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

Authors

Jennifer Stiens[1], Yen Yi Tan[1], Rosanna Joyce[1], Kristine B. Arnvig[2], Sharon L. Kendall[3], Irene Nobeli[1]

[1]Institute of Structural and Molecular Biology, Biological Sciences, Birkbeck, University of London, London, UK

[2]Institute of Structural and Molecular Biology, Division of Biosciences, University College London, London, UK

[3]Centre for Emerging, Endemic and Exotic Diseases, Pathobiology and Population Sciences, Royal Veterinary College, Hatfield, UK

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.

38 ## Abbreviations

39 CDS, coding sequence

40 ME, module eigengene

41 MM, module membership

42 Mtb, *Mycobacterium tuberculosis*

43 MTBC, *Mycobacterium tuberculosis* complex

44 ncRNA, non-coding RNA

45 ORF, open reading frame

46 RNA-seq, RNA sequencing

47 RNAP, RNA polymerase
48
49 sORF, short open reading frame

50 sRNA, short non-coding RNA

51 TSS, transcription start site

52 TTS, transcription termination site

53 UTR, untranslated region

54 WGCNA, weighted gene co-expression analysis

55

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, including non-coding RNA, short, unannotated ORFs and untranslated regions at the 5' and 3' end of protein-coding sequences, many of which are poorly annotated and understood. In this paper, we extend our focus to include 'non-coding' RNA (ncRNA), here referring to non-ribosomal RNA transcripts not known to be translated into peptides, such as short RNAs (sRNAs) acting on either distant or antisense mRNA targets and the expressed untranslated regions (UTRs) flanking coding regions (which may also contain short open reading frames (sORFs) upstream from coding regions). Non-coding RNA can alter the abundance of RNA and proteins by controlling mRNA stability, processing and access to ribosome binding sites. Discovering the contribution of the non-coding genome to specific adaptation-response pathways may improve our ability to 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; K. B. Arnvig et al., 2011; Gerrick et al., 2018; Ignatov et al., 2015). Several mycobacterial ncRNA transcripts (particularly, sRNA) have been extensively studied and found to be associated with regulatory systems controlling adaptation to stress conditions or growth phase, linked to virulence pathways and access to lipid media (Arnvig et al., 2011; Gerrick et al., 2018; Girardin & McDonough, 2020; Mai et al., 2019; Moores et al., 2017; Solans et al., 2014). 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 of ncRNA in Mtb has progressed using both molecular biology methods and high-throughput sequence-based approaches (reviewed in Schwenk & Arnvig, 2018) but uncovering the regulation and actions of a particular ncRNA is experimentally expensive and very few have been fully-characterised. 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 are impeded by non-standardised nomenclature, different standards of experimental validation, incomplete reference annotations (especially for the closely-related 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).

Using RNA-sequencing (RNA-seq) data to predict ncRNA in the compact Mtb genome is challenging. Paradoxically, more sensitive, high-depth sequencing can make it more

110 difficult to identify the small, low-abundance, functional transcripts above stochastic gene
111 expression and technical noise. Parameters of detection must therefore be carefully
112 considered for each dataset to account for variation in expression levels. Though RNA-
113 seq-based ncRNA prediction algorithms are often assumed to overpredict putative
114 ncRNAs, especially at the 5' and 3' ends of coding genes, there are biological and technical
115 reasons for detecting abundant signal in the unannotated regions of the genome.
116 Ribosome profiling (Ribo-seq) methods that sequence the ribosome-protected fragments of
117 mRNA have identified actively translated RNA in the 5' UTRs of annotated protein-coding
118 mRNA transcripts (Canestrari et al., 2020; D'Halluin et al., 2022; Sawyer et al., 2021;
119 Shell et al., 2015; Smith et al., 2022). These unannotated sORFs may represent functional
120 peptides or function to regulate the translation of the downstream transcript; however, it
121 is impossible to tell the difference between a putative ncRNA and a sORF from RNA-seq
122 signal alone. Additionally, the 3' ends in mycobacterial RNA-seq data often lack clear
123 signal termination (Dar et al., 2016; D'Halluin et al., 2022; Lejars et al., 2019) and
124 processing of transcripts at the 3' end may be the norm (Wang et al., 2019). Finally,
125 polycistronic transcripts often include non-coding sequence between the genes of an
126 operon, and this may contain functional elements and/or processing sites (Martini et al.,
127 2019).
128
129 The location of a transcription start site (TSS) in the 5' end of a predicted transcript
130 supports the biological relevance of a predicted ncRNA. However, the available lists of
131 Mtb TSS sites (Cortes et al., 2013; Shell et al., 2015) have so far been mapped only in
132 starvation and exponential growth and may not include TSSs that are expressed under
133 different experimental conditions. New TSS maps, published subsequent to this analysis
134 may increase the number of predicted transcripts with a TSS (D'Halluin et al., 2022).
135 Furthermore, functional ncRNA elements generated from the 3' UTRs of coding genes

136 through RNase processing would presumably lack a TSS. 3' UTRs that are functionally
137 independent from their cognate coding sequence (CDS) have been identified in other
138 bacteria (Desgranges et al., 2021; Menendez-Gil et al., 2020; Ponath et al., 2022).
139 Therefore, it is important to consider predicted UTRs as separate annotated elements
140 from protein-coding transcripts when quantifying differential expression.
141
142 To include a complete picture of the interaction of the non-coding genome with coding
143 genes involved in adaptation pathways, we have generated a novel set of ncRNA sequence-
144 based predictions (sRNAs and UTRs) from publicly available datasets using our in-house
145 software package, *baerhunter* (Ozuna et al., 2019). Some of these predicted non-coding
146 transcripts overlap with those of previous studies, but many represent novel predictions.
147 The expression of these transcripts is quantified along with the protein-coding genes and
148 used in network analysis to provide a more complete picture of the functional groupings
149 involved in adaptation to environmental changes. Including a variety of culture conditions
150 that replicate aspects of the host environment improves the chances that the expression
151 of any ncRNA that is restricted to one or more conditions is included in the network (Ami
152 et al., 2020).
153

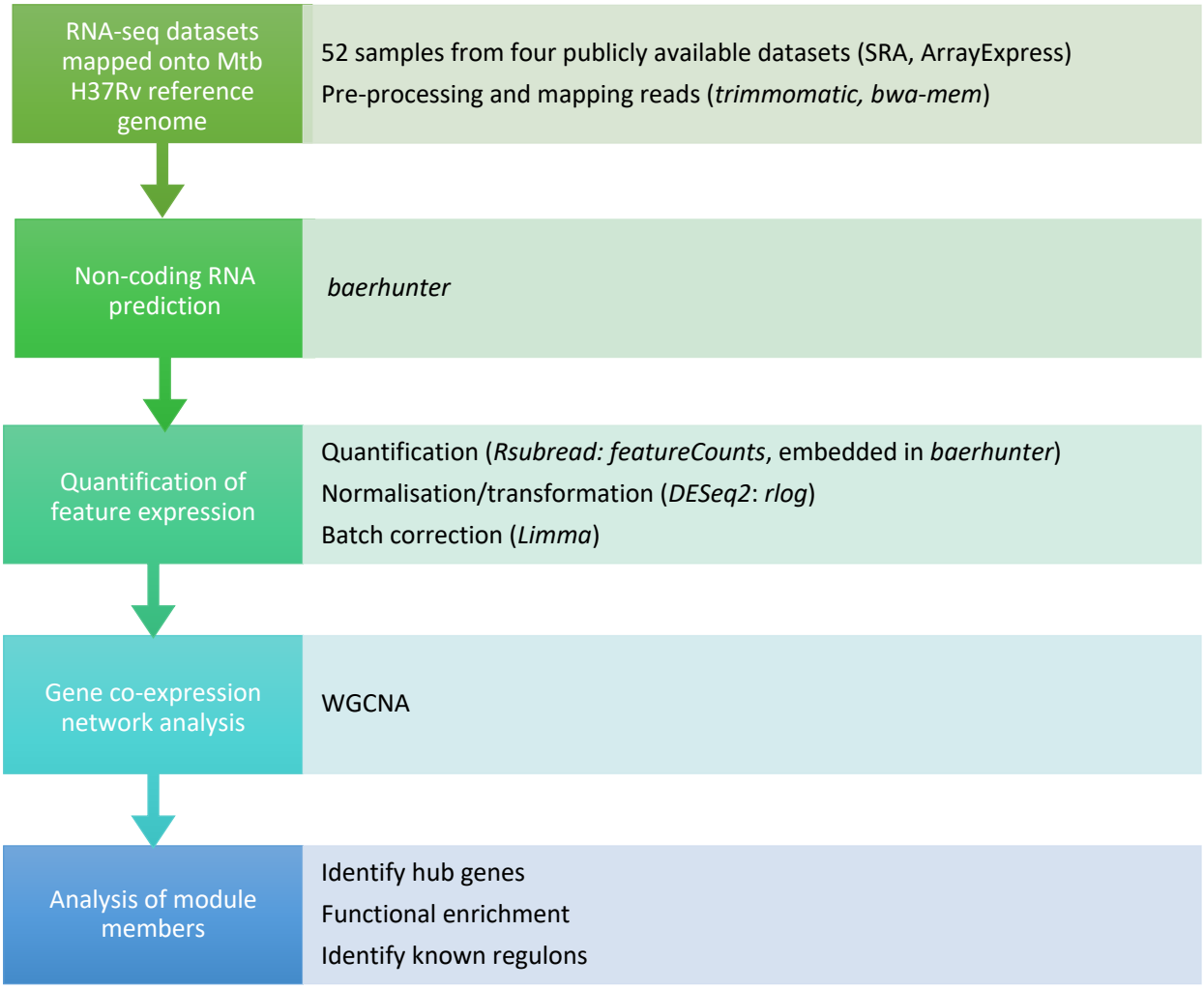
154 **Using WGCNA to implicate functional associations of non-coding RNA**
155
156 Weighted gene co-expression network analysis (WGCNA) (B. Zhang & Horvath, 2005) has
157 been widely used to identify functional groups of genes, called 'modules', through the
158 application of hierarchical clustering to differential expression levels of RNA transcripts
159 in microarray or RNA-seq experiments. Recent studies have focussed entirely on the
160 protein-coding portion of the transcriptome, using WGCNA with RNA-seq to cluster the
161 differentially expressed genes of *Mycobacterium marinum* in response to resuscitation
162 after hypoxia (Jiang et al., 2020) and *Mycobacterium aurum* infected macrophages (Lu et

163 al., 2021). Mtb microarray data have been used to cluster protein-coding genes that show
164 differential expression among clinical isoloates (Puniya et al., 2013) and in response to
165 two different hypoxic models to identify potential transcription factors (Jiang et al., 2016).
166 Another recent network analysis, using a matrix deconvolution method followed by
167 module clustering, uses a large number of RNA-seq samples including deletion mutants,
168 infection models and antibiotic-treated samples as well as restricted media and culture
169 conditions (Yoo, et al., 2022). Here the authors identify 80 modules of protein-coding genes
170 that each approximate an isolated source of variance, together estimated to account for
171 61% of the total variance seen in in the dataset. This proportion is reportedly lower than
172 results from similar analyses in other organisms, potentially due to the bias in the types
173 of conditions available in the database and/or the complex nature of regulation in Mtb
174 (Yoo, et al., 2022). However, the contribution of regulatory ncRNA elements may be a
175 considerable unexplored source of variance in this complex system. Here we use an
176 alternative, complementary approach by including ncRNA, as well as annotated protein-
177 coding genes, in the modules.
178
179 In this study, WGCNA was applied to multiple Mtb H37Rv datasets covering 15 different
180 culture conditions replicating various growth conditions, nutrient sources and stressors
181 encountered in the host environment. We present a global view of the non-coding genome
182 across an extensive WGCNA network and interrogate selected modules to identify
183 functional groupings between protein-coding and non-coding transcripts, as well as
184 between well-characterised genes and those with little functional annotation. The
185 correlation of the modules with the various conditions can identify participants in large-
186 scale transcriptomic remodelling programs in response to changes in environmental
187 conditions.
188

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://doi.org/10.5281/zenodo.7319853**.

Figure 1. Analysis workflow



Data Acquisition and Mapping

Datasets were downloaded from SRA (https://www.ncbi.nlm.nih.gov/sra/docs/) or Array Express (https://www.ebi.ac.uk/arrayexpress/) using the accession numbers listed in Table 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

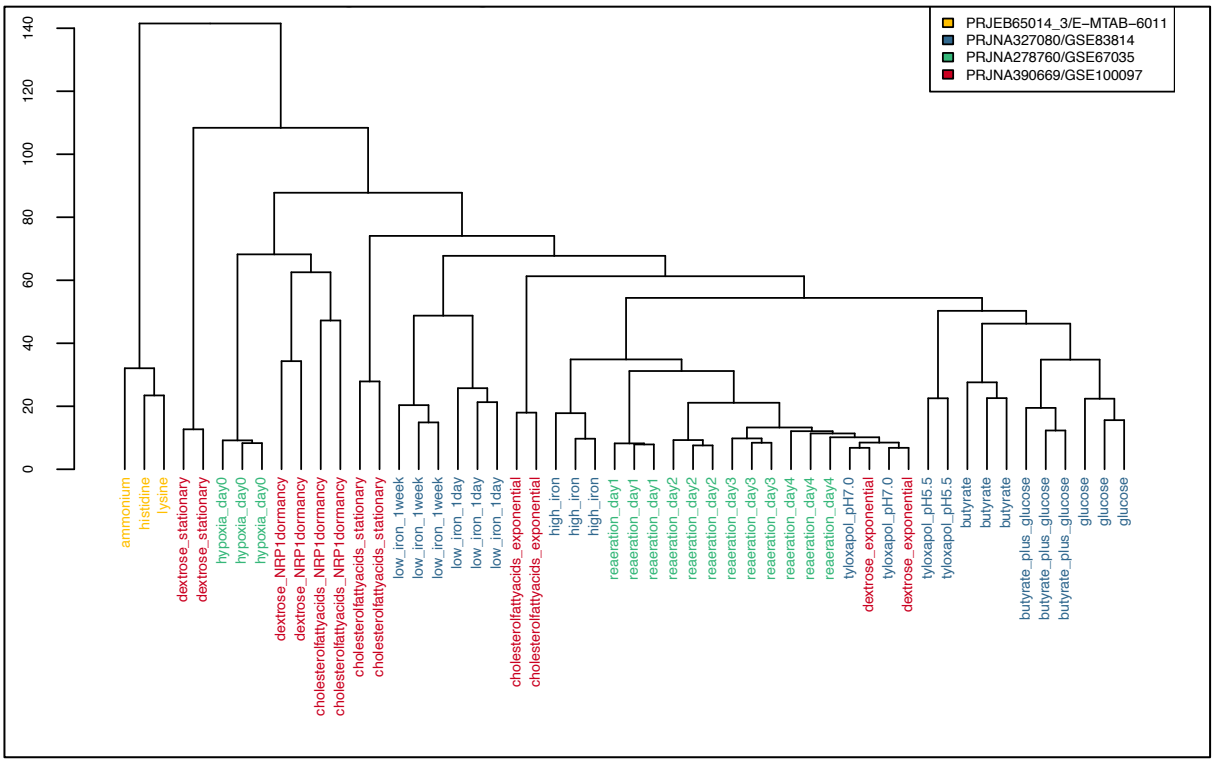
Non-coding RNA prediction

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://doi.org/10.5281/zenodo.7319853**). '*Count_features*' and '*tpm_norm_flagging*' functions were used for transcript quantification and to identify low

225 expression hits (less than or equal to 10 transcripts per million) in each dataset, which
226 were subsequently eliminated. When viewed on a genome browser, coverage at the 3' ends
227 of putative sRNA and UTRs often appears to decrease gradually, with the actual end of
228 the transcript appearing indistinct, compared to the 5' end. Prokaryotic ncRNA
229 transcripts may not demonstrate a clear fall-off of expression signal in RNA-seq due to
230 incomplete RNAP processivity and pervasive transcription regulated by the changing
231 levels of Rho protein observed in different conditions (Bidnenko & Bidnenko, 2018; Wade
232 & Grainger, 2014). These very long predictions can mask predicted transcripts in the same
233 region from other samples, obscuring potentially interesting shorter transcripts expressed
234 in different conditions. For this reason, transcripts longer than 1000 nucleotides were
235 eliminated before combining the predictions between datasets. The predicted annotations
236 for each dataset were combined into a single annotation file, adding the union of the
237 predicted boundaries to the reference genome for H37Rv (AL123456.3). Predictions that
238 overlapped with annotated ncRNAs and UTR predictions that overlapped sRNA
239 predictions from a different dataset were eliminated. Transcript quantification was
240 repeated on each dataset using the resulting combined annotation file and the count data
241 from each dataset was merged into a single counts matrix.
242
243 *DESeq2* v1.30.1 (Love et al., 2014) was used on the complete counts matrix including the
244 filtered *baerhunter* predictions to calculate size factors, estimate dispersion and
245 normalise the data with the regularised log transformation function (Supp figures, S1 and
246 S2). The normalised data was checked for potential batch effects using PCA plots and
247 hierarchical dendrograms. *Limma* v3.46.0 (Ritchie et al., 2015) *'removeBatchEffect'* was
248 applied with a single batch argument to remove batch effects associated with the first
249 component (batching the data according to dataset due to technical differences) while
250 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 = "pearson", networkType = "signed", power = 12, TOMType = "signed", minModuleSize = 25, reassignThreshold = 0, mergeCutHeight = 0.15, deepSplit = 2. In this type of network, the 'nodes' are the genes, and the 'edges', or

269 links, are created when gene expression patterns correlate. In contrast to unweighted
270 binary networks where links are assigned 0 or 1 to indicate whether or not the genes are
271 linked, in a weighted network the links are given a numeric weight based on how closely
272 correlated the expression is. *WGCNA* first calculates the signed co-expression similarity
273 for each gene pair. The absolute value of this correlation is raised to a power (determined
274 by the user, based on a scale-free topology model that mimics biological systems (Supp
275 figure S4) in order to weight the strong connections more highly than the weaker
276 connections. The resulting similarity matrix is used to cluster groups of genes with strong
277 connections to each other in a non-supervised manner (i.e., it doesn't use any previous
278 information about gene groups or connected regulons). A cluster dendrogram is created
279 (Supp figure, S8) and closely connected branches of the dendrogram are merged into
280 modules based on a cut-off value (also a parameter controlled by the user). The modules
281 are defined by a 'module eigengene' (ME), which explains most of the variance in the
282 expression values in the module. The connectivity of the MEs define the shape of the
283 overall network (Supp figure, S9). The modules can then be tested for potential
284 correlations with experimental conditions while reducing the degree of penalties for
285 multiple testing. In signed networks, correlation of the module with a condition can be in
286 either the positive or negative direction, as modules include transcripts that are similar
287 in both the degree and direction of correlation, allowing for a more fine-grained analysis
288 than with unsigned networks (Supp figure, S10).
289
290 To test correlations of modules with experimental conditions, the individual RNA-seq
291 samples were assigned to a condition based on the experimental description in the project
292 metadata. Some of these conditions were shared among the different projects, so when
293 appropriate, samples from different datasets were assigned the same condition, resulting
294 in 15 tested conditions. For example, late-stage reaeration samples were tested along with

295 exponential growth samples, and samples that tested hypoxia and cholesterol utilisation
296 together were included in multiple conditions. Models of hypoxia differed between the
297 RNA-seq projects, and these samples were assigned to different conditions: 'hypoxia'
298 versus 'extended hypoxia' (Supp Table 1, 'Condition summary' tab). All correlations were
299 made using robust biweight midcorrelation tests and all p-values were corrected for
300 multiple testing with the Benjamini-Hochberg (BH) method (Benjamini & Hochberg,
301 1995). Significance was evaluated as an adjusted p-value (p_{adj}) of < 0.05.
302

303 **Module Enrichment**

304 Modules were interrogated for enrichment for Gene Onotology (GO) terms (Ashburner et
305 al., 2000; The Gene Ontology Consortium, 2021), Clusters of Orthologous Groups (COG)
306 (Galperin et al., 2021), KEGG pathway genes (Kanehisa et al., 2022), functional categories
307 and literature searches for known regulons. GO terms, COG term and KEGG pathway
308 enrichment were accessed programmatically using the DAVID web service (Huang et al.,
309 2009b, 2009a; Jiao et al., 2012) to query the list of protein-coding genes from each module
310 for enrichment. Enrichment was determined using a modified one-sided Fisher's Exact
311 Test ('EASE' score) with BH correction for multiple testing, with $p_{adj} < 0.01$ considered
312 significantly enriched for a particular term, pathway or COG term. Enrichment for the 11
313 functional categories from Mycobrowser annotation (Kapopoulou et al., 2011) was
314 determined using a one-sided Fisher's Exact Test with BH correction for multiple testing.
315 Modules were enriched for a particular functional category if $p_{adj} < 0.01$. Lists of genes
316 associated with known regulons were mined from literature and enrichment was tested
317 using the same one-sided Fisher's Exact Test as above with a $p_{adj} < 0.01$ cut-off for
318 enrichment.
319

320 Non-coding RNA prediction, network analysis and subsequent data manipulation was
321 performed with R (v4.0.5, 2021-03-31). All plots were made in R with the following
322 packages: *WGCNA* (v1.69), *dendextend* (v1.15.2), *ggplot2* (v3.3.5). Scripts and expression
323 data are available at **https://doi.org/10.5281/zenodo.7319853.**
324
325

326 # RESULTS AND DISCUSSION

327

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

330 *Mycobacterium tuberculosis* RNA-seq datasets were selected from publicly available data
331 to find experiments using the wild-type H37Rv strain and representing a range of growth
332 conditions the pathogen may encounter in a host environment. Four datasets passing our
333 quality standards were subjected to our analysis pipeline (see Material and Methods) and
334 included 52 samples under 15 different experimental conditions (Supp Table 1, 'Samples'
335 tab). The R package, *baerhunter* (Ozuna et al., 2019), was used to predict ncRNA in
336 intergenic regions, antisense RNA (opposite a protein-coding gene) and UTRs at both the
337 5' and 3' ends of genes by searching the mapped RNA-seq data for expression peaks
338 outside of the annotated regions in the reference sequence for H37Rv. Non-coding RNA
339 predictions from each dataset were filtered for low expression and combined to create a
340 single set of non-overlapping annotations that encompassed all predictions made from any
341 sample under any experimental condition. In total, 1283 putative sRNAs were predicted
342 (including both truly intergenic transcripts as well as those antisense to a protein-coding
343 gene, or annotated RNA) and 1715 UTRs which includes all transcribed regions outside
344 of annotated protein-coding sequences at both 5' and 3' ends, as well as the non-coding
345 regions between adjacent genes in operons. All putative ncRNA transcripts (sRNAs and
346 UTRs) were searched for a TSS near the start of the predicted 5' boundary using
347 previously published annotations (Cortes et al., 2013; Shell et al., 2015). Annotated TSSs
348 were found within 20 nucleotides of the 5' end in 43% of the predicted sRNA transcripts.
349 Predicted 5' UTRs had a TSS within 10 nucleotides of the start in 42% of cases, compared
350 with 3% of the predicted 3' UTRs. Where the UTR covered the entire sequence between
351 two protein-coding regions (labelled as 'between' UTRs), 9% had a TSS in the first 10

352 nucleotides of the sequence (Table 2 and Supp Table 2 'putative_sRNAs', 'putative_UTRs'
353 tabs).
354
355 Table 2. Tally of predicted expressed elements in the *baerhunter*-generated combined annotation file. 4015
356 protein-coding genes were included in the annotation. 'Between' UTRs cover the entire sequence between
357 two protein-coding regions. *TSS predictions from (Cortes et al., 2013; Shell et al., 2015).
358

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

359
360
361 The predicted sRNAs were further annotated using the accepted nomenclature
362 (Lamichhane et al., 2013) which identifies the putative ncRNA relative to annotated gene
363 loci and differently signifies truly intergenic sRNAs and those that overlap any part of a
364 protein-coding region on the opposite strand. Most of the putative sRNAs are antisense to
365 the protein-coding region of one or more genes, but 91 putative sRNAs have predicted
366 boundaries that do not overlap an annotated transcript on either strand (or overlap an
367 annotated transcript on the opposite strand by fewer than 10 nucleotides). This number
368 is most probably an underestimate of the truly 'intergenic' sRNAs in the genome, as many
369 of the sRNA predictions appear over-estimated at the 3' end, effectively classifying them
370 as an antisense RNA even though the 5' half of the transcript does not overlap any genes
371 on the opposite strand. Isoforms of annotated sRNAs can be subject to post-transcriptional
372 processing to create an active transcript (Moores et al., 2017) and post-transcriptional
373 processing of 3' ends *in vivo* is more likely the norm for most prokaryotic transcripts

374 (Wang et al., 2019). However, for our purposes, any RNA-seq transcripts that extend to
375 overlap a protein-coding gene on the other strand in any dataset will be labelled as
376 antisense RNA.
377
378 The generated combined annotation file was used to quantify the expression of all 7043
379 expressed elements, including every annotated CDS, annotated ncRNA and predicted
380 ncRNA, in each sample. Raw counts of expression varied greatly among the datasets due
381 to different sequencing depth, as well as between some samples within datasets (as would
382 be expected with different environmental conditions), and only three protein-coding genes
383 showed no expression in any sample. The raw expression counts were transformed using
384 DESeq2's rlog function (Love et al., 2014), and plots of the dispersion of count data show
385 that the median expression level between samples and between datasets has been
386 normalised (Supp figures S1, S2). The distribution of the normalised expression levels of
387 protein-coding regions alone shows consistent median expression levels across the entire
388 dataset, however distribution of the normalised data restricted to putative sRNAs shows
389 more variability, with certain conditions showing increased or decreased expression of
390 these transcripts (Supp figures S5-S7). This is not unexpected, given that several studies
391 have identified pervasive transcription in hypoxic infection models, stationary phase and
392 dormancy. This is accompanied by a concomitant increase in non-rRNA abundance
393 (especially antisense RNA transcripts) and in the number of predicted TSSs in Mtb and
394 *M. smegmatis* (a fast-growing, non-pathogenic strain) (Arnvig et al., 2011; Ignatov et al.,
395 2015; Martini et al., 2019).
396
397
398 **Module networks represent groups of co-expressed genes and predicted non-**
399 **coding RNA**

400

401 ***Creation of the WGCNA network***

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

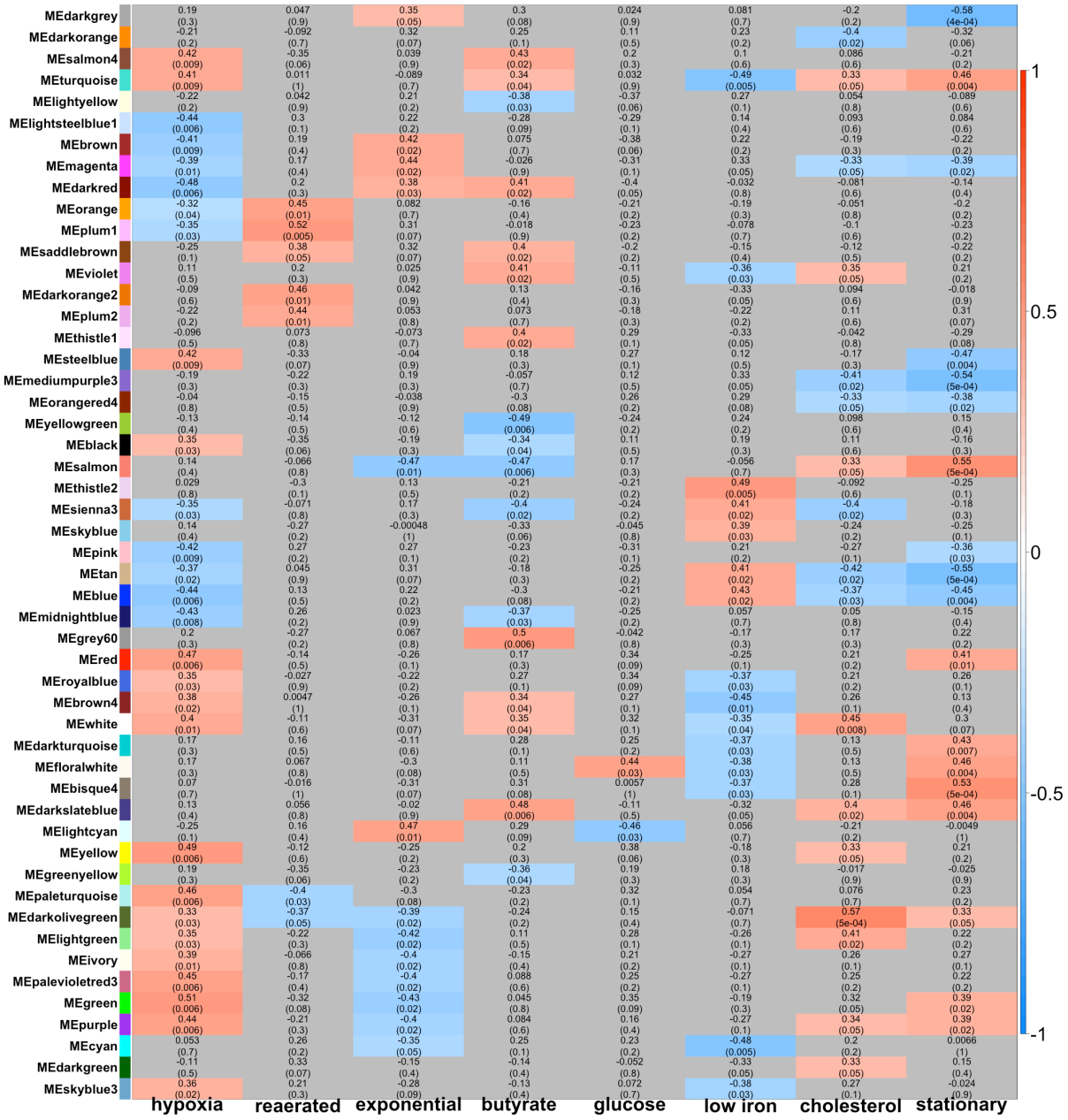
411

412 The signed co-expression network presented in this paper consists of 54 different modules,
413 assigning 99.3% of the expressed elements (CDS, putative UTRs and putative sRNAs)
414 into 53 modules, with 46 unassigned elements clustered in the 'grey' module (Supp Table
415 2, 'Module_Overview' tab). Module size ranged from 766 to 25 expressed elements. The
416 modules (using the ME) were tested for correlations with the various conditions used in
417 the RNA-seq experiments (see Materials and Methods). The RNA-seq data was
418 categorised into 15 different experimental conditions in total with varying numbers of
419 replicates (Supp Table S1, 'Condition Summary' tab), therefore, a statistically significant
420 correlation of modules with every condition was not expected. However, some modules do
421 show significant correlations with conditions such as iron restriction, cholesterol media,
422 hypoxia and growth phase and this can be informative when considering the association
423 of the gene groups with biological processes (Figure 3).

424

425

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

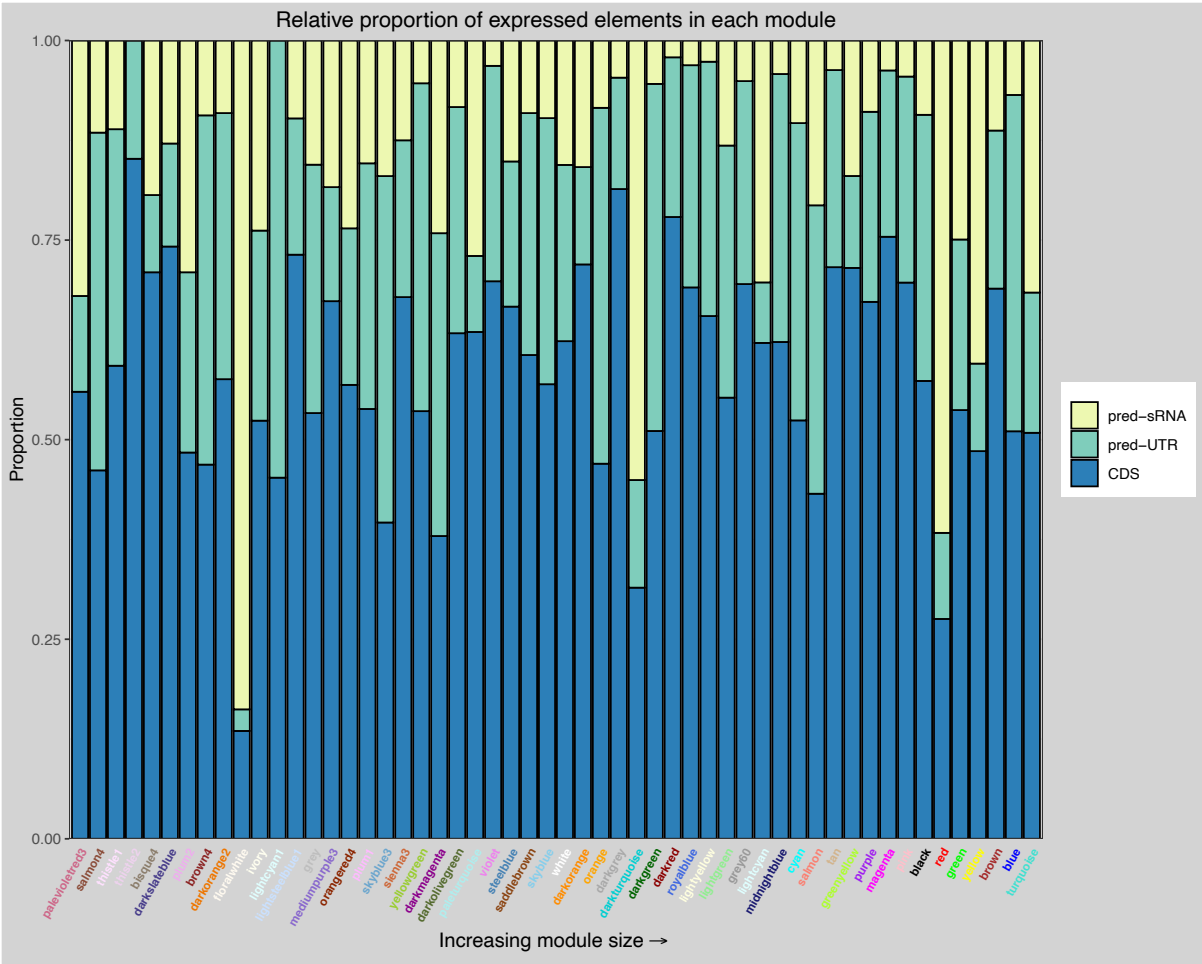


Well-established regulons cluster together in single modules

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

436 and stress responses (Du et al., 2016; Rustad et al., 2008; Voskuil et al., 2004). 47 of 48
437 previously identified DosR-regulated genes are found in a single module, '*cyan*',
438 representing statistically significant enrichment of DosR-regulated genes in the module
439 (one-sided Fisher's exact test, p_{adj}=3.81e-53). The '*cyan'* module also includes 5 genes from
440 the PhoP regulon which is associated with hypoxic response and coordination with the
441 DosR regulon (Gonzalo-Asensio et al., 2008; Singh et al., 2020) and the DosR-regulated
442 ncRNA, DrrS/MTS1338, known to be upregulated in hypoxic conditions (Ignatov et al.,
443 2015; Moores et al., 2017) . Unsurprisingly, the *'cyan'* module is enriched for the GO term,
444 'response to hypoxia', however, a statistically significant correlation was not seen with the
445 hypoxia condition (though it is negatively correlated with the exponential growth
446 condition, bicor=-0.35, p_{adj}=0.05) (Figure 3). The KstR regulon includes 74 genes under
447 control of the TetR-type transcriptional repressor, KstR, known to be involved in lipid
448 catabolism and upregulated during infection (Kendall et al., 2007, 2010; Nesbitt et al.,
449 2010). The '*royalblue*' module is significantly enriched for known KstR-regulated genes
450 (one-sided Fisher's exact test, p_{adj} = 5.06e-30) with 30 of 72 KstR-regulated genes
451 clustering together in the module. This module is enriched for genes of the KEGG pathway
452 for steroid degradation (p_{adj}= 3.32e-10) and the GO term 'steroid metabolic process' (p_{adj} =
453 5.62e-16). The module shows statistically significant positive correlation for hypoxia
454 (bicor=0.35, p_{adj}=0.03) and negative correlation with the low iron condition (bicor=-0.37,
455 p_{adj}=0.03) (Figure 3). Genes involved in mycobactin synthesis are nearly all found in the
456 *'grey60'* module (one-sided Fisher's Exact test, p_{adj}= 1.23e-17), a module enriched for the
457 KEGG pathways 'siderophore metabolic processes' and 'arginine biosynthesis'. As these
458 examples show, known associated genes are co-located in modules which represent a
459 functional group of genes that have co-regulated expression under various experimental
460 conditions. The modules can be further explored to identify novel associations.
461

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

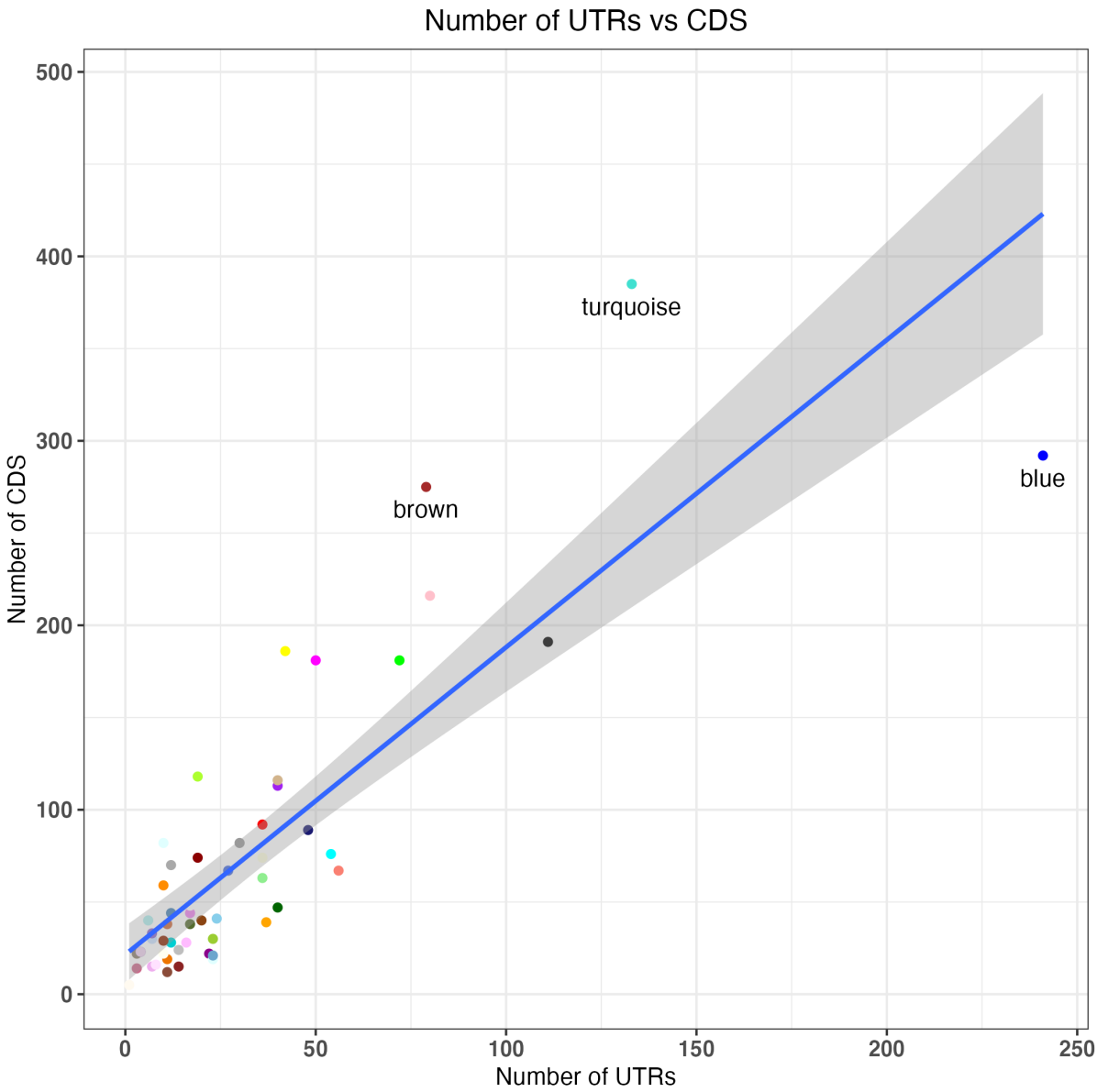


Predicted non-coding RNAs are enriched in certain modules

Putative sRNAs and/or predicted UTRs were distributed throughout all modules in the network (Figure 4, Supp Table 2, 'Module_Overview' tab). The number of predicted sRNAs were statistically enriched in seven modules and predicted UTRs enriched in another seven modules (one-sided Fisher's exact test, $p_{adj} < 0.01$, 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

477 protein-coding genes are predicted as a single UTR. As expected, the two modules that
478 include the highest number of predicted operons (from OperonDB, Chetal & Janga, 2015),
479 *'turquoise'* and *'brown'*, have a lower relative proportion of UTRs; however, the *'blue'*
480 module, which includes 15 complete predicted operons, is significantly enriched for UTRs
481 (p_{adj} = 6.79e-21) (Figure 5).

482 Figure 5. Plot of number of UTRs against number of CDS in each module. Grey shading indicates confidence
483 interval of 0.95.
484



485
486
487 Within the module sub-networks, the tight co-expression of protein-coding genes and
488 ncRNA is reflected by the number of ncRNA found among the most connected elements
489 in the module. The 'hub' elements are those with the best correlation to the ME and

490 therefore the most tightly connected elements in the individual module networks. In 14
491 modules, ncRNA (both predicted and annotated) make up more than half of the elements
492 with module membership values (MM) > = 0.80 (our threshold for identifying hub
493 elements) (Supp Table 2, 'Hub_info' tab). These associations may implicate ncRNA as co-
494 conspirators in regulatory pathways implemented to adapt to conditions such as hypoxia,
495 cholesterol media and low iron. The 30 annotated ncRNAs in the Mtb reference genome
496 (AL123456.3) are spread over 20 modules, with 10 of them hubs of the module, and one
497 unassigned ('grey' module) (Supp Table 2, 'Annotated ncRNA' tab). For example,
498 Ms1/MTS2823, observed to be the most abundantly expressed ncRNA in expression
499 studies over various stress conditions (Arnvig et al., 2011; Arnvig & Young, 2012; Ignatov
500 et al., 2015; Šiková et al., 2019), is a hub element in a module that is positively correlated
501 with cholesterol-containing media conditions (*'darkgreen',* bicor=0.35, p_{adj}=0.04) (Figure
502 3). This module is significantly enriched for KEGG pathways, including: Pyruvate
503 metabolism (p_{adj} = 3.1e-3) and two-component systems (p_{adj} = 3.8e-3), and GO terms:
504 plasma membrane respiratory chain complex II and plasma membrane fumarate
505 reductase complex. Mcr7/ncRv2395A, found to be part of the PhoP regulon (Solans et al.,
506 2014), is a hub in the *'violet'* module enriched for lipid metabolism and PE/PPE functional
507 categories, correlated positively with growth in cholesterol (bicor= 0.35, p_{adj}= 0.04) and
508 butyrate (bicor= 0.41, p_{adj}= 0.02) and negatively correlated with low iron (bicor= -0.36,
509 p_{adj}= 0.03) (Figure 3). F6/ncRv10243/SfdS, a sRNA upregulated in starvation and mouse
510 infection models, is thought to be involved in regulating lipid metabolism and long-term
511 persistence (Houghton et al., 2021). This ncRNA is a hub in a module found to be enriched
512 in 'lipid metabolism' genes (*'saddlebrown'*) and found to be correlated positively with
513 reaerated culture (bicor= 0.38, p_{adj}= 0.04) and butyrate (bicor= 0.4, p_{adj}= 0.02) conditions
514 (Figure 3).
515

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

517 We were interested to see how many of the predicted UTRs were assigned the same
518 module as the adjacent ORF—indicating whether the ORF and its adjacent UTR were co-
519 regulated. Intuitively, the UTR of a protein-coding gene would be expected to be expressed
520 as a single transcript along with the ORF and show similar expression patterns. However,
521 both 5' and 3' UTRs can act independently of the attached ORF and RNA abundance in
522 RNA-seq experiments reflects both transcription activity and transcript stability. For
523 example, some 5' UTRs are known to contain regulatory elements, such as riboswitches,
524 that alter the transcription of the downstream ORF (Dar et al., 2016; Kipkorir et al., 2021;
525 Schwenk & Arnvig, 2018; Warner et al., 2007), whereas sRNAs cleaved from 3' UTRs have
526 been shown to regulate the stability of the remaining transcript--with different half-lives
527 as a result (Chao et al., 2012; Dar & Sorek, 2018; Menendez-Gil & Toledo-Arana, 2021).
528 Of the *baerhunter* -predicted UTRs labelled 5' and 3', the UTRs co-segregated with the
529 ORF they were closest to in fewer than half of cases (Table 3). We would expect correctly-
530 identified 5' UTRs to utilise a TSS (whether or not there is a known predicted TSS),
531 whereas it appears functional 3' UTRs are more likely to be cleaved from the longer mRNA
532 transcript (Dar & Sorek, 2018; Menendez-Gil & Toledo-Arana, 2021; Ponath et al., 2022).
533 Our data confirms this: transcripts classified as 5' UTRs are much more likely to have a
534 predicted TSS in the first 10 nucleotides than transcripts classified as 3' UTRs (42% vs
535 2.7%). Approximately 9% of the UTRs predicted to be between ORFs (labelled, 'Between'
536 UTRs) have predicted TSS (Table 3). The presence of a TSS in the first 10 nucleotides of
537 the predicted UTR appeared to have little bearing on whether or not the UTR and its
538 adjacent ORF are assigned to the same module, with 43% of 5' and 19% of 3' UTRs with a
539 predicted TSS co-assigned with their adjacent ORF partner. 42% of the 'Between' UTRs
540 do not segregate with either the ORF upstream or downstream, indicating their
541 expression is, to some degree, independent of either adjacent ORF. 195 UTRs were found

542 to be hubs in modules independent of their adjacent ORF(s), with 27 including a predicted
543 TSS. All 'independent' UTRs are found in Supplementary Table 2, 'independent_UTRs'
544 tab.
545
546 Table 3. UTRs and module assignment of adjacent ORFs. DS=downstream, US=upstream. TSS indicates
547 presence of annotated TSS in first 10 nucleotides of predicted UTR (Cortes et al., 2013; Shell et al., 2015).
548

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

549
550
551 ***Antisense RNAs are hubs in modules independent of cognate ORF***
552 It has been observed that the overall abundance of antisense RNA and other non-
553 ribosomal RNA increases upon exposure to stress such as hypoxia and nutrient restriction
554 (Arnvig et al., 2011; Ignatov et al., 2015), and in our network, ncRNA are well-connected
555 in various modules that include known transcription factors and gene regulons associated
556 with stress responses. Not unexpectedly, very few (5%) of the predicted antisense
557 transcripts were assigned to the same module as the protein-coding region overlapping on
558 the opposite strand (choosing the most downstream locus in the event of multiple
559 overlapping ORFs), signifying distinct patterns of expression for transcripts on opposite
560 strands, possibly due to independent or bi-directional promoters and/or overlapping
561 transcription termination sites. Bi-directional promoters have been identified in multiple
562 prokaryotic genomes, and competition for RNA polymerase (RNAP) binding among
563 divergently transcribed sense/antisense pairs may function as a mechanism for regulation
564 of gene expression (Ju et al., 2019; Warman et al., 2021). Long 3' UTRs that overlap with
565 converging protein-coding genes on the opposite strand (or with the 3' UTR) can create an
566 'excludon' regulatory arrangement, where transcription of the two opposite mRNAs is

567 simultaneously regulated by RNase targeting, or mutually exclusive due to RNAP
568 collision (Sáenz-Lahoya et al., 2019; Toledo-Arana & Lasa, 2020). Examining the module
569 groupings of the antisense RNAs and their base-pairing target on the other strand may
570 provide insight on which genes are regulated by antisense transcription.
571

572 ## Focus on Selected Module Networks

573 The large-scale transcription analysis presented here is useful for the more global
574 analysis of the overall trends related to ncRNA and transcription, but there is a great deal
575 of information to be gleaned by more fine-grained inspection of individual module
576 groupings. To discover novel associations in such a large and complex dataset, we have
577 selected a few modules for closer examination, focussing on those that contain gene groups
578 or regulons related to the tested conditions. Many of the modules that contain interesting
579 correlations or gene regulon enrichments also include an abundance of putative sRNAs
580 and UTRs. Using the 'guilt by association' principle, we can hypothesise that the well-
581 connected ncRNAs found among the module hub elements have a role in transcriptional
582 'remodelling' in response to changes in environmental conditions such as growth on
583 cholesterol-containing media, restricted iron or hypoxia.
584

585 ### *Detoxification-linked proteins cluster in the module best correlated with cholesterol*
586 ### *media condition*

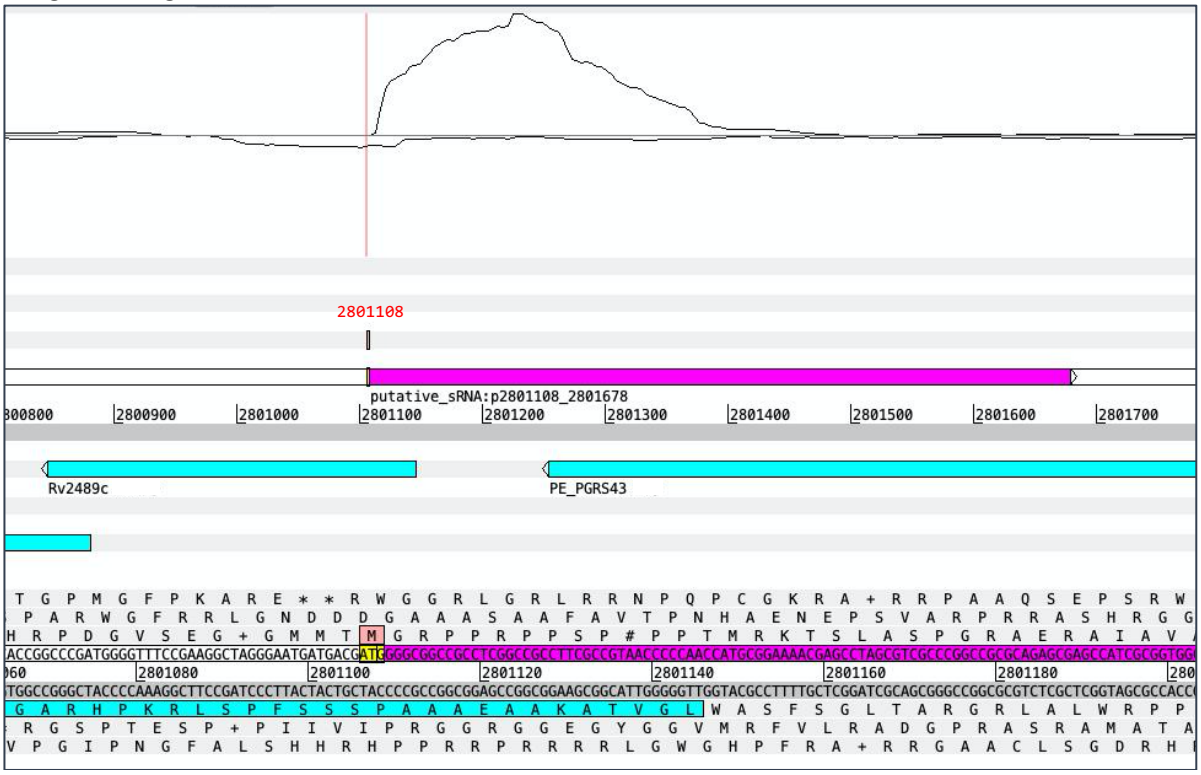
587 The *'darkolivegreen'* module showed positive correlation with the cholesterol media
588 condition (bicor=0.57, p_{adj}=5.0e-04) and negative correlation with low iron (bicor = -0.48,
589 p_{adj} = 0.001) (Figure 3). Many protein-coding genes involved in detoxification pathways
590 are hubs in the module, including several encoding transmembrane proteins such as the
591 *mmpL5-mmpS5* efflux pump operon (Rv0676c-Rv0677c), as well as the next gene
592 downstream, Rv0678, which was identified as part of a 'core lipid response' in differential

593 expression analysis in lipid-rich media (Aguilar-Ayala et al., 2017). The 5' UTR for
594 Rv0677c and 3' UTRs for Rv0676c and Rv0678 are also hubs. This operon is involved in
595 siderophore transport and expressed in cholesterol and lipid-rich environments (Aguilar-
596 Ayala, et al., 2017; Pawełczyk et al., 2021). The module contains several Type II toxin-
597 antitoxin systems including VapBC12 (Rv1720c1721c), VapBC41 (Rv2601A-2602), RBE2
598 (relFG, Rv2865-2866) and vapB36 and vapB40 which may have roles in adaptation to
599 cholesterol and the evolution of persisters (Ramage et al., 2009; Sala et al., 2014).
600 VapBC12, specifically, has been shown to inhibit translation and promote persister
601 phenotypes in response to cholesterol (Talwar et al., 2020). Other detoxification-linked
602 genes in the module, such as the ABC-family transporter efflux system, Rv1216c-1219c,
603 have also been implicated in transcriptomic remodelling in response to cholesterol
604 (Aguilar-Ayala et al., 2017; Pawełczyk et al., 2021).
605
606 Two adjacent predictions, the 3' UTR for Rv1772 (putative_UTR:p2006948_2007063)
607 followed by ncRv1773/putative_sRNA:p2007213_2007377, are hubs in the
608 *'darkolivegreen'* module. Together, they extend to overlap the antisense strand of a large
609 portion of Rv1773c, a probable transcriptional regulator in the IclR-family, found in a
610 different module *('turquoise')*. The 3' UTR for Rv1772 has been previously identified as an
611 abundant antisense transcript during exponential growth (Arnvig et al., 2011). The start
612 of the predicted sRNA transcript has no known TSS and could instead be an extension of
613 the predicted 3' UTR (Supp figure S11). (When combining predicted annotations from
614 different datasets, long predicted UTRs that overlapped shorter sRNA predictions were
615 discarded, see Methods). In *E.coli*, the IclR-family transcriptional regulators demonstrate
616 both activating and repressing activities on targets such as multidrug efflux pumps and
617 the *aceBAK* operon which regulates the glyoxylate shunt (Zhou et al., 2012). *Icl2a*
618 (Rv1915) is one of the Mtb isoforms of the isocitrate/methylocitrate lyase gene, *aceA*, and

619 may be regulated by Rv1773c, as seen in *E.coli*. Icl2a, Rv1772, its predicted UTR and the
620 antisense RNA (ncRv1773) are all hubs in the *'darkolivegreen'* module. *Icl2a* has been
621 observed to be upregulated with cholesterol as the sole carbon source and likely has a
622 second function as part of the methylcitrate cycle to convert the fatty acid metabolites
623 propionate and propionyl CoA to less toxic compounds (Bhusal et al., 2017; Pawełczyk et
624 al., 2021).
625
626 ***Module correlated with reaeration after non-replicating persistence includes genes for***
627 ***amino-acid synthesis and cell wall remodelling***
628 The module, *'saddlebrown'* is enriched for GO-terms for various amino-acid metabolic
629 processes and COG 'lipid metabolism'. It is positively correlated with reaeration after non-
630 replicating persistence (bicor= 0.38, p_{adj}= 0.04) and butyrate-containing media (bicor= 0.4,
631 p_{adj}= 0.02) (Figure 3). This pairing of upregulation of amino-acid synthesis and
632 upregulation of the synthesis of cell wall lipids has been observed in the 'lag phase' after
633 reaeration for increased protein synthesis (Du et al., 2016). The hubs of the *'saddlebrown'*
634 module include several predicted sRNAs, and the annotated sRNA, F6.
635 F6/ncRv10243/SfdS is a sigF-dependent ncRNA which has been shown to be induced in
636 nutrient starvation, oxidative stress, acid stress (Arnvig & Young, 2009; Houghton et al.,
637 2021) and the fatty acid hypoxia model (Del Portillo et al., 2019). In addition to being
638 expressed from its own promoter, F6/SfdS has been proposed to be co-transcribed with the
639 upstream gene *fadA2* (Rv0243), a probable acetyl-CoA acyltransferase; however, *fadA2* is
640 clustered in a different module from SfdS ('*darkred')*.
641
642 One of the predicted sRNAs among the '*saddlebrown'* module hubs is antisense transcript
643 ncRv2489/putative_srna:p2801108_2801678 with a TSS at 2801108. This overlaps the 3'
644 end of PE-PGRS43 (Rv2490c) (Figure 6). There is a short reading frame (30 nucleotides,

10 amino acids) initiating from a Methionine at this TSS that suggests a possible dual-function sRNA or sORF with independent function. A shorter, possibly-leadered, sORF was predicted by Shell et al. (2015) that falls within this region (2801238..2801261). The TSS for the predicted sRNA overlaps the 5' end of Rv2489c, a short, hypothetical 'alanine-rich protein'. The TSSs for these convergently overlapping transcripts are 42 nts apart

Figure 6. Antisense sRNA, ncRv2489/putative_srna:p2801108_2801678, (magenta bar) overlaps two transcripts and may encode a short peptide. TSS for sRNA indicated in red and corresponding amino acid highlighted in pink. Sample SRR5689230 from PRJNA390669, exponential growth on cholesterol and fatty acid media. Strand coverage using the 'second' read of each pair mapping to the transcript strand, visualised using Artemis genome browser (Carver et al., 2012).

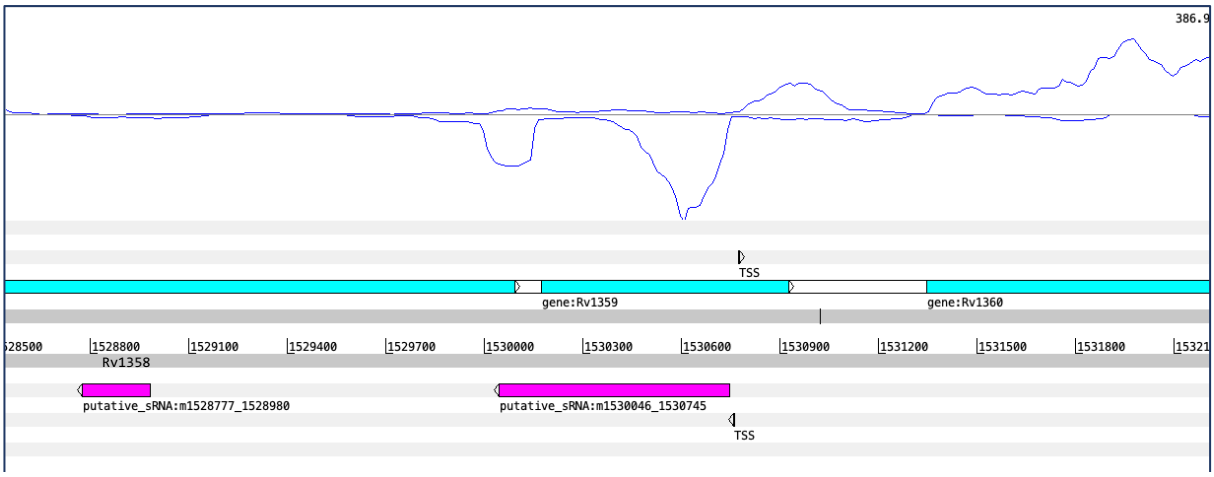


and may involve RNAP collision if both are transcribed simultaneously. Therefore, transcription of the predicted sRNA could impact either Rv2489c and/or PE-PGRS43 expression through two different mechanisms. Another hub sRNA in *'saddlebrown'* includes ncRv1450/putative_sRNA:p1630466_1631246, which has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

663 different modules. The 3' end of the prediction includes possible run-on transcription
664 antisense to the 3' end of PE-PGRS27.
665
666 The fatty-acid desaturase gene, Rv3229c (*desA3)* is a hub in the module, but without its
667 operon partner, Rv3230c. However, the module does contain an antisense sRNA in this
668 region, ncRv3230/putative_sRNA:p3607084_3607499 which is antisense to the 3' end of
669 Rv3230c, but lacks a known TSS. Interestingly, Rv3230c has an internal transcription
670 termination site predicted at 3607550 which coincides with the 3' end of the antisense
671 sRNA (D'Halluin et al., 2022) (Supp figure S12). Another hub antisense sRNA,
672 putative_sRNA:p3608313_3608866/ncRv3231c, overlaps the 3' end of the upstream gene,
673 Rv3231c, and has a predicted TSS at 3608313.
674
675 **Slow-growth correlated module is associated with transcriptional remodelling and metal**
676 **ion homeostasis and enriched for sRNAs**
677 The *'green'* module contains genes that are associated with transcriptional remodelling in
678 response to hypoxic or stationary growth conditions. It is positively correlated with
679 hypoxic (bicor=0.49, p_{adj}=0.004) and stationary (bicor=0.4, p_{adj}=0.01) growth conditions,
680 negatively correlated with exponential growth (bicor=-0.44 , p_{adj}=0.01) (Figure 3) and is
681 enriched for GO terms related to response to metal ions as well as regulation of gene
682 expression. The *'green'* module contains at least 30 known transcription factors, with 14
683 of them hubs in the module, including FurA, Zur and sigma factor, SigH, as well as being
684 enriched for SigH regulon genes. Three of the most well-connected transcription factors
685 (furA, smtB and zur) are involved in iron uptake and utilisation, and the Zur-regulated
686 ESAT-6 secretory proteins, esxR and esxS (Rv3019c, Rv3020c), are also present in the
687 module, linking metal homeostasis with response to hypoxia (Maciąg et al., 2007; Zhang
688 et al., 2020). Two chaperonin protein targets of the non-coding RNA F6/Sfds, GroES
689 (Rv3418c) and GroEL2 (Rv0440) are in the module, as well as the chaperonin protein, hsp

690 (Rv0251c), all of which are part of the phoPR virulence-regulating system (Gonzalo-
691 Asensio et al., 2008, 2014).
692
693 The *'green'* module is enriched for sRNAs (p_{adj}=0.011). Among the best-connected, are 27
694 predicted antisense RNAs. One of these hubs, putative_sRNA:p1404640_1404929/
695 ncRv1257 is antisense to the 3' end of Rv1257c, a probably oxidoreductase, and another
696 (putative_sRNA:p1771044_1771498/ ncRv1546) is antisense to the 5' end of a trehalose
697 synthetase, *treX.* Both of these sRNAs have TSSs and are expressed differentially among
698 the tested conditions. Control of reactive oxygen species and synthesis of trehalose
699 intermediates are important for cells in to survive in hypoxic conditions (Eoh et al., 2017;
700 Harold et al., 2019) and antisense RNA may be involved in fine-tuning these responses.
701 Another antisense RNA, ncRv1358c (putative_sRNA:m1530046_1530745) has a TSS near
702 its start and is found antisense to Rv1359. Rv1359 and the upstream gene, Rv1358, on
703 the opposite strand are very similar to each other (43.7% identity in 197 aa overlap) and
704 to another gene elsewhere in the genome, Rv0891c (48.5% identity in 204 aa overlap)
705 (Kapopoulou et al., 2011). All three genes are possible LuxR family transcriptional
706 regulators which are thought to be involved in quorum-sensing adaptations and contain
707 a probable ATP/GTP binding site motif (Chen & Xie, 2011; Modlin et al., 2021) and are
708 found in different modules. Expression of this antisense sRNA appears to suppress the
709 expression of the transcript on the opposite strand to varying degrees in all conditions
710 (Figure 7). In the cholesterol and fatty acid media samples, expression of a shorter
711 transcript appears to begin inside the Rv1359 ORF, where the transcript is not overlapped
712 by the antisense transcript, possibly utilising an internal TSS at 1530774.
713
714 Figure 7. Expression of antisense transcript putative_sRNA:m1530046_1530745 (magenta bar) seems to
715 suppress the expression of most of Rv1359 and Rv1358 in cholesterol and fatty acid media. An internal TSS
716 exists inside the Rv1359 CDS at 1530774 near where expression begins. Note, prediction of an individual
717 sRNA is an aggregate of predictions under different conditions, so will not always match the expression of the
718 sRNA in any particular sample. Sample SRR5689230 from PRJNA27860. Strand coverage using the 'second'

read of each pair mapping to the transcript strand, visualised using Artemis genome browser (Carver et al., 2012).



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

The *'darkred'* module is negatively correlated with the hypoxia condition (bicor=-0.46, p_{adj}=0.005, Figure 3). This module contains most of the ESX-3 genes (Rv0282-Rv0292) related to siderophore-mediated iron (and zinc) uptake in Mtb (Serafini et al., 2013; L. Zhang et al., 2020), with nine of these representing hubs in the module. The module is enriched for the PE/PPE functional category, and includes the two genes preceding the ESX-3 genes, Rv0280 (PPE3) and Rv0281 (a possible S-adenosylmethionine-dependent methyltransferase involved in lipid metabolism, though its position in the genome would suggest regulation could be linked to ESX-3 (Lunge et al., 2020)), and an ESX-5 gene, Rv1797 (*eccE5)*. The module also contains another Zur-regulated gene, Rv0106, which is a potential zinc-ion transporter (Zondervan et al., 2018). Among the hubs of the module are several genes related to lipid metabolism and fatty acid synthesis, including: probable triglyceride transporter, Rv1410; the operon consisting of Rv0241c (*htdX)*, Rv0242c (*fabG4)*, and Rv0243 (*fadA2)* (Dutta, 2018); and a gene involved in the pentose phosphate pathway, *zwf2* (Rv1447c).

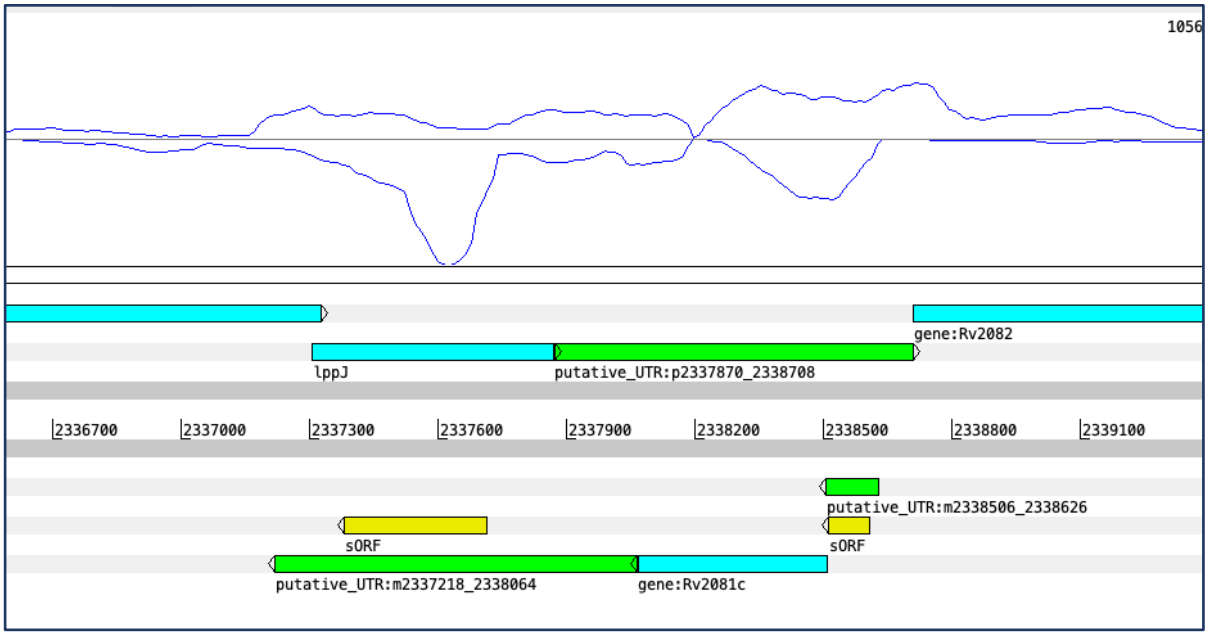
742 There are some well-connected ncRNAs in the *'darkred'* module, including a predicted
743 antisense RNA to Rv0281, 'ncRv0281c' (putative_sRNA:m341328_342075). This putative
744 sRNA has a predicted TSS at the 5' end and is transcribed divergently from Rv0282
745 *(eccA3)*. This is one of the rarer cases where the antisense transcript and cognate protein-
746 coding gene (Rv0281) are clustered in the same module. The prevailing direction of
747 transcription at this locus may be a result of competition for RNAP binding at a bi-
748 directional promoter in the predicted 5' UTR of Rv0282 which also clusters in the module.
749 There are several UTRs in the module hubs, including a 3' UTR for the gene Rv1133c,
750 *metE* (also found in the module). This UTR was previously identified as abundantly
751 expressed in exponential culture (Arnvig et al., 2011). There is a 3' UTR for Rv0292
752 *(eccE3,* also a hub in the *'darkred'* module) that is antisense to a large part of the 3' end
753 of Rv0293c which has a converging orientation to Rv0292 (Supp figure S13). Rv0293c is a
754 hub in a different module *('lightgreen')* along with its 3' UTR. Overlapping 3' ends of genes
755 could function to regulate transcription, possibly by bi-directional termination brought
756 about by RNAP collision, or function post-transcriptionally by influencing transcript
757 stability (Ju et al., 2019; Vargas-Blanco & Shell, 2020).
758
759 ***Module enriched for sRNAs and PE/PPE genes is correlated with stationary condition***
760 The *'darkturquoise'* module is enriched with sRNAs, with 33 hub sRNAs. It is negatively
761 correlated with the low iron condition (bicor = -0.37, p_{adj}=0.03) and positively correlated
762 with stationary growth (bicor= 0.43, p_{adj}=0.007). The genes of the module are enriched for
763 the PE/PPE functional category and there are several PE/PPE genes among the hubs.
764 The previously annotated ncRNA, B11 (also known as 6C or ncRv13660c), is one of the
765 most well-connected elements in the module and overexpression of B11 in *M.smegmatis*
766 has been shown to cause growth arrest and downregulation of a large set of genes
767 including those involved in cell division and virulence, including all the ESX-1 secretion

768 system genes (Mai et al., 2019). Mcr11 is also found in the module. This sRNA is known
769 to respond to the second messenger 3',5'-cyclic adenosine monophosphate and has been
770 found to be expressed in hypoxic Mtb cultures and in a mouse infection model (Girardin
771 & McDonough, 2020). Mcr11 regulates the expression of several genes that adapt central
772 carbon metabolism during slow growth conditions (Girardin & McDonough, 2020).
773
774 There are two well-connected intergenic sRNAs predicted in the '*darkturquoise'* module.
775 Putative_sRNA:p1164036_1164162 / ncRv11040 is located between PE8 and a possible
776 transposase, Rv1041c, but on the antisense strand. There is a predicted TSS at 1163697,
777 39 nucleotides upstream of the predicted start sequence. This transcript is in a converging
778 orientation to the transposase and may be instrumental in regulating horizontal gene
779 transfer (Ellis & Haniford, 2016; Lejars et al., 2019). The other intergenic hub is also
780 upstream from possible transposase, Rv3114, but in diverging orientation on the opposite
781 strand. The TSS is at 3481459, and the sRNA is within a predicted 'MT-complex-specific'
782 genomic island associated with virulence genes (Becq et al., 2007). Rv3112-14 are
783 clustered in the *'salmon'* module.
784
785 There are several interesting 'independent' UTRs that are well-connected in the module,
786 but their assumed transcriptional partner clusters in another module. There are several
787 predicted TSS's and transcriptional termination sites (TTS) (D'Halluin et al., 2022) within
788 the predicted boundaries of a 3' UTR for the gene Rv0281c
789 (putative_UTR:m2337218_2338064) and a predicted sORF based on ribosome profiling
790 (Smith et al., 2022) (Figure 8). This 3' UTR is antisense to Rv2080, *lppJ,* which is in the
791 *'cyan'* module along with most of the DosR regulated genes. The higher level of expression
792 of the antisense transcript relative to the accepted sense strand has invited speculation
793 that Rv2080 has been misannotated (K. B. Arnvig et al., 2011). The 5' UTR of Rv0281c is

794 also a hub in the module and contains predicted TSSs, TTS and sORF. It would be
795 interesting to discover whether these UTRs could have dual functions as regulatory RNA
796 elements as well as being translated into short peptides. Rv2081c is a conserved
797 membrane protein containing a simple sequence repeat of 8 C's and has been identified
798 as a source of sequence variation in Mtb sputum and culture (Shockey et al., 2019; Sreenu
799 et al., 2007).

800

801 Figure 8. 5' and 3' UTRs for Rv2081c (green bars) are overlapped by predicted sORFs (yellow bars). (Cortes
802 et al., 2013; Smith et al., 2022). Shown is sample SRR5689224, hypoxia condition from PRJNA27860. Strand
803 coverage using the 'second' read of each pair mapping to the transcript strand, visualised using Artemis
804 genome browser (Carver et al., 2012).
805



806

807

808 The best connected elements in the module are antisense sRNAs, including
809 putative_sRNA:p2081553_2082178/ ncRv1835, with a predicted TSS at its start. It is
810 antisense to Rv1835c, the gene for a putative serine esterase clustered in the
811 *'mediumpurple3'* module, in particular to the 3' end of the peptidase domain (Xaa-Pro
812 dipeptidyl-peptidase-like domain) (Blum et al., 2020).
813 Putative_sRNA:m2497549_2498369/ncRv2225c, with a TSS at 2498368, is antisense to

814 Rv2225, coding for a 3-methyl-2-oxobutanoate hydroxymethyltransferase PanB. This
815 gene clusters in the in *'turquoise'* module.
816

817 CONCLUSION

818 This paper presents a large-scale network analysis of over 7000 transcripts expressed by
819 Mtb under a variety of conditions. The modules group together clusters of co-expressed
820 protein-coding genes, as well as ncRNA transcripts predicted from RNA-Seq signals.
821 Several modules are statistically enriched for sRNAs, especially those modules positively
822 correlated with hypoxia. The abundance of antisense RNA in conditions of stress has been
823 widely observed, and it is therefore not a surprise to find them in the hubs of these
824 modules. However, it is noticeable that the complementary ORF is usually excluded,
825 which leads us to seriously consider antisense transcription as part of strategic regulation
826 of protein production in response to environmental cues through mechanisms of divergent
827 transcription, translational control or by regulating mRNA stability (Vargas-Blanco &
828 Shell, 2020; Warman et al., 2021). If these strategies actually differ among the members
829 of the MTBC, it may have implications for host specificity and virulence (Dinan, Adam M.
830 et al., 2014). By the same logic, 3' UTR transcripts clustering in modules distinct from
831 their upstream ORF implies independent function from the ORF. sRNAs generated from
832 3' UTRs have been reported in other prokaryotes and evidence points to widespread
833 mRNA processing that could release independent transcripts at the 3' end (Dar & Sorek,
834 2018; Desgranges et al., 2021; Updegrove et al., 2019; Wang et al., 2019). In compact
835 bacterial genomes, 3' UTRs are also found to overlap other 3' UTRs in a converging
836 transcription pattern which may provide a mechanism for regulating the expression or
837 stability of either transcript (Ju et al., 2019; Vargas-Blanco & Shell, 2020).
838

839 The gene modules presented here are somewhat 'blunt-force instruments' applied to
840 transcripts that are part of overlapping, coordinated responses to various environmental
841 cues, but restricted to a single module grouping. Recent work exploring differentially
842 expressed genes in response to various environmental conditions have revealed highly
843 integrated adaptation responses. In other words, a single environmental change, e.g.
844 hypoxia or growth on fatty acids or cholesterol, stimulates transcriptomic remodelling
845 across diverse cellular functions, perhaps acting as cues to stimulate anticipatory
846 pathways and ready the pathogen for the next challenge (Eoh et al., 2017; Gerrick et al.,
847 2018). Confounders such as dual-function, 'moonlighting', proteins may weaken the
848 correlation of a module with a specific condition and may create noise in otherwise well-
849 connected modules. However, focussing on the best connected transcripts in various
850 modules can uncover the unexpected connections between genes of 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 modules discussed in depth in this paper represent a limited snapshot of this
867 extensive co-expression network. Modules of interest can be identified by correlations to
868 experimental conditions, associated GO terms, functional categories, or gene group
869 enrichment. The supplementary tables (Supp Table 2) have been organised into an easily-
870 accessible spreadsheet for researchers to query particular genes or modules of interest
871 and find associated protein-coding genes or ncRNA. These spreadsheets provide
872 information about the module association, membership values, TSSs and for UTRs, the
873 module membership of the adjacent ORFs for each predicted ncRNA. We anticipate this
874 to be a useful resource for discovering ncRNA candidates for further investigation, add
875 context to the circumstances of expression of previously identified ncRNAs, identify
876 associations of genes with unknown functions and suggest roles for 'moonlighting'
877 proteins that may be associated with unexpected gene groupings.

878

879 **References**

880 Aguilar-Ayala, D. A., Tilleman, L., Van Nieuwerburgh, F., Deforce, D., Palomino, J. C.,
881 Vandamme, P., Gonzalez-Y-Merchand, J. A., & Martin, A. (2017). The transcriptome
882 of Mycobacterium tuberculosis in a lipid-rich dormancy model through RNAseq
883 analysis. *Scientific Reports*, *7*(1), 17665. https://doi.org/10.1038/s41598-017-17751-x
884 Ami, V. K. G., Balasubramanian, R., & Hegde, S. R. (2020). Genome-wide identification of
885 the context- dependent sRNA expression in Mycobacterium tuberculosis. *BMC*
886 *Genomics*, *21*(167), 1–12.
887 Arnvig, K. B., Comas, I., Thomson, N. R., Houghton, J., Boshoff, H. I., Croucher, N. J.,
888 Rose, G., Perkins, T. T., Parkhill, J., Dougan, G., & Young, D. B. (2011). Sequence-
889 Based Analysis Uncovers an Abundance of Non-Coding RNA in the Total
890 Transcriptome of Mycobacterium tuberculosis. *PLOS Pathogens*, *7*(11), e1002342.
891 https://doi.org/10.1371/journal.ppat.1002342
892 Arnvig, K. B., & Young, D. B. (2009). Identification of small RNAs in Mycobacterium
893 tuberculosis. *Molecular Microbiology*, *73*(3), 397–408.
894 https://doi.org/10.1111/j.1365-2958.2009.06777.x
895 Arnvig, K., & Young, D. (2012). Non-coding RNA and its potential role in Mycobacterium
896 tuberculosis pathogenesis. *RNA Biology*, *9*(4), 427–436.
897 https://doi.org/10.4161/rna.20105
898 Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P.,
899 Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L.,
900 Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G.
901 M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology.
902 *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

903 Becq, J., Gutierrez, M. C., Rosas-Magallanes, V., Rauzier, J., Gicquel, B., Neyrolles, O., &
904 Deschavanne, P. (2007). Contribution of horizontally acquired genomic islands to the
905 evolution of the tubercle bacilli. *Molecular Biology and Evolution*, *24*(8), 1861–1871.
906 https://doi.org/10.1093/molbev/msm111
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 Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka,
918 G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams,
919 L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., Marchler-Bauer, A., …
920 Finn, R. D. (2020). The InterPro protein families and domains database: 20 years on.
921 *Nucleic Acids Research*, *49*(D1), D344–D354. https://doi.org/10.1093/nar/gkaa977
922 Canestrari, J. G., Lasek-Nesselquist, E., Upadhyay, A., Rofaeil, M., Champion, M. M., Wade,
923 J. T., Derbyshire, K. M., & Gray, T. A. (2020). Polycysteine-encoding leaderless
924 short ORFs function as cysteine-responsive attenuators of operonic gene expression in
925 mycobacteria. *Molecular Microbiology*, *114*(1), 93–108.
926 https://doi.org/10.1111/mmi.14498

927 Chao, Y., Papenfort, K., Reinhardt, R., Sharma, C. M., & Vogel, J. (2012). An atlas of Hfq-
928 bound transcripts reveals 3' UTRs as a genomic reservoir of regulatory small RNAs.
929 *The EMBO Journal*, *31*(20), 4005–4019. https://doi.org/10.1038/emboj.2012.229
930 Chen, J., & Xie, J. (2011). Role and regulation of bacterial LuxR-like regulators. *Journal of*
931 *Cellular Biochemistry*, *112*(10), 2694–2702. https://doi.org/10.1002/jcb.23219
932 Chetal, K., & Janga, S. C. (2015). OperomeDB: A Database of Condition-Specific
933 Transcription Units in Prokaryotic Genomes. *BioMed Research International*, *2015*,
934 318217–318217. PubMed. https://doi.org/10.1155/2015/318217
935 Cortes, T., Schubert, O. T., Rose, G., Arnvig, K. B., Comas, I., Aebersold, R., & Young, D.
936 B. (2013). Genome-wide mapping of transcriptional start sites defines an extensive
937 leaderless transcriptome in Mycobacterium tuberculosis. *Cell Reports*, *5*(4), 1121–
938 1131. https://doi.org/10.1016/j.celrep.2013.10.031
939 Dar, D., Shamir, M., Mellin, J. R., Koutero, M., Stern-Ginossar, N., Cossart, P., & Sorek, R.
940 (2016). Term-seq reveals abundant ribo-regulation of antibiotics resistance in
941 bacteria. *Science*, *352*(6282), aad9822. https://doi.org/10.1126/science.aad9822
942 Dar, D., & Sorek, R. (2018). Bacterial noncoding RNAs excised from within protein-coding
943 transcripts. *MBio*, *9*(5). https://doi.org/10.1128/mBio.01730-18
944 Del Portillo, P., García-Morales, L., Menéndez, M. C., Anzola, J. M., Rodríguez, J. G.,
945 Helguera-Repetto, A. C., Ares, M. A., Prados-Rosales, R., Gonzalez-y-Merchand, J.
946 A., & García, M. J. (2019). Hypoxia Is Not a Main Stress When Mycobacterium
947 tuberculosis Is in a Dormancy-Like Long-Chain Fatty Acid Environment. *Frontiers in*
948 *Cellular and Infection Microbiology*, *8*, 449–449.
949 Desgranges, E., Barrientos, L., & Caldelari, I. (2021). The 3'UTR-derived sRNA RsaG
950 coordinates redox homeostasis and metabolism adaptation in response to glucose-6-

phosphate uptake in Staphylococcus aureus. *Molecular Microbiology*. https://doi.org/10.1111/MMI.14845

D'Halluin, A., Polgar, P., Kipkorir, T., Patel, Z., Cortes, T., & Arnvig, K. B. (2022). Term-seq reveals an abundance of conditional, Rho-dependent termination in Mycobacterium tuberculosis. *BioRxiv*, 2022.06.01.494293. https://doi.org/10.1101/2022.06.01.494293

Dinan, Adam M., Tong, Pin, Lohan, Amanda J., Conlon, Kevin M., Miranda-CasoLuengo Aleksandra A., Malone, Kerri M., Gordon, Stephen V., & Loftus, Brendan J. (2014). Relaxed Selection Drives a Noisy Noncoding Transcriptome in Members of the Mycobacterium tuberculosis Complex. *MBio*, *5*(4), e01169-14. https://doi.org/10.1128/mBio.01169-14

Du, P., Sohaskey, C. D., & Shi, L. (2016). Transcriptional and physiological changes during Mycobacterium tuberculosis reactivation from non-replicating persistence. *Frontiers in Microbiology*, *7*(AUG). https://doi.org/10.3389/fmicb.2016.01346

Dutta, D. (2018). Advance in Research on Mycobacterium tuberculosis FabG4 and Its Inhibitor. *Frontiers in Microbiology*, *9*. https://www.frontiersin.org/article/10.3389/fmicb.2018.01184

Ellis, M. J., & Haniford, D. B. (2016). Riboregulation of bacterial and archaeal transposition. *WIREs RNA*, *7*(3), 382–398. https://doi.org/10.1002/wrna.1341

Eoh, H., Wang, Z., Layre, E., Rath, P., Morris, R., Branch Moody, D., & Rhee, K. Y. (2017). Metabolic anticipation in Mycobacterium tuberculosis. *Nature Microbiology*, *2*(8), 17084. https://doi.org/10.1038/nmicrobiol.2017.84

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., & Koonin, E. V. (2021). COG database update: Focus on microbial diversity, model organisms, and

widespread pathogens. *Nucleic Acids Research*, *49*(D1), D274–D281. https://doi.org/10.1093/nar/gkaa1018

Gerrick, E. R., Barbier, T., Chase, M. R., Xu, R., François, J., Lin, V. H., Szucs, M. J., Rock, J. M., Ahmad, R., Tjaden, B., Livny, J., & Fortune, S. M. (2018). Small RNA profiling in mycobacterium tuberculosis identifies mrsi as necessary for an anticipatory iron sparing response. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(25), 6464–6469. https://doi.org/10.1073/pnas.1718003115

Girardin, R. C., & McDonough, K. A. (2020). Small RNA Mcr11 requires the transcription factor AbmR for stable expression and regulates genes involved in the central metabolism of Mycobacterium tuberculosis. *Molecular Microbiology*, *113*(2), 504–520. https://doi.org/10.1111/mmi.14436

Gonzalo-Asensio, J., Malaga, W., Pawlik, A., Astarie-Dequeker, C., Passemar, C., Moreau, F., Laval, F., Daffé, M., Martin, C., Brosch, R., & Guilhot, C. (2014). Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(31), 11491–11496. https://doi.org/10.1073/pnas.1406693111

Gonzalo-Asensio, J., Mostowy, S., Harders-Westerveen, J., Huygen, K., Hernández-Pando, R., Thole, J., Behr, M., Gicquel, B., & Martín, C. (2008). PhoP: a missing piece in the intricate puzzle of Mycobacterium tuberculosis virulence. *PloS One*, *3*(10), e3496–e3496. PubMed. https://doi.org/10.1371/journal.pone.0003496

Harold, L. K., Antoney, J., Ahmed, F. H., Hards, K., Carr, P. D., Rapson, T., Greening, C., Jackson, C. J., & Cook, G. M. (2019). FAD-sequestering proteins protect mycobacteria against hypoxic and oxidative stress. *Journal of Biological Chemistry*, *294*(8), 2903–5814. https://doi.org/10.1074/jbc.RA118.006237

1000 Houghton, Joanna, Rodgers, Angela, Rose, Graham, D'Halluin, Alexandre, Kipkorir, Terry,
1001 Barker, Declan, Waddell, Simon J., Arnvig, Kristine B., & Oglesby, Amanda G.
1002 (2021). The Mycobacterium tuberculosis sRNA F6 Modifies Expression of Essential
1003 Chaperonins, GroEL2 and GroES. *Microbiology Spectrum*, *9*(2), e01095-21.
1004 https://doi.org/10.1128/Spectrum.01095-21
1005 Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009a). Bioinformatics enrichment tools:
1006 Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids*
1007 *Research*, *37*(1), 1–13. https://doi.org/10.1093/nar/gkn923
1008 Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009b). Systematic and integrative
1009 analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*,
1010 *4*(1), 44–57. https://doi.org/10.1038/nprot.2008.211
1011 Ignatov, D. V., Salina, E. G., Fursov, M. V., Skvortsov, T. A., Azhikina, T. L., &
1012 Kaprelyants, A. S. (2015). Dormant non-culturable Mycobacterium tuberculosis
1013 retains stable low-abundant mRNA. *BMC Genomics*, *16*(1), 954.
1014 https://doi.org/10.1186/s12864-015-2197-6
1015 Jiang, J., Lin, C., Zhang, J., Wang, Y., Shen, L., Yang, K., Xiao, W., Li, Y., Zhang, L., &
1016 Liu, J. (2020). Transcriptome Changes of Mycobacterium marinum in the Process of
1017 Resuscitation From Hypoxia-Induced Dormancy. *Frontiers in Genetics*, *10*(February),
1018 1–13. https://doi.org/10.3389/fgene.2019.01359
1019 Jiang, J., Sun, X., Wu, W., Li, L., Wu, H., Zhang, L., Yu, G., & Li, Y. (2016). Construction
1020 and application of a co-expression network in Mycobacterium tuberculosis. *Scientific*
1021 *Reports*, *6*(March 2015), 1–18. https://doi.org/10.1038/srep28422
1022 Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., &
1023 Lempicki, R. A. (2012). DAVID-WS: a stateful web service to facilitate gene/protein

1024 list analysis. *Bioinformatics*, *28*(13), 1805–1806.
1025 https://doi.org/10.1093/bioinformatics/bts251
1026 Ju, X., Li, D., & Liu, S. (2019). Full-length RNA profiling reveals pervasive bidirectional
1027 transcription terminators in bacteria. *Nature Microbiology*, *4*(11), 1907–1918.
1028 https://doi.org/10.1038/s41564-019-0500-z
1029 Kanehisa, M., Sato, Y., & Kawashima, M. (2022). KEGG mapping tools for uncovering
1030 hidden features in biological data. *Protein Science*, *31*(1), 47–53.
1031 https://doi.org/10.1002/pro.4172
1032 Kapopoulou, A., Lew, J. M., & Cole, S. T. (2011). The MycoBrowser portal: A
1033 comprehensive and manually annotated resource for mycobacterial genomes.
1034 *Tuberculosis*, *91*(1), 8–13. https://doi.org/10.1016/j.tube.2010.09.006
1035 Kendall, S. L., Burgess, P., Balhana, R., Withers, M., Ten Bokum, A., Lott, J. S., Gao, C.,
1036 Uhia-Castro, I., & Stoker, N. G. (2010). Cholesterol utilization in mycobacteria is
1037 controlled by two TetR-type transcriptional regulators: KstR and kstR2.
1038 *Microbiology*, *156*(5), 1362–1371. https://doi.org/10.1099/mic.0.034538-0
1039 Kendall, S. L., Withers, M., Soffair, C. N., Moreland, N. J., Gurcha, S., Sidders, B., Frita, R.,
1040 Ten Bokum, A., Besra, G. S., Lott, J. S., & Stoker, N. G. (2007). A highly conserved
1041 transcriptional repressor controls a large regulon involved in lipid degradation in
1042 Mycobacterium smegmatis and Mycobacterium tuberculosis. *Molecular*
1043 *Microbiology*, *65*(3), 684–699. https://doi.org/10.1111/j.1365-2958.2007.05827.x
1044 Kipkorir, Terry, Mashabela, Gabriel T., de Wet, Timothy J., Koch, Anastasia, Dawes
1045 Stephanie S., Wiesner, Lubbe, Mizrahi, Valerie, Warner, Digby F., & Henkin, Tina
1046 M. (2021). De Novo Cobalamin Biosynthesis, Transport, and Assimilation and
1047 Cobalamin-Mediated Regulation of Methionine Biosynthesis in Mycobacterium

1048 smegmatis. *Journal of Bacteriology*, *203*(7), e00620-20.
1049 https://doi.org/10.1128/JB.00620-20
1050 Lamichhane, G., Arnvig, K. B., & McDonough, K. A. (2013). Definition and annotation of
1051 (myco)bacterial non-coding RNA. *Tuberculosis*, *93*(1), 26–29.
1052 https://doi.org/10.1016/j.tube.2012.11.010
1053 Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation
1054 network analysis. *BMC Bioinformatics*, *9*. https://doi.org/10.1186/1471-2105-9-559
1055 Lejars, M., Kobayashi, A., & Hajnsdorf, E. (2019). Physiological roles of antisense RNAs in
1056 prokaryotes. *Biochimie*, *164*, 3–16. https://doi.org/10.1016/j.biochi.2019.04.015
1057 Li, Heng. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-*
1058 *MEM*. https://doi.org/10.48550/arXiv.1303.3997
1059 Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and
1060 dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21.
1061 https://doi.org/10.1186/s13059-014-0550-8
1062 Lu, L., Wei, R., Bhakta, S., Waddell, S. J., & Boix, E. (2021). Weighted gene co-expression
1063 network analysis to identify key modules and hub genes associated with
1064 paucigranulocytic asthma. *Antibiotics*, *10*(97).
1065 https://doi.org/10.3390/antibiotics10020097
1066 Lunge, A., Gupta, R., Choudhary, E., & Agarwal, N. (2020). The unfoldase ClpC1 of
1067 Mycobacterium tuberculosis regulates the expression of a distinct subset of proteins
1068 having intrinsically disordered termini. *Journal of Biological Chemistry*, *295*(28),
1069 9455–9473. https://doi.org/10.1074/jbc.RA120.013456
1070 Maciąg Anna, Dainese Elisa, Rodriguez G. Marcela, Milano Anna, Provvedi Roberta, Pasca
1071 Maria R., Smith Issar, Palù Giorgio, Riccardi Giovanna, & Manganelli Riccardo.

1072 (2007). Global Analysis of the Mycobacterium tuberculosis Zur (FurB) Regulon.
1073 *Journal of Bacteriology*, *189*(3), 730–740. https://doi.org/10.1128/JB.01190-06
1074 Mai, J., Rao, C., Watt, J., Sun, X., Lin, C., Zhang, L., & Liu, J. (2019). Mycobacterium
1075 tuberculosis 6C sRNA binds multiple mRNA targets via C-rich loops independent of
1076 RNA chaperones. *Nucleic Acids Research*, *47*(8), 4292–4307.
1077 https://doi.org/10.1093/nar/gkz149
1078 Martini, M. C., Zhou, Y., Sun, H., & Shell, S. S. (2019). Defining the Transcriptional and
1079 Post-transcriptional Landscapes of Mycobacterium smegmatis in Aerobic Growth and
1080 Hypoxia. In *Frontiers in Microbiology* (Vol. 10).
1081 https://www.frontiersin.org/article/10.3389/fmicb.2019.00591
1082 Menendez-Gil, P., Caballero, C., Catalan-Moreno, A., Irurzun, N., Barrio-Hernandez, I.,
1083 Caldelari, I., & Toledo-Arana, A. (2020). Differential evolution in 3'UTRs leads to
1084 specific gene expression in Staphylococcus. *Nucleic Acids Research*, *48*.
1085 https://doi.org/10.1093/nar/gkaa047
1086 Menendez-Gil, P., & Toledo-Arana, A. (2021). Bacterial 3′UTRs: A Useful Resource in Post-
1087 transcriptional Regulation. *Frontiers in Molecular Biosciences*, *7*.
1088 https://www.frontiersin.org/article/10.3389/fmolb.2020.617633
1089 Modlin, S. J., Afif, E., Deepika, G., Zlotnicki, A. M., Dillon, N. A., Dhillon, N., Kuo, N.,
1090 Robinhold, C., Chan, C. K., Baughn, A. D., & Valafar, F. (2021). Structure-Aware
1091 Mycobacterium tuberculosis Functional Annotation Uncloaks Resistance, Metabolic,
1092 and Virulence Genes. *MSystems*, *0*(0), e00673-21.
1093 https://doi.org/10.1128/mSystems.00673-21
1094 Moores, A., Riesco, A. B., Schwenk, S., & Arnvig, K. B. (2017). Expression, maturation and
1095 turnover of DrrS, an unusually stable, DosR regulated small RNA in Mycobacterium

1096 tuberculosis. *PLOS ONE*, *12*(3), e0174079.
1097 https://doi.org/10.1371/journal.pone.0174079
1098 Nesbitt, N. M., Yang, X., Fontán, P., Kolesnikova, I., Smith, I., Sampson, N. S., & Dubnau,
1099 E. (2010). A Thiolase of Mycobacterium tuberculosis Is Required for Virulence and
1100 Production of Androstenedione and Androstadienedione from Cholesterol. *Infection*
1101 *and Immunity*, *78*(1), 275 LP – 282. https://doi.org/10.1128/IAI.00893-09
1102 Ozuna, A., Liberto, D., Joyce, R. M., Arnvig, K. B., & Nobeli, I. (2019). baerhunter: An R
1103 package for the discovery and analysis of expressed non-coding regions in bacterial
1104 RNA-seq data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz643
1105 Pawełczyk, J., Brzostek, A., Minias, A., Płociński, P., Rumijowska-Galewicz, A., Strapagiel,
1106 D., Zakrzewska-Czerwińska, J., & Dziadek, J. (2021). Cholesterol-dependent
1107 transcriptome remodeling reveals new insight into the contribution of cholesterol to
1108 Mycobacterium tuberculosis pathogenesis. *Scientific Reports*, *11*(1), 12396.
1109 https://doi.org/10.1038/s41598-021-91812-0
1110 Ponath, F., Hör, J., & Vogel, J. (2022). An overview of gene regulation in bacteria by small
1111 RNAs derived from mRNA 3' ends. *FEMS Microbiology Reviews*, fuac017.
1112 https://doi.org/10.1093/femsre/fuac017
1113 Puniya, B. L., Kulshreshtha, D., Verma, S. P., Kumar, S., & Ramachandran, S. (2013).
1114 Integrated gene co-expression network analysis in the growth phase of
1115 Mycobacterium tuberculosis reveals new potential drug targets. *Molecular*
1116 *BioSystems*, *9*(11), 2798–2815. https://doi.org/10.1039/c3mb70278b
1117 Ramage, H. R., Connolly, L. E., & Cox, J. S. (2009). Comprehensive functional analysis of
1118 Mycobacterium tuberculosis toxin-antitoxin systems: Implications for pathogenesis,
1119 stress responses, and evolution. *PLoS Genetics*, *5*(12).
1120 https://doi.org/10.1371/journal.pgen.1000767

1121 Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015).
1122 Limma powers differential expression analyses for RNA-sequencing and microarray
1123 studies. *Nucleic Acids Research*, *43*(7), e47–e47. https://doi.org/10.1093/nar/gkv007
1124 Rustad, T. R., Harrell, M. I., Liao, R., & Sherman, D. R. (2008). The enduring hypoxic
1125 response of Mycobacterium tuberculosis. *PLoS ONE*, *3*(1), 1–8.
1126 https://doi.org/10.1371/journal.pone.0001502
1127 Saelens, W., Cannoodt, R., & Saeys, Y. (2018). A comprehensive evaluation of module
1128 detection methods for gene expression data. *Nature Communications*, *9*(1), 1090.
1129 https://doi.org/10.1038/s41467-018-03424-4
1130 Sáenz-Lahoya S., Bitarte N., García B., Burgui S., Vergara-Irigaray M., Valle J., Solano C.,
1131 Toledo-Arana A., & Lasa I. (2019). Noncontiguous operon is a genetic organization
1132 for coordinating bacterial gene expression. *Proceedings of the National Academy of*
1133 *Sciences*, *116*(5), 1733–1738. https://doi.org/10.1073/pnas.1812746116
1134 Sala, A., Bordes, P., & Genevaux, P. (2014). Multiple toxin-antitoxin systems in
1135 Mycobacterium tuberculosis. *Toxins*, *6*(3), 1002–1020.
1136 https://doi.org/10.3390/toxins6031002
1137 Sawyer, E. B., Phelan, J. E., Clark, T. G., & Cortes, T. (2021). A snapshot of translation in
1138 Mycobacterium tuberculosis during exponential growth and nutrient starvation
1139 revealed by ribosome profiling. *Cell Reports*, *34*(5).
1140 https://doi.org/10.1016/j.celrep.2021.108695
1141 Schwenk, S., & Arnvig, K. B. (2018). Regulatory RNA in Mycobacterium tuberculosis, back
1142 to basics. *Pathogens and Disease*, *76*(4). https://doi.org/10.1093/femspd/fty035
1143 Serafini, A., Pisu, D., Palù, G., Rodriguez, G. M., & Manganelli, R. (2013). The ESX-3
1144 Secretion System Is Necessary for Iron and Zinc Homeostasis in Mycobacterium
1145 tuberculosis. *PLoS ONE*, *8*(10), 1–15. https://doi.org/10.1371/journal.pone.0078351

1146 Shell, S. S., Wang, J., Lapierre, P., Mir, M., Chase, M. R., Pyle, M. M., Gawande, R.,
1147 Ahmad, R., Sarracino, D. A., Ioerger, T. R., Fortune, S. M., Derbyshire, K. M., Wade,
1148 J. T., & Gray, T. A. (2015). Leaderless Transcripts and Small Proteins Are Common
1149 Features of the Mycobacterial Translational Landscape. *PLOS Genetics*, *11*(11),
1150 e1005641. https://doi.org/10.1371/journal.pgen.1005641
1151 Shockey, A. C., Dabney, J., & Pepperell, C. S. (2019). Effects of Host, Sample, and in vitro
1152 Culture on Genomic Diversity of Pathogenic Mycobacteria. In *Frontiers in Genetics*
1153 (Vol. 10). https://www.frontiersin.org/article/10.3389/fgene.2019.00477
1154 Šiková, M., Janoušková, M., Ramaniuk, O., Páleníková, P., Pospíšil, J., Bartl, P., Suder, A.,
1155 Pajer, P., Kubičková, P., Pavliš, O., Hradilová, M., Vítovská, D., Šanderová, H.,
1156 Převorovský, M., Hnilicová, J., & Krásný, L. (2019). Ms1 RNA increases the amount
1157 of RNA polymerase in Mycobacterium smegmatis. *Molecular Microbiology*, *111*(2),
1158 354–372. https://doi.org/10.1111/mmi.14159
1159 Singh Prabhat Ranjan, Vijjamarri Anil Kumar, Sarkar Dibyendu, & Federle Michael J.
1160 (2020). Metabolic Switching of Mycobacterium tuberculosis during Hypoxia Is
1161 Controlled by the Virulence Regulator PhoP. *Journal of Bacteriology*, *202*(7),
1162 e00705-19. https://doi.org/10.1128/JB.00705-19
1163 Smith, C., Canestrari, J. G., Wang, A. J., Champion, M. M., Derbyshire, K. M., Gray, T. A.,
1164 & Wade, J. T. (2022). Pervasive translation in Mycobacterium tuberculosis. *ELife*, *11*,
1165 e73980. https://doi.org/10.7554/eLife.73980
1166 Solans, L., Gonzalo-Asensio, J., Sala, C., Benjak, A., Uplekar, S., Rougemont, J., Guilhot,
1167 C., Malaga, W., Martín, C., & Cole, S. T. (2014). The PhoP-Dependent ncRNA Mcr7
1168 Modulates the TAT Secretion System in Mycobacterium tuberculosis. *PLOS*
1169 *Pathogens*, *10*(5), e1004183. https://doi.org/10.1371/journal.ppat.1004183

1170 Sreenu, V. B., Kumar, P., Nagaraju, J., & Nagarajaram, H. A. (2007). Simple sequence
1171 repeats in mycobacterial genomes. *Journal of Biosciences*, *32*(1), 3–15.
1172 https://doi.org/10.1007/s12038-007-0002-7
1173 Stiens, J., Arnvig, K. B., Kendall, S. L., & Nobeli, I. (2022). Challenges in defining the
1174 functional, non-coding, expressed genome of members of the Mycobacterium
1175 tuberculosis complex. *Molecular Microbiology*, *117*(1), 20–31.
1176 https://doi.org/10.1111/mmi.14862
1177 Talwar, S., Pandey, M., Sharma, C., Kutum, R., Lum, J., Carbajo, D., Goel, R., Poidinger,
1178 M., Dash, D., Singhal, A., & Pandey, A. K. (2020). Role of VapBC12 Toxin-
1179 Antitoxin Locus in Cholesterol-Induced Mycobacterial Persistence. *MSystems*, *5*(6).
1180 https://doi.org/10.1128/msystems.00855-20
1181 The Gene Ontology Consortium. (2021). The Gene Ontology resource: Enriching a GOld
1182 mine. *Nucleic Acids Research*, *49*(D1), D325–D334.
1183 https://doi.org/10.1093/nar/gkaa1113
1184 Toledo-Arana, A., & Lasa, I. (2020). Advances in bacterial transcriptome understanding:
1185 From overlapping transcription to the excludon concept. *Molecular Microbiology*,
1186 *113*(3), 593–602. https://doi.org/10.1111/mmi.14456
1187 Updegrove, T. B., Kouse, A. B., Bandyra, K. J., & Storz, G. (2019). Stem-loops direct precise
1188 processing of 3′ UTR-derived small RNA MicL. *Nucleic Acids Research*, *47*(3),
1189 1482–1492. https://doi.org/10.1093/nar/gky1175
1190 Vargas-Blanco, D. A., & Shell, S. S. (2020). Regulation of mRNA Stability During Bacterial
1191 Stress Responses. *Frontiers in Microbiology*, *11*(September).
1192 https://doi.org/10.3389/fmicb.2020.02111

1193 Voskuil, M. I., Visconti, K. C., & Schoolnik, G. K. (2004). Mycobacterium tuberculosis gene
1194 expression during adaptation to stationary phase and low-oxygen dormancy.
1195 *Tuberculosis*, *84*(3–4), 218–227. https://doi.org/10.1016/j.tube.2004.02.003
1196 Wade, J. T., & Grainger, D. C. (2014). Pervasive transcription: Illuminating the dark matter
1197 of bacterial transcriptomes. *Nature Reviews Microbiology*, *12*(9), 647–653.
1198 https://doi.org/10.1038/nrmicro3316
1199 Wang, X., Monford Paul Abishek, N., Jeon, H. J., Lee, Y., He, J., Adhya, S., & Lim, H. M.
1200 (2019). Processing generates 3′ ends of RNA masking transcription termination
1201 events in prokaryotes. *Proceedings of the National Academy of Sciences of the United*
1202 *States of America*, *116*(10), 4440–4445. https://doi.org/10.1073/pnas.1813181116
1203 Warman, E. A., Forrest, D., Guest, T., Haycocks, J. J. R. J., Wade, J. T., & Grainger, D. C.
1204 (2021). Widespread divergent transcription from bacterial and archaeal promoters is a
1205 consequence of DNA-sequence symmetry. *Nature Microbiology*, *6*(6), 746–756.
1206 https://doi.org/10.1038/s41564-021-00898-9
1207 Warner, D. F., Savvi, S., Mizrahi, V., & Dawes, S. S. (2007). A Riboswitch Regulates
1208 Expression of the Coenzyme B12-Independent Methionine Synthase in
1209 Mycobacterium tuberculosis: Implications for Differential Methionine Synthase
1210 Function in Strains H37Rv and CDC1551. *Journal of Bacteriology*, *189*(9), 3655 LP
1211 – 3659. https://doi.org/10.1128/JB.00040-07
1212 World Health Organization. (2021, October 14). *Tuberculosis Fact Sheet*. Tuberculosis.
1213 https://www.who.int/news-room/fact-sheets/detail/tuberculosis
1214 Yoo, Reo, Rychel, Kevin, Poudel, Saugat, Al-bulushi, Tahani, Yuan Yuan, Chauhan,
1215 Siddharth, Lamoureux, Cameron, Palsson, Bernhard O., Sastry, Anand, & Tringe,
1216 Susannah Green. (2022). Machine Learning of All Mycobacterium tuberculosis
1217 H37Rv RNA-seq Data Reveals a Structured Interplay between Metabolism, Stress

1218 Response, and Infection. *MSphere*, *0*(0), e00033-22.
1219 https://doi.org/10.1128/msphere.00033-22
1220 Zhang, B., & Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression
1221 Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1).
1222 https://doi.org/10.2202/1544-6115.1128
1223 Zhang, L., Hendrickson, R. C., Meikle, V., Lefkowitz, E. J., Ioerger, T. R., & Niederweis, M.
1224 (2020). Comprehensive analysis of iron utilization by Mycobacterium tuberculosis.
1225 *PLOS Pathogens*, *16*(2), e1008337. https://doi.org/10.1371/journal.ppat.1008337
1226 Zhou, Y., Huang, H., Zhou, P., & Xie, J. (2012). Molecular mechanisms underlying the
1227 function diversity of transcriptional factor IclR family. *Cellular Signalling*, *24*(6),
1228 1270–1275. https://doi.org/10.1016/j.cellsig.2012.02.008
1229 Zondervan, N. A., Van Dam, J. C. J., Schaap, P. J., Martins dos Santos, V. A. P., & Suarez-
1230 Diez, M. (2018). Regulation of Three Virulence Strategies of Mycobacterium
1231 tuberculosis: A Success Story. *International Journal of Molecular Sciences*, *19*(2).
1232 https://doi.org/10.3390/ijms19020347
1233