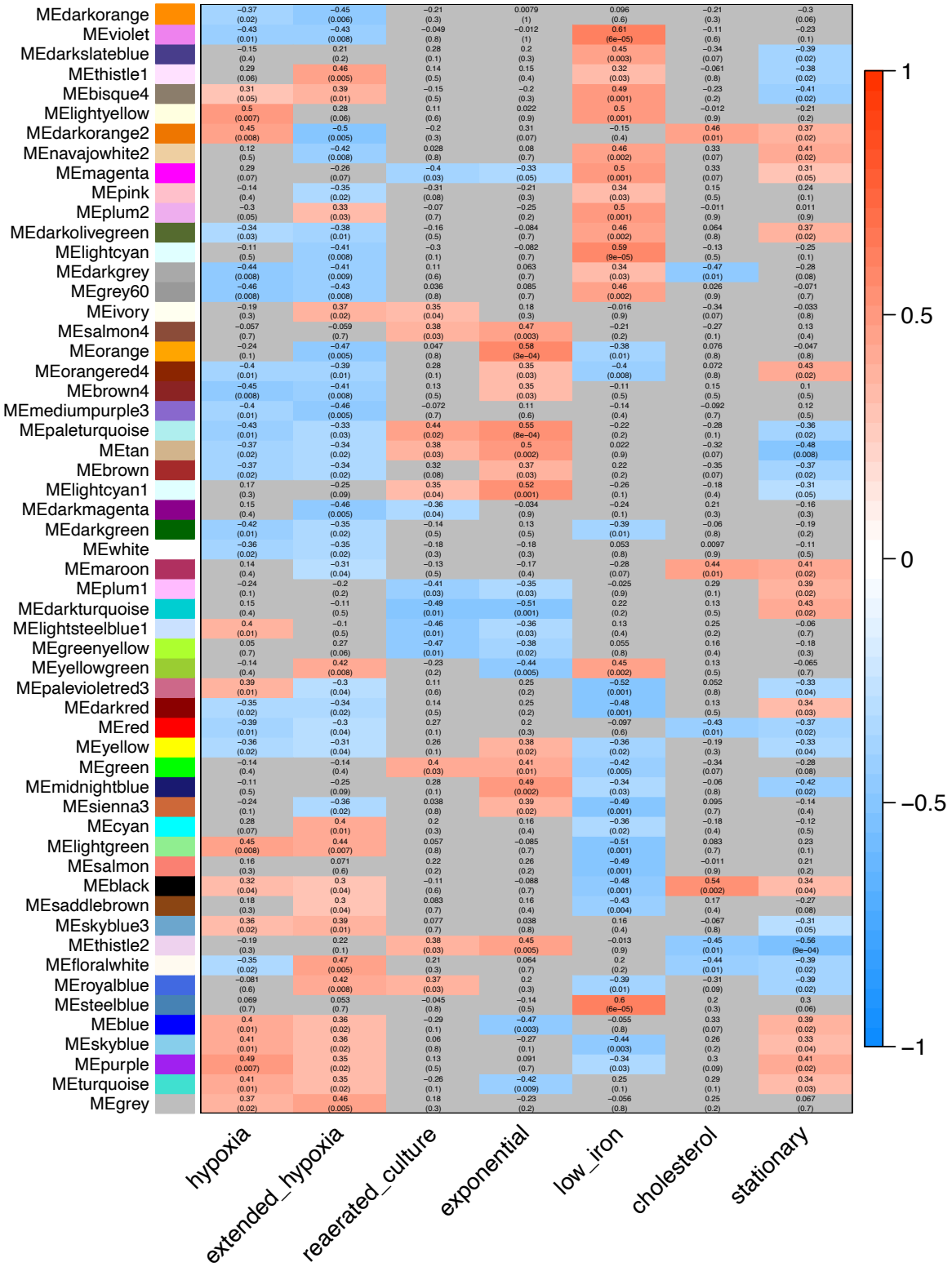
444 a functional group of genes that have co-regulated expression under various experimental

445 conditions. The modules can be further explored to identify novel associations.

446

447

448 Figure 3. Heat map of correlation of module eigengene (ME) of each module with selected
449 experimental conditions. Correlation was calculated using biweight midcorrelation (bicor)
450 and p-values were adjusted for multiple testing (BH-fdr). Positive correlation is red,
451 negative correlation is blue. Non-significant correlations in grey ($p_{adj} < 0.05$).



452

453

454 ***Predicted non-coding RNAs are enriched in certain modules***

Putative sRNAs and predicted UTRs were distributed throughout all modules in the network (Figure 4, Supp Table 2, 'Module_Overview' tab). The number of predicted elements were enriched in certain modules: the two largest modules, *'turquoise'* and *'blue',* are significantly enriched for predicted sRNAs, and eight modules are statistically enriched for predicted UTRs (one-sided Fisher's exact test, $p_{adj} < 0.01$, Supp Table 2, 'Module_Overview' tab). A roughly linear relationship between the number of CDS and the number of UTRs, is to be expected, given that UTRs are defined by the *baerhunter* algorithm by their position at the start or end of protein-coding genes (Ozuna et al., 2019). However, if the UTRs are positioned in an operon, there will be a smaller increase in the relative number of UTRs with an increasing number of protein-coding genes, as UTRs between two protein-coding genes are predicted as a single UTR. As a result, the two modules with the highest number of predicted operons (from OperonDB, Chetal & Janga, 2015), *'turquoise'* and '*brown*', have a lower relative proportion of UTRs (Figure 5).

Within the module sub-networks, the tight co-expression of protein-coding genes and ncRNA is reflected by the number of ncRNA found among the most connected elements in the module. The 'hub' elements are those with the best correlation to the ME and therefore the most tightly connected elements in the individual module networks. In 12 modules, ncRNA (both predicted and annotated) make up more than half of the elements with module membership values (MM) > = 0.80 (our threshold for identifying hub elements) (Supp Table 2, 'Hub_info' tab). This implicates ncRNA as important members of the regulatory pathways implemented to adapt to conditions such as hypoxia, cholesterol media and low iron. The 30 annotated ncRNAs in the Mtb reference genome (AL123456.3) are spread over 15 modules, with 10 of them hubs of the modules, and one unassigned ('grey' module) (Supp Table 2, 'Annotated ncRNA' tab). For example, Ms1/MTS2823, observed to be the most abundantly expressed ncRNA in expression

studies over various stress conditions (Arnvig et al., 2011; Arnvig & Young, 2012; Ignatov et al., 2015; Šiková et al., 2019), is a hub element in a module that is positively correlated with hypoxia and negatively correlated with exponential and reaerated culture conditions (*'lightsteelblue1',* Figure 3). Mcr7/ncRv2395A, found to be part of the PhoP regulon (Solans et al., 2014), is a hub in the *'magenta'* module enriched for elements in the KEGG pathway for valine, leucine and isoleucine degradation and correlated positively with the low iron condition (Figure 3). F6/ncRv10243/SfdS, a sRNA upregulated in starvation and mouse infection models, is thought to be involved in regulating lipid metabolism and long-term persistence (Houghton et al., 2021). This ncRNA is a hub in a module found to be enriched for the GO terms 'lipid metabolism' and 'biosynthesis of fatty acids' (*'lightcyan'*) and found to be correlated positively with low iron and negatively with extended hypoxia conditions (Figure 3).

Figure 4. Relative proportion of annotated CDS, predicted UTRs and predicted sRNAs in each module, ordered by module size.

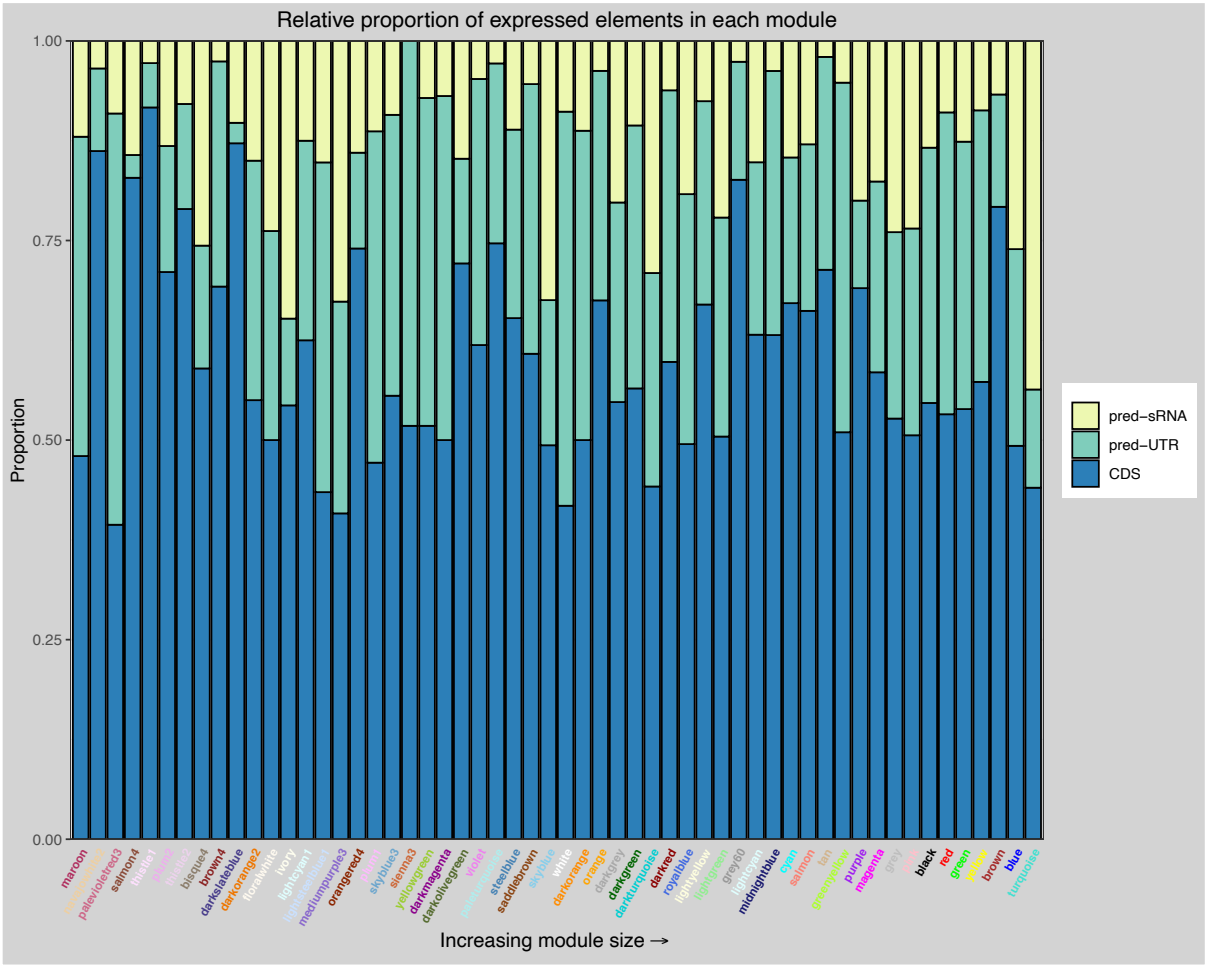
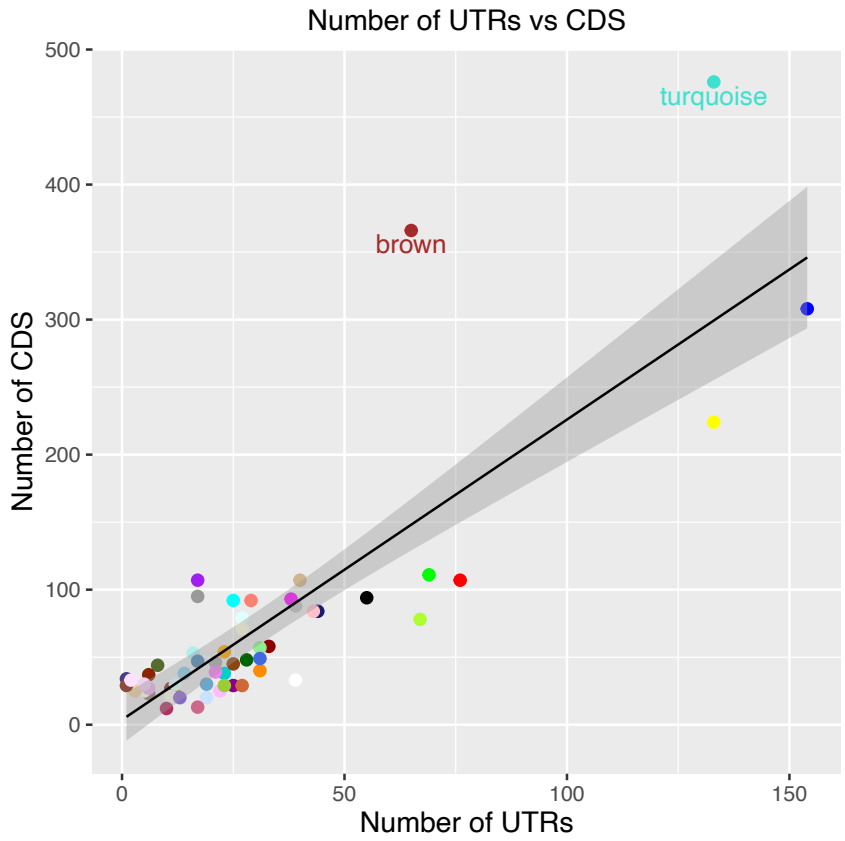


Figure 5. Plot of number of UTRs against number of CDS in each module. Grey shading indicates confidence interval of 0.95.



505
506

507

508 ***UTR and adjacent ORF expression differ in nearly 50% of cases***

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

Of the *baerhunter*-predicted UTRs labelled 5' and 3', the UTRs co-segregated with the ORF they were closest to approximately half the time (Table 3). We would expect correctly-identified 5' UTRs to utilise a TSS (whether or not there is a known predicted TSS), whereas it appears functional 3' UTRs are more likely to be cleaved from the longer mRNA transcript (Dar & Sorek, 2018; Menendez-Gil & Toledo-Arana, 2021; Ponath et al., 2022). Our data confirms this: transcripts classified as 5' UTRs are much more likely to have a predicted TSS in the first 10 nucleotides than transcripts classified as 3' UTRs (42% vs 2.7%). Approximately 11% of the UTRs predicted to be between ORFs (labelled, 'Between' UTRs) have predicted TSS (Table 3). The presence of a TSS in the first 10 nucleotides of the predicted UTR appeared to have little bearing on whether or not the UTR and its adjacent ORF are assigned to the same module, with 56% of 5' and 44% of 3' UTRs with a predicted TSS co-assigned with their adjacent ORF partner. A similar proportion of the 'Between' UTRs (38%) do not segregate with either the ORF upstream or downstream, indicating their expression is, to some degree, independent of either adjacent ORF. All UTRs that are in modules independent of their adjacent ORF(s) are found in Supplementary Table 2, 'independent_UTRs' tab.

Table 3. UTRs and module assignment of adjacent ORFs. DS=downstream, US=upstream. TSS indicates presence of annotated TSS in first 10 nucleotides of predicted UTR (Cortes et al., 2013; Shell et al., 2015).

	Total (excluding grey)	Number with TSS	Number in same module as adjacent ORF	Proportion of UTRs in same module as ORF
5' UTR	462	196	227 DS	48%
3' UTR	592	16	296 US	49%
BTWN UTR	622	55	117 DS 140 US 126 both 239 none	19% 23% 20% 38%

Antisense RNAs are hubs in modules independent of cognate ORF

It has been observed that the overall abundance of antisense RNA and other non-ribosomal RNA increases upon exposure to stress such as hypoxia and nutrient restriction (Arnvig et al., 2011; Ignatov et al., 2015), and in our network, ncRNA, including antisense RNAs, were found to be well-connected hubs in module sub-networks associated with known gene regulons, such as DosR and KstR. This supports the view that antisense RNA may be part of specific regulatory networks, especially those that are involved in adaptation to environmental conditions, rather than products of indiscriminate pervasive transcription (Arnvig et al., 2011; Lloréns-Rico et al., 2016). Not unexpectedly, very few (7%) of the predicted antisense transcripts were assigned to the same module as the protein-coding region overlapping on the opposite strand (choosing the most downstream locus in the event of multiple overlapping ORFs), signifying distinct patterns of expression for transcripts on opposite strands, possibly due to independent or bi-directional promoters and/or overlapping transcription termination sites. Bi-directional promoters have been identified in multiple prokaryotic genomes, and competition for RNA polymerase (RNAP) binding among divergently transcribed sense/antisense pairs may function as a mechanism for regulation of gene expression (Ju et al., 2019; Warman et al., 2021). Long 3' UTRs that overlap with converging protein-coding genes on the opposite strand (or with the 3' UTR) can create an 'excludon' regulatory arrangement, where transcription of the two opposite mRNAs is simultaneously regulated by RNase targeting, or mutually exclusive due to RNAP collision (Sáenz-Lahoya S. et al., 2019; Toledo-Arana & Lasa, 2020). Examining the module groupings of the antisense RNAs and their base-pairing target on the other strand may provide insight on which genes are regulated by antisense transcription.

Focus on Selected Module Networks

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

One condition that causes a major shift in the transcriptome is the adaptation of Mtb to a cholesterol or lipid-rich environment, a process that involves a multitude of gene pathways to facilitate the pathogen's survival and persistence in the infected macrophage (Del Portillo et al., 2019; Pandey Amit K. & Sassetti Christopher M., 2008; Pawełczyk et al., 2021). In fact, a recent study, published after this analysis, observes differential expression of over 500 protein-coding genes with a switch from glycerol to cholesterol as the carbon source (Pawełczyk et al., 2021). Our network includes transcriptomes from several samples that utilise cholesterol and fatty-acid containing media over a range of growth conditions including hypoxia (SRA project: PRJNA390669) (Aguilar-Ayala et al., 2017) and although several modules have a significant correlation with the cholesterol media trait, other modules with clusters of genes related to cholesterol catabolism are correlated to hypoxia or extended hypoxia conditions. All of these modules are found to contain a large number of predicted non-coding elements, confirming studies that show increased ncRNA expression levels in response to lipid conditions and cholesterol,

596 especially when combined with hypoxia; conditions meant to most resemble those
597 encountered in host infection models (Aguilar-Ayala et al., 2017; Del Portillo et al., 2019;
598 Soto-Ramirez et al., 2017).
599
600 Several modules correlating with the low iron condition show enrichment of genes
601 associated with siderophore synthesis, transport and regulation, along with redox sensors
602 and genes known to be upregulated in response to cholesterol media. Restricting iron
603 availability to growing cells is meant to replicate a host response to infection and will
604 stimulate a cascade of pathways to enable the pathogen to survive in a slow-growing, or
605 latent state. The co-expression of genes involved in metal ion homeostasis and genes
606 known to be involved in adaptation to cholesterol and lipids is supported by observations
607 in a recent study that the presence of cholesterol causes changes in metal ion metabolism
608 (Pawełczyk et al., 2021) and closer inspection may uncover gene interactions related to
609 the metabolic changes made in anticipation of re-entry from hypoxic environments when
610 bacteria are particularly vulnerable to oxidative stress (Eoh et al., 2017; Gerrick et al.,
611 2018).
612
613 The data have been organised into an easily-accessible spreadsheet for researchers to
614 query particular genes or modules of interest and find associated protein-coding genes or
615 ncRNA (Supp Table 2). We anticipate this to be a useful resource to find ncRNA
616 candidates for further study, to identify associations of genes with unknown functions,
617 and to suggest roles for 'moonlighting' proteins that may be associated with unexpected
618 gene groupings.
619
620 ***The largest module includes the kstR regulon and is enriched for predicted sRNAs***

621 The *'turquoise' module* contains more than 1,000 expressed elements, with over 50% of
622 the hubs being predicted sRNAs. It contains 461 protein-coding genes, including 34 of the
623 71 KstR regulon genes and 52 transcription factors (Rustad et al., 2014). 26 of the 32 *kstR*
624 regulon genes found to be differentially expressed in Mtb grown with cholesterol versus
625 glycerol as the main carbon source (Pawełczyk et al., 2021) are found in the '*turquoise*'
626 module, with 15 of them hubs. The hubs also include 10 transcription factors and DNA
627 binding proteins, including IdeR, FurA, KstR, KstR2 and SigB, anti-sigma factor ResA,
628 two annotated sRNAs (mcr11/ncRv11264c, and mpr6/ncRv1222) and many predicted
629 ncRNA elements including 131 predicted sRNAs and 26 UTRs (Supp Table 2, 'CDS hubs,
630 'srna_hubs' tabs). The module has 46 complete predicted operons from OperomeDB
631 (Chetal & Janga, 2015), and the highest number of consecutive ORFs in the genome of all
632 the modules.
633
634 The size of the *'turquoise'* module, and the fact that it has resisted splintering into smaller
635 modules during the tree-cutting process, indicates that it includes many highly connected
636 gene operons involved in multiple interconnected stress response pathways. The module
637 shows enrichment for the GO terms 'regulation of transcription' and 'cholesterol catabolic
638 process', as well as for the KEGG pathway for steroid degradation (Supp Table 2, 'Module
639 Overview' tab). Despite the inclusion of genes linked specifically to cholesterol
640 metabolism, a significant correlation of the *'turquoise'* module with the cholesterol-
641 containing media condition was not established; rather, the module shows positive
642 correlations with hypoxia (bicor = 0.41, p_{adj} = 0.001), extended hypoxia (bicor = 0.035, p_{adj}
643 = 0.002) and stationary (bicor = 0.34, p_{adj} = 0.03) conditions, and a negative correlation
644 with exponential growth (bicor = -0.42, p_{adj} = 0.009) (Figure 3). Transcriptomic changes in
645 response to lipid degradation include many genes related to redox maintenance which are

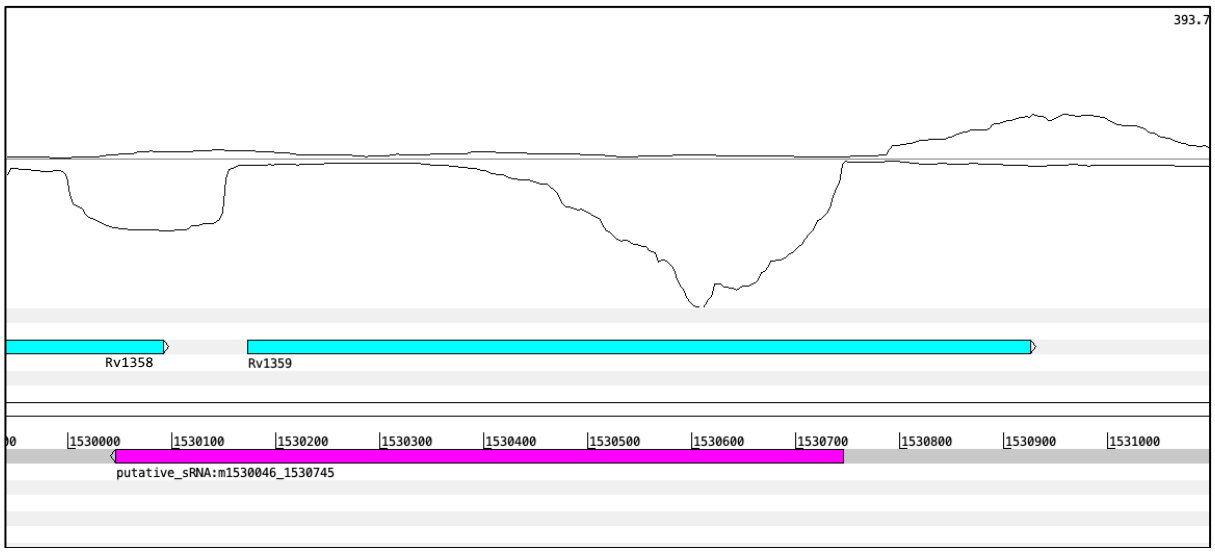
found in the module, including redox-sensing *whiB3* and *whiB4* (Larsson et al., 2012; Mehta & Singh, 2019).

Among the module hubs, are annotated sRNAs such as mcr11/ncRv11264c, which has been associated with dormancy and hypoxic conditions and shown to regulate the expression of genes related to the metabolic remodelling associated with persistence and slow growth states in Mtb (Girardin & McDonough, 2020). Other annotated ncRNA in *'turquoise'* include: mpr6 (ncRv1222), G2 (ncRv11689c), mcr16 (ncRv2243c), C8/4.5S RNA (ncRv13722Ac), and another experimentally-verified ncRNA, mrsI (ncRv11846) that was predicted as a somewhat longer transcript in this study (and in a previous study, (Arnvig et al., 2011)) which extends antisense to the gene Rv1847 (putative_sRNA:m2096739_2097122 / ncRv1847c). MrsI has been found to be upregulated in several growth states and stress conditions and is implicated in anticipatory regulation of iron acquisition (Gerrick et al., 2018). Most of the predicted sRNAs in the *'turquoise'* hubs are classified as antisense transcripts, with 82 having predicted TSSs within 10 nt of the start. In addition, 7 strictly 'intergenic' predicted sRNAs are among the hubs. Four of these have predicted TSS within 20 nucleotides of the start. (Supp Table 2, 'intergenic_putative_sRNAs' tab).

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

The *'black'* module showed positive correlation with the cholesterol media condition (bicor=0.54, p_{adj}=0.002) and negative correlation with low iron (bicor = -0.48, p_{adj} = 0.001) (Figure 3). Many protein-coding genes involved in detoxification pathways are hubs in the module, including several encoding transmembrane proteins such as the *mmpL5-mmpS5* efflux pump operon (Rv0676c-Rv0677c), as well as the next gene downstream, Rv0678,

672 which was identified as part of a 'core lipid response' in differential expression analysis in
673 lipid-rich media (Aguilar-Ayala, et al., 2017). The 5' UTR for Rv0677c and 3' UTRs for
674 Rv0676c and Rv0677c are also hubs. This operon is involved in siderophore transport and
675 expressed in cholesterol and lipid-rich environments (Aguilar-Ayala, et al., 2017;
676 Pawełczyk et al., 2021). Other detoxification-linked genes in the module, such as the ABC-
677 family transporter efflux system, Rv1216c-1219c and the operon including PPE53
678 (Rv3159c), Rv3160c and Rv3161c, have also been implicated in transcriptomic
679 remodelling in response to cholesterol (Aguilar-Ayala et al., 2017; Pawełczyk et al., 2021).
680
681 Among the hubs are three predicted antisense RNAs. One antisense RNA, ncRv1358c
682 (putative_sRNA:m1530046_1530745) has a TSS near its start and is found antisense to
683 Rv1359. Rv1359 and the upstream gene, Rv1358, on the opposite strand are very similar
684 to each other (43.7% identity in 197 aa overlap) and to another gene elsewhere in the
685 genome, Rv0891c (48.5% identity in 204 aa overlap) (Kapopoulou et al., 2011). All three
686 genes are possible LuxR family transcriptional regulators which are thought to be
687 involved in quorum-sensing adaptations and contain a probable ATP/GTP binding site
688 motif (Chen & Xie, 2011; Modlin et al., 2021). Expression of this antisense sRNA appears
689 to suppress the expression of the transcript on the opposite strand to varying degrees in
690 all conditions (Figure 6). Expression of a shorter transcript appears to begin inside the
691 Rv1359 ORF, where the transcript is not overlapped by the antisense transcript, possibly
692 utilising an internal TSS at 1530774.
693
694
695 Figure 6. Expression of antisense transcript putative_sRNA:m1530046_1530745 (magenta bar) seems to
696 suppress the expression of most of Rv1359 and Rv1358. An internal TSS exists inside the Rv1359 CDS at
697 1530774 near where expression begins. Sample SRR5689230 from PRJNA27860, exponential growth on
698 cholesterol and fatty acid media. Strand coverage using the 'second' read of each pair mapping to the
699 transcript strand, visualised using Artemis genome browser (Carver et al., 2012).
700



Two adjacent predictions, the 3' UTR for Rv1772 (putative_UTR:p2006948_2007063) followed by ncRv1773/ putative_sRNA:p2007213_2007377, are hubs in the *'black'* module. Together, they extend to overlap the antisense strand of a large portion of Rv1773c, a probable transcriptional regulator in the IclR-family, found in a different module *('navajowhite2')*. The 3' UTR for Rv1772 was previously identified as an abundant antisense transcript during exponential growth (Arnvig et al., 2011). The start of the predicted sRNA transcript has no known TSS and could instead be an extension of the predicted 3' UTR (Supp figure S11). (When combining predicted annotations from different datasets, long predicted UTRs that overlapped shorter sRNA predictions were discarded, see Methods). In *E.coli*, the IclR-family transcriptional regulators demonstrate both activating and repressing activities on targets such as multidrug efflux pumps and the *aceBAK* operon which regulates the glyoxylate shunt (Zhou et al., 2012). *Icl2a* (Rv1915) is one of the Mtb isoforms of the isocitrate/methylocitrate lyase gene, *aceA*, and may be regulated by Rv1773c, as seen in *E.coli*. Icl2a, Rv1772, its predicted UTR and the antisense RNA (ncRv1773) are all hubs in the *'black'* module. *Icl2a* has been observed to be upregulated with cholesterol as the sole carbon source and likely has a second function as part of the methylcitrate cycle to convert the fatty acid metabolites propionate and

propionyl CoA to less toxic compounds (Bhusal et al., 2017; Pawełczyk et al., 2021). Another predicted antisense RNA in the *'black'* hubs, ncRv0027c/ putative_sRNA:m31259_31967, has a TSS near its start (31967) and is antisense both to Rv0027 and Rv0028, conserved hypothetical proteins with no known function found in different modules.

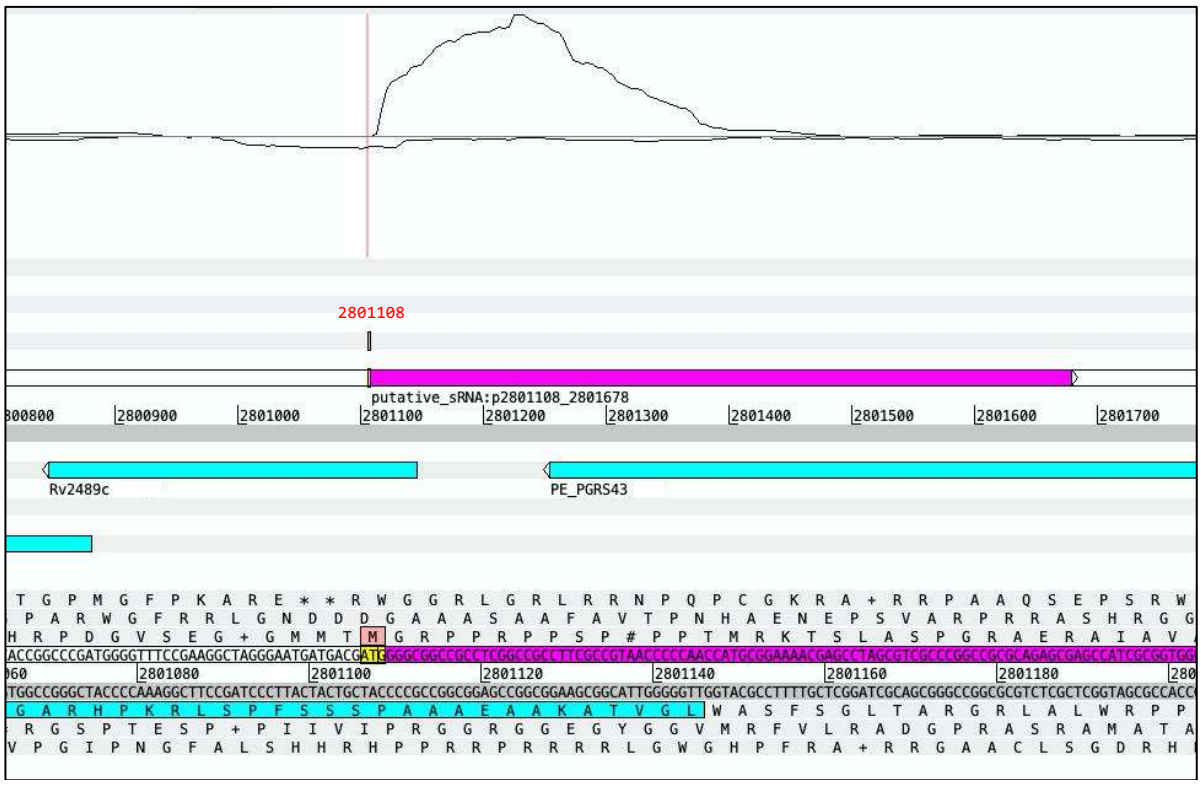
The module including transcriptional regulator whiB1 and genes of kstR2 regulon, links metal ion balance with cholesterol utilisation

The '*lightcyan*' module is significantly enriched for genes under control of another TetR-type repressor, KstR2, (one-sided Fisher's exact test, p_{adj} = 4.27e-06) with 7 of the 15 known regulon genes found in the module. KstR2-regulated genes are known to be involved in cholesterol utilisation (Kendall et al., 2010) and the protein-coding genes of this module were enriched for the COG term, 'lipid metabolism', and KEGG pathways, 'Biosynthesis of unsaturated fatty acids' and 'Fatty acid metabolism'. However this module did not significantly correlate with the cholesterol media condition. Instead, the ME was positively-correlated with the low iron condition (bicor = 0.59, p_{adj} = 9e-5) and negatively-correlated with the extended hypoxia condition (bicor = 0.41, p_{adj} = 0.008) (Figure 3). The correlation of this ME with the low iron condition implies that there are expressed elements within the module that are involved in iron homeostasis, possibly in tandem with adaptation to cholesterol. Intriguingly, one of the hub genes of this module encodes the redox-sensing transcriptional regulator, WhiB1. This transcription factor is known to be stimulated by a variety of stress conditions and *in vivo,* and binds an iron-sulfur cluster (Larsson et al., 2012; L. J. Smith et al., 2010).

The hubs of the *'lightcyan'* module include several predicted sRNAs, and the annotated sRNA, F6. F6/ncRv10243/SfdS is a sigF-dependent ncRNA which has been shown to be

747 induced in nutrient starvation, oxidative stress, acid stress (Arnvig & Young, 2009;
748 Houghton et al., 2021) and the fatty acid hypoxia model (Del Portillo et al., 2019). In
749 addition to being expressed from its own promoter, F6/SfdS has been proposed to be co-
750 transcribed with the upstream gene *fadA2* (Rv0243), a probable acetyl-CoA
751 acyltransferase; however, *fadA2* is clustered in a different module from SfdS, one
752 associated with iron acquisition (*'violet',* see below). One of the predicted sRNAs in the
753 module hubs is antisense transcript ncRv2489/putative_srna:p2801108_2801678 with a
754 TSS at 2801108. This overlaps the 3' end of PE-PGRS43 (Rv2490c) (Figure 7). There is a
755 short reading frame (30 nucleotides, 10 amino acids) initiating from a Methionine at this
756 TSS that suggests a possible dual-function sRNA or sORF with independent function. The
757 TSS for the predicted sRNA overlaps the 5' end of Rv2489c, a short, hypothetical 'alanine-
758 rich protein'. The TSSs for these convergently overlapping transcripts are 42 nts apart
759 (Rv2489c appears to be a leaderless transcript based on dRNA-seq and position of TSS)
760 and may involve RNAP collision if both are transcribed simultaneously. Therefore,
761 transcription of the predicted sRNA could impact either Rv2489c and/or PE-PGRS43
762 expression through two different mechanisms. Other hub sRNAs in *'lightcyan'* include
763 ncRv1450/putative_sRNA:p1630466_1631246, which has a TSS at 1630466 and is
764 antisense to the 3' end of PE-PGRS27 (Rv1450c) and putative_sRNA:p3936733_3936893
765 / ncRv3509 which includes a predicted TSS at 3936720 which overlaps the 3' end (and
766 predicted 3' UTR) of Rv3509c (*ilvX),* a probable acetolactate synthase (found in the *'violet'*
767 module).
768

769 Figure 7. Antisense sRNA, ncRv2489/putative_srna:p2801108_2801678, (magenta bar) overlaps two
770 transcripts and may encode a short peptide. TSS for sRNA indicated in red and corresponding amino acid
771 highlighted in pink. Sample SRR5689230 from PRJNA390669, exponential growth on cholesterol and fatty
772 acid media. Strand coverage using the 'second' read of each pair mapping to the transcript strand, visualised
773 using Artemis genome browser (Carver et al., 2012).



774
775

776

777 **The module best correlated with the low iron condition includes genes related to metal**
778 **ion and fatty acid homeostasis**

779 Another module that is positively correlated to the low iron condition is the *'violet'* module
780 (bicor=0.61, p_{adj}=6e-05, Figure 3). This module contains most of the ESX-3 genes (Rv0282-
781 Rv0292) related to siderophore-mediated iron (and zinc) uptake in Mtb (Serafini et al.,
782 2013; L. Zhang et al., 2020), with two of these representing hubs in the module. The gene
783 preceding the ESX-3 genes, Rv0281, a possible S-adenosylmethionine-dependent
784 methyltransferase involved in lipid metabolism (though its position in the genome would
785 suggest regulation could be linked to ESX-3 (Lunge et al., 2020)), is in the module, as well
786 as an ESX-5 gene, Rv1797 (*eccE5)*. The module also contains another Zur-regulated gene,
787 Rv0106, which is a potential zinc-ion transporter (Zondervan et al., 2018). Among the
788 hubs of the module are several genes related to lipid metabolism and fatty acid synthesis,
789 including: probable triglyceride transporter, Rv1410; the operon consisting of Rv0241c

790 (*htdX)*, Rv0242c (*fabG4)*, and Rv0243 (*fadA2)* (Dutta, 2018); and a gene involved in the
791 pentose phosphate pathway, *zwf2* (Rv1447c).
792
793 There are some well-connected ncRNAs in the *'violet'* module, including a predicted
794 antisense RNA to Rv0281, 'ncRv0281c'. This putative sRNA has a predicted TSS at the 5'
795 end and is transcribed divergently from Rv0282 *(eccA3)*. This is one of the rarer cases
796 where the antisense transcript and cognate protein-coding gene (Rv0281) are clustered in
797 the same module. The prevailing direction of transcription at this locus may be a result of
798 competition for RNAP binding at a bi-directional promoter in the predicted 5' UTR of
799 Rv0282 which also clusters in the module. Another predicted sRNA in the module,
800 ncRv3508/putative_sRNA:m3932046_3932369 has a predicted TSS at 3932369 and
801 transcribed opposite to a central region of Rv3508c, PE_PGRS54, a gene in the
802 *'darkolivegreen'* module which is enriched for PE/PPE genes (p_{adj} = 4.12e-09).
803
804 There are several UTRs in the module hubs, including a 3' UTR for the gene Rv1133c,
805 *metE;* the gene is found in another module, *'grey60'*. This might be an example of a sRNA
806 differentially transcribed or cleaved from the 3' UTR of a protein-coding gene. This UTR
807 was also identified as abundantly expressed in exponential culture (Arnvig et al., 2011).
808 There is a 3' UTR for Rv0292 *(eccE3,* also a hub in the *'violet'* module) that is antisense to
809 a large part of the 3' end of Rv0293c which has a converging orientation to Rv0292 (Supp
810 figure S12). Rv0293c is found in a different module *('turquoise')* and has a 3' UTR in the
811 '*lightsteelblue1'* module. The overlapping 3' ends of the genes could function to regulate
812 transcription, possibly to facilitate bi-directional termination brought about by RNAP
813 collision.
814
815

CONCLUSION

This paper presents a large-scale network analysis of over 7000 transcripts expressed by Mtb under a variety of conditions. The modules group together clusters of co-expressed protein-coding genes, as well as ncRNA transcripts predicted from RNA-Seq signals. The ncRNAs are unevenly distributed among modules; modules with the highest proportion of sRNAs correlated negatively to exponential growth and correlated positively to hypoxia and the extended hypoxia model (*'turquoise', 'blue', 'skyblue'*) (Figures 3 and 4), supporting the observation that high levels of ncRNA are associated with Mtb's response to hypoxic stress (Arnvig et al., 2011; Ignatov et al., 2015; Martini et al., 2019). The prevalence of antisense RNA in the hubs of these and other modules, and the fact that the complementary ORF is usually excluded, implicates antisense transcription as part of a regulation strategy through mechanisms of divergent transcription or in order to regulate mRNA stability (Vargas-Blanco & Shell, 2020; Warman et al., 2021); strategies that may differ among the members of the MTBC (Dinan, Adam M. et al., 2014). 3' UTR transcripts in modules distinct from their upstream ORF implies independent function from the ORF. sRNAs generated from 3' UTRs have been reported in other prokaryotes and evidence points to widespread mRNA processing that could release independent transcripts at the 3' end (Dar & Sorek, 2018; Desgranges et al., 2021; Updegrove et al., 2019; Wang et al., 2019). In compact bacterial genomes, 3' UTRs are also found to overlap other 3' UTRs in a converging transcription pattern which may provide a mechanism for regulating the expression or stability of either transcript.

The gene modules presented here are somewhat 'blunt-force instruments' applied to transcripts that are part of overlapping, coordinated responses to various environmental cues, but restricted to a single module grouping. Recent work exploring differentially expressed genes in response to various environmental conditions have revealed highly

842 integrated adaptation responses. In other words, a single environmental change, e.g.
843 hypoxia or growth on fatty acids or cholesterol, stimulates transcriptomic remodelling
844 across diverse cellular functions, perhaps acting as cues to stimulate anticipatory
845 pathways and ready the pathogen for the next challenge (Aguilar-Ayala et al., 2017; Eoh
846 et al., 2017; Gerrick et al., 2018). Confounders such as dual-function, 'moonlighting',
847 proteins may weaken the correlation of a module with a specific condition and may create
848 noise in otherwise well-connected modules. However, focussing on the best connected
849 transcripts in various modules can uncover the unexpected connections between genes of
850 diverse pathways.
851
852 Other methods of network analysis, such as those using deconvolution methods, allow
853 genes to be members of more than one module and are considered less 'noisy' than
854 clustering methods, such as WGCNA. However, these methods require extremely large
855 numbers of samples to perform well, may be subject to batch effect issues between
856 experimental datasets and characterise a limited proportion of the protein-coding
857 transcripts expressed by Mtb (Saelens et al., 2018; Yoo, et al., 2022). Predicting ncRNA
858 from different datasets involves a significant degree of quality control, parameter
859 adjustment and manual curation, limiting the number of datasets that could be included
860 in our analysis. Including more data would most likely strengthen the correlations with
861 certain conditions and improve the overall specificity of the modules. However, the work
862 presented here confirms that ncRNA are important players in adaptation responses, and
863 their associations with the protein-coding genes in their assigned modules provides
864 context for their activity.
865
866 The few modules discussed in depth in this paper represent a very limited snapshot of
867 this extensive co-expression network. Modules of interest can be identified by correlations

868 to experimental conditions, associated GO terms, functional categories, or gene group
869 enrichment. The supplementary tables provide information about the module association,
870 membership values, TSSs and for UTRs, the module membership of the adjacent ORFs
871 for each predicted ncRNA. This analysis can add context to the circumstances of
872 expression of previously identified ncRNAs and conserved hypothetical proteins by
873 associating their expression with functionally-characterised protein-coding genes in the
874 same module, as well as identifying novel ncRNA candidates for further investigation
875 such as structural analysis, target prediction and ultimately, experimental validation.
876

877 **References**

878 Aguilar-Ayala, D. A., Tilleman, L., Van Nieuwerburgh, F., Deforce, D., Palomino, J. C.,
879 Vandamme, P., Gonzalez-Y-Merchand, J. A., & Martin, A. (2017a). The
880 transcriptome of Mycobacterium tuberculosis in a lipid-rich dormancy model through
881 RNAseq analysis. *Scientific Reports*, *7*(1), 17665. https://doi.org/10.1038/s41598-
882 017-17751-x

883 Aguilar-Ayala, D. A., Tilleman, L., Van Nieuwerburgh, F., Deforce, D., Palomino, J. C.,
884 Vandamme, P., Gonzalez-Y-Merchand, J. A., & Martin, A. (2017b). The
885 transcriptome of Mycobacterium tuberculosis in a lipid-rich dormancy model through
886 RNAseq analysis. *Scientific Reports*, *7*(1), 17665–17665. PubMed.
887 https://doi.org/10.1038/s41598-017-17751-x

888 Ami, V. K. G., Balasubramanian, R., & Hegde, S. R. (2020). Genome-wide identification of
889 the context- dependent sRNA expression in Mycobacterium tuberculosis. *BMC*
890 *Genomics*, *21*(167), 1–12.

891 Arnvig, K. B., Comas, I., Thomson, N. R., Houghton, J., Boshoff, H. I., Croucher, N. J.,
892 Rose, G., Perkins, T. T., Parkhill, J., Dougan, G., & Young, D. B. (2011). Sequence-
893 Based Analysis Uncovers an Abundance of Non-Coding RNA in the Total
894 Transcriptome of Mycobacterium tuberculosis. *PLOS Pathogens*, *7*(11), e1002342.
895 https://doi.org/10.1371/journal.ppat.1002342

896 Arnvig, K. B., & Young, D. B. (2009). Identification of small RNAs in Mycobacterium
897 tuberculosis. *Molecular Microbiology*, *73*(3), 397–408.
898 https://doi.org/10.1111/j.1365-2958.2009.06777.x

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

902 Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P.,
903 Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L.,
904 Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G.
905 M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology.
906 *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556
907 Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and
908 Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society:*
909 *Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-
910 6161.1995.tb02031.x
911 Bhusal, R. P., Bashiri, G., Kwai, B. X. C., Sperry, J., & Leung, I. K. H. (2017). Targeting
912 isocitrate lyase for the treatment of latent tuberculosis. *Drug Discovery Today*, *22*(7),
913 1008–1016. https://doi.org/10.1016/j.drudis.2017.04.012
914 Bidnenko, E., & Bidnenko, V. (2018). Transcription termination factor Rho and microbial
915 phenotypic heterogeneity. *Current Genetics*, *64*(3), 541–546.
916 https://doi.org/10.1007/s00294-017-0775-7
917 Canestrari, J. G., Lasek-Nesselquist, E., Upadhyay, A., Rofaeil, M., Champion, M. M., Wade,
918 J. T., Derbyshire, K. M., & Gray, T. A. (2020). Polycysteine-encoding leaderless
919 short ORFs function as cysteine-responsive attenuators of operonic gene expression in
920 mycobacteria. *Molecular Microbiology*, *114*(1), 93–108.
921 https://doi.org/10.1111/mmi.14498
922 Chao, Y., Papenfort, K., Reinhardt, R., Sharma, C. M., & Vogel, J. (2012). An atlas of Hfq-
923 bound transcripts reveals 3' UTRs as a genomic reservoir of regulatory small RNAs.
924 *The EMBO Journal*, *31*(20), 4005–4019. https://doi.org/10.1038/emboj.2012.229
925 Chen, J., & Xie, J. (2011). Role and regulation of bacterial LuxR-like regulators. *Journal of*
926 *Cellular Biochemistry*, *112*(10), 2694–2702. https://doi.org/10.1002/jcb.23219

927 Chetal, K., & Janga, S. C. (2015). OperomeDB: A Database of Condition-Specific
928 Transcription Units in Prokaryotic Genomes. *BioMed Research International*, *2015*,
929 318217–318217. PubMed. https://doi.org/10.1155/2015/318217
930 Cortes, T., Schubert, O. T., Rose, G., Arnvig, K. B., Comas, I., Aebersold, R., & Young, D.
931 B. (2013). Genome-wide mapping of transcriptional start sites defines an extensive
932 leaderless transcriptome in Mycobacterium tuberculosis. *Cell Reports*, *5*(4), 1121–
933 1131. https://doi.org/10.1016/j.celrep.2013.10.031
934 Dar, D., Shamir, M., Mellin, J. R., Koutero, M., Stern-Ginossar, N., Cossart, P., & Sorek, R.
935 (2016). Term-seq reveals abundant ribo-regulation of antibiotics resistance in
936 bacteria. *Science*, *352*(6282), aad9822. https://doi.org/10.1126/science.aad9822
937 Dar, D., & Sorek, R. (2018). Bacterial noncoding RNAs excised from within protein-coding
938 transcripts. *MBio*, *9*(5). https://doi.org/10.1128/mBio.01730-18
939 Del Portillo, P., García-Morales, L., Menéndez, M. C., Anzola, J. M., Rodríguez, J. G.,
940 Helguera-Repetto, A. C., Ares, M. A., Prados-Rosales, R., Gonzalez-y-Merchand, J.
941 A., & García, M. J. (2019). Hypoxia Is Not a Main Stress When Mycobacterium
942 tuberculosis Is in a Dormancy-Like Long-Chain Fatty Acid Environment. *Frontiers in*
943 *Cellular and Infection Microbiology*, *8*, 449–449.
944 Desgranges, E., Barrientos, L., & Caldelari, I. (2021). The 3'UTR-derived sRNA RsaG
945 coordinates redox homeostasis and metabolism adaptation in response to glucose-6-
946 phosphate uptake in Staphylococcus aureus. *Molecular Microbiology*.
947 https://doi.org/10.1111/MMI.14845
948 D'Halluin, A., Polgar, P., Kipkorir, T., Patel, Z., Cortes, T., & Arnvig, K. B. (2022). Term-
949 seq reveals an abundance of conditional, Rho-dependent termination in
950 Mycobacterium tuberculosis. *BioRxiv*, 2022.06.01.494293.
951 https://doi.org/10.1101/2022.06.01.494293

952 Dinan, Adam M., Tong, Pin, Lohan, Amanda J., Conlon, Kevin M., Miranda-CasoLuengo
953 Aleksandra A., Malone, Kerri M., Gordon, Stephen V., & Loftus, Brendan J. (2014).
954 Relaxed Selection Drives a Noisy Noncoding Transcriptome in Members of the
955 Mycobacterium tuberculosis Complex. *MBio*, *5*(4), e01169-14.
956 https://doi.org/10.1128/mBio.01169-14
957 Du, P., Sohaskey, C. D., & Shi, L. (2016). Transcriptional and physiological changes during
958 Mycobacterium tuberculosis reactivation from non-replicating persistence. *Frontiers*
959 *in Microbiology*, *7*(AUG). https://doi.org/10.3389/fmicb.2016.01346
960 Dutta, D. (2018). Advance in Research on Mycobacterium tuberculosis FabG4 and Its
961 Inhibitor. *Frontiers in Microbiology*, *9*.
962 https://www.frontiersin.org/article/10.3389/fmicb.2018.01184
963 Eoh, H., Wang, Z., Layre, E., Rath, P., Morris, R., Branch Moody, D., & Rhee, K. Y. (2017).
964 Metabolic anticipation in Mycobacterium tuberculosis. *Nature Microbiology*, *2*(8),
965 17084. https://doi.org/10.1038/nmicrobiol.2017.84
966 Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., & Koonin, E.
967 V. (2021). COG database update: Focus on microbial diversity, model organisms, and
968 widespread pathogens. *Nucleic Acids Research*, *49*(D1), D274–D281.
969 https://doi.org/10.1093/nar/gkaa1018
970 Gerrick, E. R., Barbier, T., Chase, M. R., Xu, R., François, J., Lin, V. H., Szucs, M. J., Rock,
971 J. M., Ahmad, R., Tjaden, B., Livny, J., & Fortune, S. M. (2018). Small RNA
972 profiling in mycobacterium tuberculosis identifies mrsi as necessary for an
973 anticipatory iron sparing response. *Proceedings of the National Academy of Sciences*
974 *of the United States of America*, *115*(25), 6464–6469.
975 https://doi.org/10.1073/pnas.1718003115

976 Girardin, R. C., & McDonough, K. A. (2020). Small RNA Mcr11 requires the transcription
977 factor AbmR for stable expression and regulates genes involved in the central
978 metabolism of Mycobacterium tuberculosis. *Molecular Microbiology*, *113*(2), 504–
979 520. https://doi.org/10.1111/mmi.14436
980 Gonzalo-Asensio, J., Mostowy, S., Harders-Westerveen, J., Huygen, K., Hernández-Pando,
981 R., Thole, J., Behr, M., Gicquel, B., & Martín, C. (2008). PhoP: a missing piece in the
982 intricate puzzle of Mycobacterium tuberculosis virulence. *PloS One*, *3*(10), e3496–
983 e3496. PubMed. https://doi.org/10.1371/journal.pone.0003496
984 Houghton, Joanna, Rodgers, Angela, Rose, Graham, D'Halluin, Alexandre, Kipkorir, Terry,
985 Barker, Declan, Waddell, Simon J., Arnvig, Kristine B., & Oglesby, Amanda G.
986 (2021). The Mycobacterium tuberculosis sRNA F6 Modifies Expression of Essential
987 Chaperonins, GroEL2 and GroES. *Microbiology Spectrum*, *9*(2), e01095-21.
988 https://doi.org/10.1128/Spectrum.01095-21
989 Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009a). Bioinformatics enrichment tools:
990 Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids*
991 *Research*, *37*(1), 1–13. https://doi.org/10.1093/nar/gkn923
992 Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009b). Systematic and integrative
993 analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*,
994 *4*(1), 44–57. https://doi.org/10.1038/nprot.2008.211
995 Ignatov, D. V., Salina, E. G., Fursov, M. V., Skvortsov, T. A., Azhikina, T. L., &
996 Kaprelyants, A. S. (2015). Dormant non-culturable Mycobacterium tuberculosis
997 retains stable low-abundant mRNA. *BMC Genomics*, *16*(1), 954.
998 https://doi.org/10.1186/s12864-015-2197-6
999 Jiang, J., Lin, C., Zhang, J., Wang, Y., Shen, L., Yang, K., Xiao, W., Li, Y., Zhang, L., &
1000 Liu, J. (2020). Transcriptome Changes of Mycobacterium marinum in the Process of

1001 Resuscitation From Hypoxia-Induced Dormancy. *Frontiers in Genetics*, *10*(February),
1002 1–13. https://doi.org/10.3389/fgene.2019.01359
1003 Jiang, J., Sun, X., Wu, W., Li, L., Wu, H., Zhang, L., Yu, G., & Li, Y. (2016). Construction
1004 and application of a co-expression network in Mycobacterium tuberculosis. *Scientific*
1005 *Reports*, *6*(March 2015), 1–18. https://doi.org/10.1038/srep28422
1006 Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., &
1007 Lempicki, R. A. (2012). DAVID-WS: a stateful web service to facilitate gene/protein
1008 list analysis. *Bioinformatics*, *28*(13), 1805–1806.
1009 https://doi.org/10.1093/bioinformatics/bts251
1010 Ju, X., Li, D., & Liu, S. (2019). Full-length RNA profiling reveals pervasive bidirectional
1011 transcription terminators in bacteria. *Nature Microbiology*, *4*(11), 1907–1918.
1012 https://doi.org/10.1038/s41564-019-0500-z
1013 Kanehisa, M., Sato, Y., & Kawashima, M. (2022). KEGG mapping tools for uncovering
1014 hidden features in biological data. *Protein Science*, *31*(1), 47–53.
1015 https://doi.org/10.1002/pro.4172
1016 Kapopoulou, A., Lew, J. M., & Cole, S. T. (2011). The MycoBrowser portal: A
1017 comprehensive and manually annotated resource for mycobacterial genomes.
1018 *Tuberculosis*, *91*(1), 8–13. https://doi.org/10.1016/j.tube.2010.09.006
1019 Kendall, S. L., Burgess, P., Balhana, R., Withers, M., Ten Bokum, A., Lott, J. S., Gao, C.,
1020 Uhia-Castro, I., & Stoker, N. G. (2010). Cholesterol utilization in mycobacteria is
1021 controlled by two TetR-type transcriptional regulators: KstR and kstR2.
1022 *Microbiology*, *156*(5), 1362–1371. https://doi.org/10.1099/mic.0.034538-0
1023 Kendall, S. L., Withers, M., Soffair, C. N., Moreland, N. J., Gurcha, S., Sidders, B., Frita, R.,
1024 Ten Bokum, A., Besra, G. S., Lott, J. S., & Stoker, N. G. (2007). A highly conserved
1025 transcriptional repressor controls a large regulon involved in lipid degradation in

1026 Mycobacterium smegmatis and Mycobacterium tuberculosis. *Molecular*
1027 *Microbiology*, *65*(3), 684–699. https://doi.org/10.1111/j.1365-2958.2007.05827.x
1028 Kipkorir, Terry, Mashabela, Gabriel T., de Wet, Timothy J., Koch, Anastasia, Dawes
1029 Stephanie S., Wiesner, Lubbe, Mizrahi, Valerie, Warner, Digby F., & Henkin, Tina
1030 M. (2021). De Novo Cobalamin Biosynthesis, Transport, and Assimilation and
1031 Cobalamin-Mediated Regulation of Methionine Biosynthesis in Mycobacterium
1032 smegmatis. *Journal of Bacteriology*, *203*(7), e00620-20.
1033 https://doi.org/10.1128/JB.00620-20
1034 Lamichhane, G., Arnvig, K. B., & McDonough, K. A. (2013). Definition and annotation of
1035 (myco)bacterial non-coding RNA. *Tuberculosis*, *93*(1), 26–29.
1036 https://doi.org/10.1016/j.tube.2012.11.010
1037 Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation
1038 network analysis. *BMC Bioinformatics*, *9*. https://doi.org/10.1186/1471-2105-9-559
1039 Larsson, C., Luna, B., Ammerman, N. C., Maiga, M., Agarwal, N., & Bishai, W. R. (2012).
1040 Gene Expression of Mycobacterium tuberculosis Putative Transcription Factors
1041 whiB1-7 in Redox Environments. *PLOS ONE*, *7*(7), e37516.
1042 https://doi.org/10.1371/journal.pone.0037516
1043 Li, Heng. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-*
1044 *MEM*. https://doi.org/10.48550/arXiv.1303.3997
1045 Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W.-H., Bork, P., Glass,
1046 J. I., Serrano, L., & Lluch-Senar, M. (2016). Bacterial antisense RNAs are mainly the
1047 product of transcriptional noise. *Science Advances*, *2*(3), e1501363.
1048 https://doi.org/10.1126/sciadv.1501363

1049 Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and
1050 dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21.
1051 https://doi.org/10.1186/s13059-014-0550-8
1052 Lu, L., Wei, R., Bhakta, S., Waddell, S. J., & Boix, E. (2021). Weighted gene co-expression
1053 network analysis to identify key modules and hub genes associated with
1054 paucigranulocytic asthma. *Antibiotics*, *10*(97).
1055 https://doi.org/10.3390/antibiotics10020097
1056 Lunge, A., Gupta, R., Choudhary, E., & Agarwal, N. (2020). The unfoldase ClpC1 of
1057 Mycobacterium tuberculosis regulates the expression of a distinct subset of proteins
1058 having intrinsically disordered termini. *Journal of Biological Chemistry*, *295*(28),
1059 9455–9473. https://doi.org/10.1074/jbc.RA120.013456
1060 Mai, J., Rao, C., Watt, J., Sun, X., Lin, C., Zhang, L., & Liu, J. (2019). Mycobacterium
1061 tuberculosis 6C sRNA binds multiple mRNA targets via C-rich loops independent of
1062 RNA chaperones. *Nucleic Acids Research*, *47*(8), 4292–4307.
1063 https://doi.org/10.1093/nar/gkz149
1064 Martini, M. C., Zhou, Y., Sun, H., & Shell, S. S. (2019). Defining the Transcriptional and
1065 Post-transcriptional Landscapes of Mycobacterium smegmatis in Aerobic Growth and
1066 Hypoxia. In *Frontiers in Microbiology* (Vol. 10).
1067 https://www.frontiersin.org/article/10.3389/fmicb.2019.00591
1068 Mehta, M., & Singh, A. (2019). Mycobacterium tuberculosis WhiB3 maintains redox
1069 homeostasis and survival in response to reactive oxygen and nitrogen species. *Free*
1070 *Radical Biology and Medicine*, *131*, 50–58.
1071 https://doi.org/10.1016/j.freeradbiomed.2018.11.032
1072 Menendez-Gil, P., Caballero, C., Catalan-Moreno, A., Irurzun, N., Barrio-Hernandez, I.,
1073 Caldelari, I., & Toledo-Arana, A. (2020). Differential evolution in 3'UTRs leads to

specific gene expression in Staphylococcus. *Nucleic Acids Research*, *48*. https://doi.org/10.1093/nar/gkaa047

Menendez-Gil, P., & Toledo-Arana, A. (2021). Bacterial 3′UTRs: A Useful Resource in Post-transcriptional Regulation. *Frontiers in Molecular Biosciences*, *7*. https://www.frontiersin.org/article/10.3389/fmolb.2020.617633

Modlin, S. J., Afif, E., Deepika, G., Zlotnicki, A. M., Dillon, N. A., Dhillon, N., Kuo, N., Robinhold, C., Chan, C. K., Baughn, A. D., & Valafar, F. (2021). Structure-Aware Mycobacterium tuberculosis Functional Annotation Uncloaks Resistance, Metabolic, and Virulence Genes. *MSystems*, *0*(0), e00673-21. https://doi.org/10.1128/mSystems.00673-21

Moores, A., Riesco, A. B., Schwenk, S., & Arnvig, K. B. (2017). Expression, maturation and turnover of DrrS, an unusually stable, DosR regulated small RNA in Mycobacterium tuberculosis. *PLOS ONE*, *12*(3), e0174079. https://doi.org/10.1371/journal.pone.0174079

Ozuna, A., Liberto, D., Joyce, R. M., Arnvig, K. B., & Nobeli, I. (2019). baerhunter: An R package for the discovery and analysis of expressed non-coding regions in bacterial RNA-seq data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz643

Pandey Amit K. & Sassetti Christopher M. (2008). Mycobacterial persistence requires the utilization of host cholesterol. *Proceedings of the National Academy of Sciences*, *105*(11), 4376–4380. https://doi.org/10.1073/pnas.0711159105

Pawełczyk, J., Brzostek, A., Minias, A., Płociński, P., Rumijowska-Galewicz, A., Strapagiel, D., Zakrzewska-Czerwińska, J., & Dziadek, J. (2021). Cholesterol-dependent transcriptome remodeling reveals new insight into the contribution of cholesterol to Mycobacterium tuberculosis pathogenesis. *Scientific Reports*, *11*(1), 12396. https://doi.org/10.1038/s41598-021-91812-0

1099 Ponath, F., Hör, J., & Vogel, J. (2022). An overview of gene regulation in bacteria by small
1100 RNAs derived from mRNA 3' ends. *FEMS Microbiology Reviews*, fuac017.
1101 https://doi.org/10.1093/femsre/fuac017
1102 Puniya, B. L., Kulshreshtha, D., Verma, S. P., Kumar, S., & Ramachandran, S. (2013).
1103 Integrated gene co-expression network analysis in the growth phase of
1104 Mycobacterium tuberculosis reveals new potential drug targets. *Molecular*
1105 *BioSystems*, *9*(11), 2798–2815. https://doi.org/10.1039/c3mb70278b
1106 Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015).
1107 Limma powers differential expression analyses for RNA-sequencing and microarray
1108 studies. *Nucleic Acids Research*, *43*(7), e47–e47. https://doi.org/10.1093/nar/gkv007
1109 Rustad, T. R., Harrell, M. I., Liao, R., & Sherman, D. R. (2008). The enduring hypoxic
1110 response of Mycobacterium tuberculosis. *PLoS ONE*, *3*(1), 1–8.
1111 https://doi.org/10.1371/journal.pone.0001502
1112 Rustad, T. R., Minch, K. J., Ma, S., Winkler, J. K., Hobbs, S., Hickey, M., Brabant, W.,
1113 Turkarslan, S., Price, N. D., Baliga, N. S., & Sherman, D. R. (2014). Mapping and
1114 manipulating the Mycobacterium tuberculosis transcriptome using a transcription
1115 factor overexpression-derived regulatory network. *Genome Biology*, *15*(11), 502.
1116 https://doi.org/10.1186/s13059-014-0502-3
1117 Saelens, W., Cannoodt, R., & Saeys, Y. (2018). A comprehensive evaluation of module
1118 detection methods for gene expression data. *Nature Communications*, *9*(1), 1090.
1119 https://doi.org/10.1038/s41467-018-03424-4
1120 Sáenz-Lahoya S., Bitarte N., García B., Burgui S., Vergara-Irigaray M., Valle J., Solano C.,
1121 Toledo-Arana A., & Lasa I. (2019). Noncontiguous operon is a genetic organization
1122 for coordinating bacterial gene expression. *Proceedings of the National Academy of*
1123 *Sciences*, *116*(5), 1733–1738. https://doi.org/10.1073/pnas.1812746116

1124 Sawyer, E. B., Phelan, J. E., Clark, T. G., & Cortes, T. (2021). A snapshot of translation in
1125 Mycobacterium tuberculosis during exponential growth and nutrient starvation
1126 revealed by ribosome profiling. *Cell Reports*, *34*(5).
1127 https://doi.org/10.1016/j.celrep.2021.108695
1128 Schwenk, S., & Arnvig, K. B. (2018). Regulatory RNA in Mycobacterium tuberculosis, back
1129 to basics. *Pathogens and Disease*, *76*(4). https://doi.org/10.1093/femspd/fty035
1130 Serafini, A., Pisu, D., Palù, G., Rodriguez, G. M., & Manganelli, R. (2013). The ESX-3
1131 Secretion System Is Necessary for Iron and Zinc Homeostasis in Mycobacterium
1132 tuberculosis. *PLoS ONE*, *8*(10), 1–15. https://doi.org/10.1371/journal.pone.0078351
1133 Shell, S. S., Wang, J., Lapierre, P., Mir, M., Chase, M. R., Pyle, M. M., Gawande, R.,
1134 Ahmad, R., Sarracino, D. A., Ioerger, T. R., Fortune, S. M., Derbyshire, K. M., Wade,
1135 J. T., & Gray, T. A. (2015). Leaderless Transcripts and Small Proteins Are Common
1136 Features of the Mycobacterial Translational Landscape. *PLOS Genetics*, *11*(11),
1137 e1005641. https://doi.org/10.1371/journal.pgen.1005641
1138 Šiková, M., Janoušková, M., Ramaniuk, O., Páleníková, P., Pospíšil, J., Bartl, P., Suder, A.,
1139 Pajer, P., Kubičková, P., Pavliš, O., Hradilová, M., Vítovská, D., Šanderová, H.,
1140 Převorovský, M., Hnilicová, J., & Krásný, L. (2019). Ms1 RNA increases the amount
1141 of RNA polymerase in Mycobacterium smegmatis. *Molecular Microbiology*, *111*(2),
1142 354–372. https://doi.org/10.1111/mmi.14159
1143 Singh Prabhat Ranjan, Vijjamarri Anil Kumar, Sarkar Dibyendu, & Federle Michael J.
1144 (2020). Metabolic Switching of Mycobacterium tuberculosis during Hypoxia Is
1145 Controlled by the Virulence Regulator PhoP. *Journal of Bacteriology*, *202*(7),
1146 e00705-19. https://doi.org/10.1128/JB.00705-19

1147 Smith, C., Canestrari, J. G., Wang, A. J., Champion, M. M., Derbyshire, K. M., Gray, T. A.,
1148 & Wade, J. T. (2022). Pervasive translation in Mycobacterium tuberculosis. *ELife*, *11*,
1149 e73980. https://doi.org/10.7554/eLife.73980
1150 Smith, L. J., Stapleton, M. R., Fullstone, G. J. M., Crack, J. C., Thomson, A. J., Le Brun, N.
1151 E., Hunt, D. M., Harvey, E., Adinolfi, S., Buxton, R. S., & Green, J. (2010).
1152 Mycobacterium tuberculosis WhiB1 is an essential DNA-binding protein with a nitric
1153 oxide-sensitive iron–sulfur cluster. *Biochemical Journal*, *432*(3), 417–427.
1154 https://doi.org/10.1042/BJ20101440
1155 Solans, L., Gonzalo-Asensio, J., Sala, C., Benjak, A., Uplekar, S., Rougemont, J., Guilhot,
1156 C., Malaga, W., Martín, C., & Cole, S. T. (2014). The PhoP-Dependent ncRNA Mcr7
1157 Modulates the TAT Secretion System in Mycobacterium tuberculosis. *PLOS*
1158 *Pathogens*, *10*(5), e1004183. https://doi.org/10.1371/journal.ppat.1004183
1159 Soto-Ramirez, M. D., Aguilar-Ayala, D. A., Garcia-Morales, L., Rodriguez-Peredo, S. M.,
1160 Badillo-Lopez, C., Rios-Muñiz, D. E., Meza-Segura, M. A., Rivera-Morales, G. Y.,
1161 Leon-Solis, L., Cerna-Cortes, J. F., Rivera-Gutierrez, S., Helguera-Repetto, A. C., &
1162 Gonzalez-y-Merchand, J. A. (2017). Cholesterol plays a larger role during
1163 Mycobacterium tuberculosis in vitro dormancy and reactivation than previously
1164 suspected. *Tuberculosis*, *103*(November 2020), 1–9.
1165 https://doi.org/10.1016/j.tube.2016.12.004
1166 Stiens, J., Arnvig, K. B., Kendall, S. L., & Nobeli, I. (2022). Challenges in defining the
1167 functional, non-coding, expressed genome of members of the Mycobacterium
1168 tuberculosis complex. *Molecular Microbiology*, *117*(1), 20–31.
1169 https://doi.org/10.1111/mmi.14862

1170 The Gene Ontology Consortium. (2021). The Gene Ontology resource: Enriching a GOld
1171 mine. *Nucleic Acids Research*, *49*(D1), D325–D334.
1172 https://doi.org/10.1093/nar/gkaa1113
1173 Toledo-Arana, A., & Lasa, I. (2020). Advances in bacterial transcriptome understanding:
1174 From overlapping transcription to the excludon concept. *Molecular Microbiology*,
1175 *113*(3), 593–602. https://doi.org/10.1111/mmi.14456
1176 Updegrove, T. B., Kouse, A. B., Bandyra, K. J., & Storz, G. (2019). Stem-loops direct precise
1177 processing of 3′ UTR-derived small RNA MicL. *Nucleic Acids Research*, *47*(3),
1178 1482–1492. https://doi.org/10.1093/nar/gky1175
1179 Vargas-Blanco, D. A., & Shell, S. S. (2020). Regulation of mRNA Stability During Bacterial
1180 Stress Responses. *Frontiers in Microbiology*, *11*(September).
1181 https://doi.org/10.3389/fmicb.2020.02111
1182 Voskuil, M. I., Visconti, K. C., & Schoolnik, G. K. (2004). Mycobacterium tuberculosis gene
1183 expression during adaptation to stationary phase and low-oxygen dormancy.
1184 *Tuberculosis*, *84*(3–4), 218–227. https://doi.org/10.1016/j.tube.2004.02.003
1185 Wade, J. T., & Grainger, D. C. (2014). Pervasive transcription: Illuminating the dark matter
1186 of bacterial transcriptomes. *Nature Reviews Microbiology*, *12*(9), 647–653.
1187 https://doi.org/10.1038/nrmicro3316
1188 Wang, X., Monford Paul Abishek, N., Jeon, H. J., Lee, Y., He, J., Adhya, S., & Lim, H. M.
1189 (2019). Processing generates 3′ ends of RNA masking transcription termination
1190 events in prokaryotes. *Proceedings of the National Academy of Sciences of the United*
1191 *States of America*, *116*(10), 4440–4445. https://doi.org/10.1073/pnas.1813181116
1192 Warman, E. A., Forrest, D., Guest, T., Haycocks, J. J. R. J., Wade, J. T., & Grainger, D. C.
1193 (2021). Widespread divergent transcription from bacterial and archaeal promoters is a

consequence of DNA-sequence symmetry. *Nature Microbiology*, *6*(6), 746–756. https://doi.org/10.1038/s41564-021-00898-9

Warner, D. F., Savvi, S., Mizrahi, V., & Dawes, S. S. (2007). A Riboswitch Regulates Expression of the Coenzyme B12-Independent Methionine Synthase in Mycobacterium tuberculosis: Implications for Differential Methionine Synthase Function in Strains H37Rv and CDC1551. *Journal of Bacteriology*, *189*(9), 3655 LP – 3659. https://doi.org/10.1128/JB.00040-07

World Health Organization. (2021, October 14). *Tuberculosis Fact Sheet*. Tuberculosis. https://www.who.int/news-room/fact-sheets/detail/tuberculosis

Yoo, Reo, Rychel, Kevin, Poudel, Saugat, Al-bulushi, Tahani, Yuan Yuan, Chauhan, Siddharth, Lamoureux, Cameron, Palsson, Bernhard O., Sastry, Anand, & Tringe, Susannah Green. (2022). Machine Learning of All Mycobacterium tuberculosis H37Rv RNA-seq Data Reveals a Structured Interplay between Metabolism, Stress Response, and Infection. *MSphere*, *0*(0), e00033-22. https://doi.org/10.1128/msphere.00033-22

Zhang, B., & Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1). https://doi.org/10.2202/1544-6115.1128

Zhang, L., Hendrickson, R. C., Meikle, V., Lefkowitz, E. J., Ioerger, T. R., & Niederweis, M. (2020). Comprehensive analysis of iron utilization by Mycobacterium tuberculosis. *PLOS Pathogens*, *16*(2), e1008337. https://doi.org/10.1371/journal.ppat.1008337

Zhou, Y., Huang, H., Zhou, P., & Xie, J. (2012). Molecular mechanisms underlying the function diversity of transcriptional factor IclR family. *Cellular Signalling*, *24*(6), 1270–1275. https://doi.org/10.1016/j.cellsig.2012.02.008

1218 Zondervan, N. A., Van Dam, J. C. J., Schaap, P. J., Martins dos Santos, V. A. P., & Suarez-
1219 Diez, M. (2018). Regulation of Three Virulence Strategies of Mycobacterium
1220 tuberculosis: A Success Story. *International Journal of Molecular Sciences*, *19*(2).
1221 https://doi.org/10.3390/ijms19020347
1222