Challenges in defining the functional, non-coding, expressed genome of pathogenic mycobacteria

Jennifer Stiens,¹ Kristine B. Arnvig,² Sharon L. Kendall³ and Irene Nobeli¹\$

¹Institute of Structural and Molecular Biology, Biological Sciences, Birkbeck, University of London, Malet Street, London, WC1E 7HX, UK

²Institute of Structural and Molecular Biology, Division of Biosciences, University College London, London, WC1E 6BT, UK

³Centre for Emerging, Endemic and Exotic Diseases, Pathobiology and Population Sciences, Royal Veterinary College, Hawkshead Lane, North Mymms, Hatfield, AL9 7TA, UK

\$ Corresponding author: i.nobeli@bbk.ac.uk

Keywords

Mycobacteria; *Mycobacterium tuberculosis*; non-coding RNA; RNA-seq; transcriptome

Abbreviations

Mtb = *Mycobacterium tuberculosis*

MTBC = *Mycobacterium tuberculosis* complex

ncRNA = non-coding RNA

sRNA = short RNA

asRNA = antisense RNA

UTR = untranslated region

nt = nucleotide(s)

RBS = ribosome binding site

TSS = transcription start site

TTS = transcription termination site

Abstract

A definitive transcriptome atlas for the non-coding expressed elements of pathogenic mycobacteria does not exist. Incomplete lists of non-coding transcripts can be obtained for some of the reference genomes (e.g. *Mycobacterium tuberculosis* H37Rv) but to what extent these transcripts have homologues in closely related species or even strains is not clear. This has implications for the analysis of transcriptomic data; non-coding parts of the transcriptome are often ignored in the absence of formal, reliable annotation. Here, we review the state of our knowledge of non-coding RNAs in pathogenic mycobacteria, emphasising the disparities in the information included in commonly used databases. We then proceed to review ways of combining computational solutions for predicting the non-coding transcriptome with experiments that can help refine and confirm these predictions.

Introduction

A definitive atlas of expressed non-coding elements in pathogenic mycobacteria does not exist. The lists available from databases and publications overlap only partially and are only available for the reference genomes of key representatives of the *Mycobacterium tuberculosis* complex (MTBC), such as *Mycobacterium tuberculosis* (*Mtb*) H37Rv. This gap in our knowledge impacts the successful analysis of the copious amounts of genomic and transcriptomic data that have become available in the last decade. For example, in the absence of a formal annotation of the non-coding transcriptome, the easiest and most common approach to call differential expression events is to largely, or entirely, ignore information that does not relate to regions currently annotated as coding (CDS); this issue is more acute in studies focusing on non-reference *Mtb* strains or their close relatives, where non-coding annotation is scarce or non-existent. In this commentary, inspired by our own struggles to compile a definitive atlas of 'non-coding' RNA (using the term here to represent regulatory RNAs such as short RNAs, antisense RNAs and the untranslated parts of mRNA transcripts) in pathogenic mycobacteria, we present a summary of the current information from publicly available sources, highlighting the existing gaps in the knowledge and the computational approaches used to attempt to uncover this less well understood part of the mycobacterial genome.

Why pathogenic mycobacteria and why non-coding RNA?

Prior to the COVID-19 pandemic, mycobacterial disease was the leading cause of death by a single pathogen; causing over 1.4 million deaths, and infecting over 10 million people in 2019, worldwide (<https://www.who.int/news-room/fact-sheets/detail/tuberculosis>). The different members of the MTBC include both human-adapted (*Mtb*) and animal-adapted (*Mycobacterium bovis*, *Mycobacterium caprae*, among others) species which show distinct host preference (Brites et al., 2018). Each of these species is uniquely adapted to cause disease within their preferred hosts, and to navigate a complicated lifecycle, which requires rapid response to changing environmental conditions. Pathogenic bacteria have different programs for invasion, proliferation and survival in particular host environments. The pathogen can rapidly and transiently adapt to environmental changes brought about by host

defences by regulating the effect and stability of the transcripts through the parsimonious action of post-transcriptional regulation (Chakravarty & Massé, 2019).

Though the different members of the MTBC have different tropisms, involving specific virulence profiles and metabolic changes made in response to the host environment, nearly 99% of the genomic sequence is conserved among the MTBC members (Malone and Gordon, 2017). The minor variations such as deletions and single nucleotide polymorphisms (SNPs) that vary among species members of the MTBC, and between species-specific strains, seem to have an outsized role determining these preferences (Dinan et al., 2014); (Malone et al., 2018); (Chiner-Oms et al., 2019); (Cheng et al., 2019). Therefore, it is logical to hypothesize that non-coding RNA (ncRNA) is involved at both the transcriptional and post-transcriptional level of gene regulation, magnifying small genetic changes and creating the documented level of phenotypic diversity (Schwenk & Arnvig, 2018).

Advances made in recent years exploring the non-coding genome, especially in the model organisms, have shown how flexible and adaptive riboregulation can be. As a full description of ncRNAs is outside the scope of this commentary, we present instead a graphical summary of the main types, mechanisms of action and targets in Table 1. There are several recent and comprehensive reviews that describe different aspects of the constantly evolving roster of non-coding elements in bacterial genomes, but most of them focus on what has been discovered in the model organisms (Table 2). Mycobacteria are different, in genome, physiology and lifestyle; and it appears that non-coding regulation in MTBC does not use the same accessory proteins or have the same sequence signatures as the model systems. Indeed, efforts to find an Hfq or ProQ analog acting as an RNA chaperone in mycobacteria have so far been unsuccessful (Gerrick, 2018). These differences impact not only on our ability to transfer knowledge from model organisms to the MTBC species, but also on how applicable current experimental and computational methods are to discovering new regulators in mycobacteria.

Table 1. Origins and targets of non-coding RNA types in bacteria.

ncRNA Type	Description	Origin	Target/ Mechanism
<p>sRNA</p>	Short RNA fragments, 30-300nt; involved in binding interactions with distant gene targets ('trans-acting') to regulate translation	Intergenic regions or UTRs; transcribed from own promoter, or by cleavage of longer transcripts	mRNAs of other genes (e.g. UTRs of transcription factors), other sRNAs (known as 'sponge sRNAs' or 'ceRNAs') to regulate sRNA abundance
<p>asRNA</p>	RNA transcript, 75-10,000nt; regulates by binding to mRNA transcript with perfect base-pairing, forming duplex RNA: altering sensitivity to RNases, action of terminators, or access to RBS (ribosome binding site)	Complementary strand of UTR or coding sequence of regulated gene; transcribed from own promoter	Cognate RNA strand
<p>Riboswitch</p>	Structured 5' UTR that changes secondary structure in response to ligand binding	In 5'UTR of gene under regulatory control	Regulates expression of cognate gene through formation of intrinsic terminator, or by regulating access to Rho-dependent terminator or RBS
<p>5'/3' UTR</p>	<p>5-100s nt long and regulate the downstream/upstream ORFs ('cis-regulatory')</p> <p>5' includes the RBS, cleavage sites for different transcript lengths, sRNAs and riboswitches and other cis-regulatory elements</p> <p>3' can include sRNAs binding sites and RNase cleavage sites to regulate transcript length</p>	<p>5': upstream sequence of coding sequence, between TSS and start codon (alternative TSS may exist in gene locus)</p> <p>3': downstream sequence between stop codon and transcription terminator (alt. TTS may exist in gene locus)</p>	<p>Binding interactions with sRNAs, proteins, metabolites, second messengers</p> <p>Antisense binding with other UTRs</p>

Table 2. Recent reviews of non-coding RNA in bacteria and their focus.

Author	Organism	Focus
Adams and Storz, 2020	Model organisms	sRNA origins/discovery
Bossi et al., 2020	Model organisms	sRNAs and Rho-dependent termination
Breaker, 2018	Model organisms	Riboswitches and translational control
Chakravarty and Massé, 2019	Model organisms	Regulatory RNA and virulence
Denham, 2020	Model organisms	Sponge RNAs (post-transcriptional regulation of sRNAs)
Hör et al., 2018	Model organisms	All regulatory RNA, global RNA-seq methods
Jørgenson et al., 2020	Gram positive and negative bacteria, including mycobacteria	sRNA-mediated regulation
Schwenk and Arnvig, 2018.	Model organisms and mycobacteria	All regulatory RNA, especially in mycobacteria
Toledo-Arana and Lasa, 2020	Model organisms	Overlapping transcripts in transcriptome organisation: asRNA, excludons

How many functional non-coding RNAs are there in mycobacteria?

Only a handful of sRNAs have been functionally characterised in the mycobacteria literature (Table 3). In most cases, top-down approaches, such as differential expression studies and ChIP-seq (chromatin immunoprecipitation with sequencing), have been employed to discover *Mtb* sRNAs, such as the RNAP-associated, Ms1 (Arnvig et al., 2011; Šiková et al., 2019) and the PhoP-regulated, Mcr7 (Solans et al., 2014). It is curious, that even among the six well-characterised examples in Table 3, there is one (MrsI) not listed in the current official annotation of the reference H37Rv genome, available from the corresponding NCBI annotation (GFF) file (GCF_000195955.2_ASM19595v2_genomic.gff), most likely because it was a recent discovery. This annotation file currently includes 20 features labelled as non-

coding RNAs, 15 of which are listed in (Arnvig et al., 2011) and 9 in (DiChiara et al., 2010)- 4 are listed in both. It also includes 10 “sequence features” which are annotated as fragments of putative small regulatory RNAs (8 matching information from DiChiara (DiChiara et al., 2010) and 2 matching information from Pelly (Pelly et al., 2012), and two “misc RNA” including a tmRNA and the ribonuclease P RNA. Although twenty or even thirty non-coding elements is almost certainly an underestimate of the total number of ncRNAs in *Mtb*, we note here that the corresponding *E. coli* reference genome annotation (GCF_000005845.2_ASM584v2_genomic.gff) contains currently 72 elements labelled ncRNAs, suggesting that either functional non-coding elements are not very common in bacteria, or that, even for a well-studied organism, our understanding of non-coding regulation is incomplete.

Table 3. Functionally characterised sRNA in mycobacteria. *Annotation according to Lamichlane et al., 2013.

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

Whereas functional characterisation is ultimately needed to create a reliable list of non-coding RNAs, homology to known families of RNAs from other organisms remains the most popular approach for predicting non-coding RNAs in the absence of experimental evidence. The RNA families described in the RFAM database (Kalvari et al., 2021) derive from the

application of covariance models (and where structure information is not available, Hidden Markov Models) representing meticulously curated multiple sequence and secondary structure alignments of homologous RNAs. RFAM thus represents some of the most reliable predictions for non-coding elements in genomes and its predictions for *Mtb* H37Rv are summarised in Table 4. As conservation of structure is at the heart of RFAM families, non-coding RNAs with few or no known relatives in other species, and those that do not fold into strongly conserved structures, are unlikely to be found in RFAM. Hence, this database too is likely to miss elements that are specific to a small number of pathogenic mycobacteria or that are too short to fold into a stable structure. In general, homology-based approaches to discovering *novel* non-coding elements will be limited in pathogenic mycobacteria as there are few closely-related genomes outside the phyla. One notable exception, 6C sRNA, is well-conserved among gram-positive bacteria with over-expression leading to altered growth phenotypes in *M. tuberculosis*, *M. smegmatis* and another GC-rich bacterium, *Corynebacterium glutamicum*. Perhaps as a result, it is one of the few sRNAs for which target molecules have been identified and experimentally validated (Mai et al., 2019).

Table 4. Conserved non-coding RNA families and sequence listings from the RFAM database (<https://rfam.xfam.org>). Ribozymes (Group II catalytic introns and Bacterial RNase P class A), tRNAs and rRNAs have not been included in this table.

RNA Type	Family Name	Rfam ID	Number Sequences in RFAM	Length (nt)
Riboswitch	Cobalamin (B ₁₂)	RF00174	2	173-218
Riboswitch	ykok leader/Mbox (Mg ⁺)	RF00380	2	169-174
Riboswitch	TPP/Thi-box (thiamine)	RF00059	2	110
Riboswitch	ydaO/yuA leader (Cyclic di-AMP)	RF00379	1	222
Riboswitch	Glycine	RF00504	2	90-97
Riboswitch	S-adenosyl methionine (SAM-IV)	RF00634	1	119
sRNA	Mcr7	RF02671	1	348
sRNA	npcTB_6715	RF02886	2	211
sRNA	Ms1	RF02566	1	301
sRNA	ncRv12659	RF02659	1	171

sRNA	ncrMT1302	RF02341	1	108
sRNA	b55	RF01783	1	60
sRNA/asRNA	ASdes	RF0781	2	67
sRNA	F6	RF01791	1	57
sRNA	Ms_AS-5	RF02465	1	44
sRNA/5'UTR	5_ureB_sRNA	RF02514	1	294
asRNA	ASpks	RF01782	5	68-77

Expanding our exploration to resources beyond the official NCBI annotation, further complicates the question of what is known about functional, non-coding RNAs in mycobacteria. Mycobrowser (Kapopoulou et al., 2011), arguably the most popular internet resource for the exploration of representative mycobacterial genomes, currently lists 92 non-coding RNAs, labelled as 'ncRNA' (including sRNAs and asRNAs under this moniker) for H37Rv: 40 overlap the official NCBI GFF annotation and originate from the four key publications listing experimentally-verified non-coding RNAs (DiChiara et al., 2010); (Arnvig et al., 2011); (Arnvig & Young, 2009); (Pelly et al., 2012) and the remaining 52 overlap the list compiled by Gerrick et al. (2018) using their in-house computational tool, *BS_finder*, applied to RNA-seq data derived with a small-RNA sequencing protocol. Despite including annotations from nine other species and strains, including *M. bovis*, *M. smegmatis* and *M. tuberculosis 18b*, non-coding RNAs annotated with the tag "ncRNA" appear in the GFF files of only three additional species/strains in Mycobrowser, and only *M. tuberculosis 18b* has more than two ncRNAs listed. Strikingly, *M. bovis*, sharing more than >99.95% sequence identity to *M. tuberculosis*, has no other entries for RNAs apart from rRNAs, tRNAs and the same two RNAs tagged "misc_RNA" in the *Mtb* annotation; it is highly unlikely that many of the ncRNAs present in *M. tuberculosis* do not have a counterpart in *M. bovis*; and thus the list must be assumed to be incomplete. In fact, at least 41 of experimentally-verified sRNAs found in various mycobacterial species, including in the above studies, can be mapped to the *M. bovis* genome (Dinan et al., 2014). Perhaps more worryingly, the RNase P RNA component is alternatively tagged as *ncRNA* in *M. haemophilium* and *M. oryzae*, *RNase_P_RNA* in *M. tuberculosis 18b* and *misc_RNA* in *M. bovis*. The lack of standardised tags and incomplete listings of non-coding

elements (even within the same resource), together with the absence of a clear justification for which elements are included and why, likely adds to the confusion about non-coding regulation in mycobacteria. A more systematic approach to the annotation tags of these elements, similar to approaches suggested for consistent naming of non-coding RNA (Lamichhane et al., 2013), could go some way towards eliminating this confusion.

Completing the non-coding transcript atlas: computational predictions from genomic and transcriptomic data

The most extensive lists of putative non-coding RNAs in mycobacteria are the result of computational predictions based on genomic or transcriptomic data (or sometimes both). Genomics-based methods rely on the conservation of non-coding elements across several species and, like RFAM, are likely to miss elements specific to a small subset of the genus or unique to a species. Such comparative genomics methods are typically enhanced by the search for characteristic sequence features and other signals of regulatory RNAs such as promoters, terminator structures and transcription-factor binding sites. For example, SIPHT begins with conserved intergenic sequences (defined as the sequence between two annotated genes/ORFs on one strand) and looks for characteristic features of sRNAs in these regions, such as conserved promoters and rho-independent terminator motifs (Livny et al., 2008). Other genomics-based programs rely entirely on sequence features and genomic context (ignoring conservation). sRNAScanner determines intergenic sequences using genome annotation files and differentiates coding from non-coding sequences using position scoring matrices for sequence signals such as RBS and start codons (Sridhar et al., 2010). A recently published tool, the Pred-GsRNA feature of the PresRAT server, extracts intergenic sequences, also based on genome annotation, and excludes candidates that have an 8 nt sequence found to be depleted in known sRNAs. It scores each predicted sequence with weighted Minimum Free Energy scores for predicted paired and loop regions and scores for the predicted U-rich consensus sequences typical of intrinsic terminators (Kumar et al., 2020). The server offers 405 possible 'non-genic sRNA' predictions for the *M. tuberculosis* H37Rv genome (<http://www.hpppi.iicb.res.in/presrat/>). We compared the predicted sRNA coordinates with the coordinates of the 92 'stable' RNAs in the H37Rv genome on Mycobrowser (<https://mycobrowser.epfl.ch>). There were no PresRAT predicted sRNAs overlapping the boundaries of the Mycobrowser listed RNAs, except for low-ranking

predictions that were over 4000 bp long, indicating that this method has limited power to recognise intergenic sRNA elements in mycobacterial genomes. It is clear that searching the intergenic regions for conserved sequence features based on sRNAs discovered in the model organisms is less effective in mycobacteria, as promoters are very diverse, and apart from the SigA -10 consensus motif, not as well-conserved (Newton-Foot & Gey van Pittius, 2013). Mycobacterial sRNAs also often lack the recognisable intrinsic terminator motifs at their 3' ends typical of chaperone protein-binding sRNAs (Moores et al., 2017), (DiChiara et al., 2010), (Arnvig et al., 2014). Furthermore, narrowing the search strictly to intergenic regions (with no annotated genes on either strand) effectively ignores sRNAs and asRNAs generated from coding or antisense regions.

Transcriptomics-based methods are usually versions of sliding window approaches looking for abrupt increases and drops in the expression signal and using such changes to delineate the limits of putative non-coding elements. High-throughput RNA sequencing (RNA-seq) has uncovered a multitude of short transcripts from intergenic sequences, 5' and 3' UTRs and antisense to coding regions. Sorting the signal from the noise is the main challenge when using these data in non-coding RNA discovery. For example, sensitive methods are able to pick up expressed elements in regions of low read coverage; this signal may represent true low-abundance transcripts but it can also be the result of either technical or transcriptional noise (due to stochastic gene expression). The more sensitive computational methods will therefore inevitably over-predict putative non-coding elements. Ironically, high-depth sequencing has magnified this problem (Tarazona et al., 2011), (Mao et al., 2015). Non-fragmented, size-selected libraries, where small transcripts remain intact, are superior for discerning between signal and noise for small RNA transcripts (Leonard et al., 2019), (Wang et al., 2016). For all the reasons discussed above, detecting the existence of sRNAs expressed in low levels against very strongly expressed coding genes remains a computational challenge. Here, we also suggest caution when using publicly available transcriptomic data, some of which dates back to the early use of RNA-seq technologies. In particular, using strand-specific cDNA libraries sequencing, where the information about which strand the transcript originates from is preserved, is invaluable to the discovery of new ncRNAs. Preservation of the strand information avoids mis-mapping asRNAs or other overlapping sRNAs that might otherwise be mapped to a coding gene on the opposite strand.

Many labs have developed their own computational pipelines and scripts to map RNA-seq data, normalise signals and identify ncRNA transcripts across the genome. (Ami et al., 2020), (Dejesus et al., 2017), (Wang et al., 2016), (Gómez-Lozano et al., 2014), (Miotto et al., 2012) whereas others have carried out this process semi-manually (Arnvig et al., 2011). Progress in the field, and an easy comparison between approaches, has been hindered by the fact that few of the labs publishing computational predictions have made their code readily available. In response to this challenge, several groups have created publicly-available prediction programs or workflows such as *Rockhopper* (McClure et al., 2013), *DETR'PROK* (Toffano-Nioche et al., 2013), *ANNOgesic* (Yu et al., 2018), *APER0* (Leonard et al., 2019) and *baerhunter* (Ozuna et al., 2019). Users of all of these transcriptomics-based methods are required to set thresholds for separating background noise (whatever its origin) from signal in the data. Indeed, most programs need adjustment to their default parameters in order to respond to sequencing depth and signal abundance (Figure 1) but tuning these parameters can be a matter of art rather than science.

Figure 1. Challenges of predicting non-coding expressed elements from transcriptomic signal alone. Coverage views from two real and one hypothetical *Mtb* transcriptomic dataset: Illumina high-throughput sequencing datasets from Bioproject accession numbers PRJNA278760, sample SRR1917694 (A) and PRJNA390669, sample

SRR5689230 **(B)**; diagrammatic illustration of long-read sequencing coverage of the same region **(C)**. The two RNA-seq samples differ in their sequencing depths: average, non-zero, sequencing depth for the region displayed is 55.6 for (A) and 312.8 for (B). The blue rectangle below the x-axis indicates the genomic region covered by *DrrS* (MTS1338), an annotated *M.tb* sRNA of 109 nts. This stable (and by far most abundant) form is cleaved from longer transcripts found in Northern Blots of 160-400+ nts (Moores et al., 2017). Both RNA-seq datasets (A & B) display a gradual drop in coverage at the 3' end. In such cases, automatic computational prediction of the correct transcript length is challenging for any algorithm but here the prediction is further complicated by the fact that multiple overlapping transcripts of different length most likely co-exist in the data. Even in the deeply sequenced sample, where the presence of overlapping transcripts could be conjectured, most algorithms would call a single transcript, without additional knowledge of transcription start and termination sites for the refinement of computational predictions. In the absence of such additional data, long-read sequencing might be helpful: as illustrated in diagram (C), long reads whose starts and ends can be unambiguously defined should be helpful in identifying the presence of multiple overlapping transcripts expressed from a single locus. Image created using the Integrated Genome Viewer (Robinson et al., 2011) and BioRender.com.

The more sophisticated among transcriptomics-based approaches use a combination of sources, such as TSS data or conservation across species, to reduce false positives. DETR'PROK is a Galaxy-based workflow, coordinating over 40 publicly-available Galaxy sequence comparison tools into a pipeline which streamlines the number of user-defined parameters. However, there are still 14 different user inputs, most of which concern filtering to account for read depth and transcriptional noise (Toffano-Nioche et al., 2013). The recently published ANNOgesic suite of tools utilises multiple third-party software packages, as well as its own scripts to analyse RNA-seq data and filter predictions. Although, the suite includes an *sRNA-finder* module, using this module in isolation on user-generated alignment files requires specific file formats for the alignment (.wig) and several reference annotation files. Multiple levels of filtering are possible to identify *bona fide* ncRNAs, but such filtering requires downloading of tools and databases such as RNAfold (Denman, 1993), BSRD (Li et al., 2013) and the NCBI nr protein database (NCBI Resource Coordinators, 2014). In the context of validating mycobacteria ncRNA predictions, such databases may possibly be less relevant, given the lack of homology or shared sequence features between mycobacterial and other bacterial ncRNAs. Additionally, fine-tuning cut-off parameters to distinguish signal from noise is ultimately still up to the user. Somewhat surprisingly, the added complexity of such methods does not always translate into more accurate results: in limited comparisons between methods that use additional information and our own simpler, signal-only-based method, we found that our naïve approach performs comparatively well, most likely because more sophisticated methods often require more tuning of their parameters to take advantage of their added complexity (Ozuna et al., 2019). As the responsibility of parameter tuning is left up to the user, it is obvious that methods with fewer parameters, such as Rockhopper,

baerhunter or APERO, may be less error-prone and, ultimately, more appealing, especially to non-computational users looking for quick and easy to implement solutions. Rockhopper is an independent, Java-based tool designed for bacterial RNAseq data (McClure et al., 2013). To eliminate guesswork by the user to adjust for noise vs. signal, the program normalises for read counts using the upper quartile of non-zero gene expression values and generates a transcriptional map of the predicted non-coding elements. *Baerhunter* (Ozuna et al., 2019) and APERO (Leonard et al., 2019) are lighter tools to install, both written in R and requiring only the most commonly used BAM format alignment files and relevant reference annotations. Like Rockhopper, the output of *baerhunter* is a transcriptional map (in .gff format), and can consolidate annotations from multiple samples. APERO exploits improvements in sequencing technology by requiring paired-end reads (where each fragment is sequenced from both ends, creating two barcoded reads for each fragment) and optimising parameters for non-fragmented libraries. The output consists of a set of flat files of the predicted transcript 5' and 3' ends for each sample that can then be filtered for read counts and assembled into a genomic context.

Steps can be taken to lend support to computational predictions of sRNAs and 5' UTRs in mycobacteria. In a recent study to identify differentially expressed, verifiable sRNAs in *M. tuberculosis*, software predictions based on RNA-seq produced over 200 candidate sRNAs (Dejesus et al., 2017), 82 of which were differentially expressed by 6-fold in at least one experimental condition (Gerrick et al., 2018). In an alternative approach, we filtered the predicted sRNAs by comparing the 5' boundaries of these predictions with a compendium of published predicted TSSs (Cortes et al., 2013)(Shell et al., 2015) and narrowed the original list to 67—all of which were also differentially expressed (Appendix 1). Mapping RNase cleavage sites in *M.tuberculosis*, as they were for *M.smegmatis*, could also lend support to the existence of other sRNA candidates cleaved from longer transcripts or otherwise processed (Martini et al., 2019). No matter how comprehensive and complex the prediction algorithm, true validation of candidate ncRNAs requires experimental confirmation such as using RACE (Rapid Amplification of cDNA Ends) (Frohman et al., 1988) to identify transcript boundaries; and Northern blot to confirm the existence and size(s) of actual RNA transcripts, and to confirm expression of orthologous transcripts in related genomes. As computational tools

become more specialised for the mycobacterial genome, laboratory resources can be more confidently directed to their predictions.

A further complication in defining the non-coding transcriptome is that putative non-coding elements predicted by computational algorithms may actually be (or contain) as yet unannotated ORFs; there is no way of asserting from the RNA-seq signal alone whether a transcript is coding or non-coding. Early ribosome profiling studies pointed to the presence of hundreds of small peptides encoded in the 5' UTR of mycobacterial transcripts (Shell et al., 2015), and more recent efforts have shown pervasive translation in *Mtb*, uncovering over 1000 novel ORFs (Smith et al., 2019). The majority of these were short ORFs with non-canonical features that would thus be missed by regular gene prediction algorithms. Although translation of these transcripts does not necessarily render them functional, they may constitute a pool of peptides that are available to use under the right conditions. The observation that leaderless transcripts are translated more efficiently under stress conditions (Sawyer et al., 2021) also points to the fact that mycobacterial non-canonical ORFs may play increasingly important roles in conditions of nutrient starvation or other stresses.

Can we improve the identification of non-coding elements in mycobacteria?

There is limited scope for improving the computational methods used to predict non-coding RNA from the currently available mycobacterial genomic and transcriptomic data. In our experience, both lack of specificity and sensitivity of current methods can be accounted for by the signal (or absence of it) in the raw data. One problem is that in a compact mycobacterial genome, overlapping signal from UTRs and ORFs may confuse algorithms and stop them from correctly predicting the limits of transcripts. In such cases, some level of manual curation is often needed, guided by visualisation on a genome viewer such as Artemis (Carver et al., 2012) or IGV (Robinson et al., 2011). Another source of problems is the use of short reads in the currently most popular sequencing protocol. Typical Illumina RNA-seq fragments are 75-150 base pairs long and are mapped in overlapping segments, preferentially using paired-ends, to infer a longer transcriptional unit. Many genes in mycobacteria, including sRNA and asRNA, are transcribed as polycistronic transcripts, where multiple sequential genes are transcribed into a single mRNA transcript. The individual overlapping transcripts of varying

lengths are often difficult to detect with standard RNA-seq (Figure 1). The development of specialised RNA-seq methods, such as dRNA-seq (Sharma et al., 2010) to enrich for the 5' end of primary transcripts and map TSSs, and Term-seq (D. Dar et al., 2016) to find 3' termini, offers information that can be used to address the issue of overlapping signal from distinct transcripts. Moreover, ribosomal profiling (Ingolia et al., 2009) will continue to be instrumental in resolving ambiguities in annotation of ORFs versus non-coding elements in untranslated regions (Shell et al., 2015), (Smith et al., 2019). Although such information can already be integrated in a subset of computational pipelines (Yu et al., 2018), the corresponding data is only available for a limited number of reference mycobacterial strains.

Perhaps one of the most promising new technologies for studying whole transcriptomes are based on long-read sequencing. Pac-Bio SMRT or Oxford Nanopore Technologies sequencing can achieve reads several thousand nucleotides long, resolving issues associated with errors in the assembly of short reads. The selection of primary RNA transcripts that have not been fragmented in cDNA library preparation make it possible to reconstruct an entire transcriptome with a high level of confidence. In addition, the ability to sequence full polycistronic transcripts allows the surveying of dynamic changes to the structure of bacterial operons in response to a change in the conditions of growth (Yan et al., 2018). Nanopore sequencing goes one step further in making it possible for native RNA molecules to be sequenced, allowing post-transcriptional modifications of individual nucleotides to be detected, that would otherwise be lost during reverse transcription to cDNA (Grünberger et al., 2020). The technologies are still evolving, but with bioinformatics improvements to resolve technical issues of noise from the nanopore and saturation (Soneson et al., 2019), long-read sequencing will become a valuable tool for studying transcriptomes. The longer reads will certainly improve our mapping of 5' and 3' UTRs, as well as our understanding of the dynamic nature of bacterial transcriptional units, but issues with discriminating coding from non-coding elements, sRNAs from UTRs and identifying original versus processed or degraded transcripts will remain a problem.

One, perhaps less obvious, way in which new sequencing technologies may prove instrumental in improving the prediction of non-coding RNAs is their role in improving the assembly of genomic sequences against which RNA-seq reads are mapped. Currently, *Mtb*

transcript mapping relies on the cultured genome, H37Rv, which shows considerable differences compared to clinical and field strains, or isolates, that have adapted to different environmental pressures (Shockey et al., 2019) (O'Neill et al., 2015). As SNPs in promoter regions and small insertions/deletions may play a major role in regulating the expression of non-coding elements (Dinan et al., 2014), it is clear that using the correct genomic sequence is important when analysing transcriptomes of non-reference strains. Sequencing and assembly of potentially thousands more strains are being facilitated by technologies offering portable sequencing platforms and we can expect the number of available mycobacterial genomes to increase manifold in public databases in the next few years, as a result of the increase in popularity of such methods.

Having the correct genomic sequence available is important but correct annotation is arguably just as important, given how many algorithms rely on the annotation of coding elements to make predictions of non-coding ones. Homologous predicted sRNAs are sometimes annotated as protein-coding or non-coding in different genomes, and could, in fact, be dual-function sRNAs (Vanderpool et al., 2011). This is especially obvious when trying to compare non-coding elements or small ORFs in different lineages of *Mtb* (Arnvig & Young, 2012). To improve annotation efforts, the idea of assembling MTBC pangenomes that differentiate core genes (including non-coding ones) present in all lineages, from accessory genes present in a subset and unique genes present in only one strain or lineage, is an appealing one (Vernikos et al., 2015). Although members of the MTBC are assumed to share very high sequence identity, this assumption is rooted primarily on comparisons of reference sequences and less so on circulating strains. For example, the sequencing of *M. bovis* strains that cannot be classified in current clonal complexes, suggested that diversity within this species may be higher than previously thought (Zimpel et al., 2020). Pangenomic projects to date have primarily focussed on identifying differences in antibiotic liability/resistance among clinical strains of *Mtb* (H. A. Dar et al., 2020) (Rufai et al., 2020), but whole genome sequencing projects of clinical and field strains to assemble lineage-specific pangenomes for both human and animal-adapted MTBC members would allow comparisons and provide a more accurate picture of the extent of riboregulation and its effect on host-specificity and other phenotypic differences (Zimpel et al., 2017).

Conclusions

A definitive answer to the question, “How many non-coding RNA elements exist in mycobacterial genomes?”, is not yet possible. Although several computational methods have been developed to support this area of research, our knowledge is currently limited by the availability and quality of raw data. We believe the key to constructing an atlas of the mycobacterial non-coding universe is recognising both the diversity of the individual genomes, and the dynamic nature of the corresponding transcriptomes. Integration of existing and new sequencing technologies and close collaboration between experimental and computational groups should allow us to progress faster towards this goal.

Funding information

J.S. is funded by a Bloomsbury Colleges PhD Studentship (LIDo programme).

Conflicts of interest

The authors declare that there are no conflicts of interest.

Acknowledgement

The mapped read (bam) files used to create panels A and B of figure 1 were obtained from Mr Yen Yi Tan’s MSc dissertation work in the Nobeli group (UCL, 2020). The authors gratefully acknowledge this contribution.

Appendix 1. List of *M. tuberculosis* 'stable regulatory RNAs' created from Mycobrowser (<https://mycobrowser.epfl.ch>) H37Rv annotation file, release 4 (*Mycobacterium_tuberculosis_H37Rv_gff_v4.gff*), and RFAM sRNA entries. RNA-seq predicted ncRNAs were filtered to include only those with either a predicted TSS (Cortes et al., 2013)(Shell et al., 2015) within 10 nts of the start position, or by record of experimental validation. *nc Locus ID refers to annotation as Lamichhane et al., (Lamichhane et al., 2013), and missing locus names were assigned based on mapping coordinates to H37Rv reference sequence (AL12345.5).

Name	nc Locus ID*	Tuberculist	Rfam	Start	End	Strd	TSS <10 nt	Experimental verification
	ncRv10071	MTB000100		80240	80440	+	TSS	
	ncRv10071c	MTB000101		80254	80344	-	TSS	
F6	ncRv10243A	MTB000051	RF01791	293604	293705	+	TSS	Northern/RLM- RACE
	ncRv10243B	MTB000107		293603	293663	+	TSS	
	ncRv10467	MTB000109		558815	558884	+	TSS	
mcr19	ncRv0485	MTB000060		575033	575069	+		Northern
	ncRv0490	MTB000110		579290	579408	+	TSS	
B55	ncRv10609	MTB000052	RF01783	704185	704246	+	TSS	Northern/RLM- RACE
	ncRv10637	MTB000112		733361	733459	+	TSS	
	ncRv10666	MTB000115		759479	759610	+	TSS	
	ncRv0810c	MTB000119		905075	905164	-	TSS	
ASdes	ncRv0824	MTB000053	RF01781	918264	918458	+	TSS	Northern/RLM- RACE
mpr5	ncRv11051	MTB000061		1175225	1175315	+	TSS	Northern
	ncRv1072	MTB000125		1197082	1197179	+	TSS	
MTS0858	ncRv1092c	MTB000074		1220388	1220487	-	TSS	Northern/RLM- RACE
ASdes2	ncRv1094c		RF01781	1222277	1222344	-		Northern/RLM- RACE
mcr10	ncRv1157	MTB000072		1283693	1283815	+		Northern/RLM- RACE
	ncRv11179c	MTB000128		1313343	1313452	-	TSS	
mpr6	ncRv1222	MTB000062		1365274	1365365	+		Northern
mcr11	ncRv11264c	MTB000063	RF02341	1413094	1413224	-	TSS	Northern/RLM- RACE
MTS0930	ncRv11147c	MTB000130		1413105	1413227	-	TSS	Northern

mcr3	ncRv11315A	MTB000064		1471619	1471742	+		RLM-RACE
	ncRv11315B	MTB000133		1473385	1473503	+	TSS	
mcr15	ncRv1364c	MTB000065		1535417	1535716	-		RLM-RACE
MTS1082	ncRv11373	MTB000076		1547129	1547268	+	TSS	Northern
	ncRv1501	MTB000136		1692646	1692731	+	TSS	
G2	ncRv11689c	MTB000054	RF01798	1914962	1915190	-	TSS	Northern/RLM-RACE
DrrS/ MTS1338	ncRv11733A	MTB000077		1960667	1960783	+	TSS	Northern/RLM-RACE
	ncRv11733B	MTB000139		1960667	1960774	+	TSS	
MrsI	ncRv11846	MTB000142		2096766	2096867	+	TSS	
	ncRv12023	MTB000143		2268164	2268231	+	TSS	
ASpks	ncRv2048c	MTB000057	RF01782	2299796	2299873	-		Northern/RLM-RACE
mcr5	ncRv2175c	MTB000066		2437823	2437866	-		Northern
mcr16	ncRv2243c	MTB000073		2517032	2517134	-		Northern
mcr7	ncRv0024	MTB000067	RF02571	2692172	2692521	+	TSS	RLM-RACE
mpr11	ncRv12560	MTB000068		2881252	2881320	+		Northern
mpr12	ncRv12562	MTB000069		2882185	2882276	+		Northern
MTS2048	ncRv12659		RF0259	2980911	2981083	+	TSS	Northern/RLM-RACE
	ncRv12783c	MTB000148		3092761	3092886	-	TSS	
	ncRv2986c	MTB000149		3343113	3343216	-	TSS	
	ncRv2993c	MTB000150		3350950	3351074	-	TSS	
	ncRv13003c	MTB000151		3363029	3363152	-	TSS	
	ncRv13418cA	MTB000154		3837297	3837458	-	TSS	
	ncRv13418cB	MTB000155		3837346	3837458	-	TSS	
mpr17	ncRv13596	MTB000070		4040879	4040938	+	TSS	Northern
mpr18	ncRv13651	MTB000071		4093468	4093522	+		Northern
6CsRNA/B 11	ncRv13660c	MTB000058	RF01066	4099386	4099478	-	TSS	Northern/RLM-RACE
Ms1/MTS 2823	ncRv13661	MTB000078	RF02566	4100669	4100968	+	TSS	Northern/RLM-RACE
C8/4.5S RNA	ncRv13722c	MTB000059		4168154	4168281	-	TSS	Northern/RLM-RACE
	ncRv13722c	MTB000160		4168192	4168281	-	TSS	
	ncRv3804c	MTB000161		4265583	4265765	-	TSS	
MTS2975	ncRv13943	MTB000080		4317073	4317165	+	TSS	Northern

References

- Ami, V. K. G., Balasubramanian, R., & Hegde, S. R. (2020). Genome-wide identification of the context- dependent sRNA expression in *Mycobacterium tuberculosis*. *BMC Genomics*, 21(167), 1–12.
- Arnvig, K. B., Comas, I., Thomson, N. R., Houghton, J., Boshoff, H. I., Croucher, N. J., Rose, G., Perkins, T. T., Parkhill, J., Dougan, G., & Young, D. B. (2011). Sequence-Based Analysis Uncovers an Abundance of Non-Coding RNA in the Total Transcriptome of *Mycobacterium tuberculosis*. *PLOS Pathogens*, 7(11), e1002342.
- Arnvig, K. B., Cortes, T., & Young, D. B. (2014). Noncoding RNA in *Mycobacteria*. *Microbiology Spectrum*, 2(2), 1–16. <https://doi.org/10.1128/microbiolspec>
- Arnvig, K. B., & Young, D. B. (2009). Identification of small RNAs in *Mycobacterium tuberculosis*. *Molecular Microbiology*, 73(3), 397–408.
<https://doi.org/10.1111/j.1365-2958.2009.06777.x>
- Brites, D., Loiseau, C., Menardo, F., Borrell, S., Boniotti, M. B., Warren, R., Dippenaar, A., Parsons, S. D. C., Beisel, C., Behr, M. A., Fyfe, J. A., Coscolla, M., & Gagneux, S. (2018). A new phylogenetic framework for the animal-adapted *mycobacterium tuberculosis* complex. *Frontiers in Microbiology*, 9(NOV), 1–14.
<https://doi.org/10.3389/fmicb.2018.02820>
- Carver, T., Harris, S. R., Berriman, M., Parkhill, J., & McQuillan, J. A. (2012). Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, 28(4), 464–469.
<https://doi.org/10.1093/bioinformatics/btr703>

- Chakravarty, S., & Massé, E. (2019). RNA-Dependent Regulation of Virulence in Pathogenic Bacteria. In *Frontiers in Cellular and Infection Microbiology* (Vol. 9).
<https://www.frontiersin.org/article/10.3389/fcimb.2019.00337>
- Cheng, G., Hussain, T., Sabir, N., Ni, J., Li, M., Zhao, D., & Zhou, X. (2019). Comparative study of the molecular basis of pathogenicity of *M. Bovis* strains in a mouse model. *International Journal of Molecular Sciences*, 20(1).
<https://doi.org/10.3390/ijms20010005>
- Chiner-Oms, Á., Berney, M., Boinett, C., González-Candelas, F., Young, D. B., Gagneux, S., Jacobs, W. R., Parkhill, J., Cortes, T., & Comas, I. (2019). Genome-wide mutational biases fuel transcriptional diversity in the *Mycobacterium tuberculosis* complex. *Nature Communications*, 10(1), 1–11. <https://doi.org/10.1038/s41467-019-11948-6>
- Cortes, T., Schubert, O. T., Rose, G., Arnvig, K. B., Comas, I., Aebbersold, R., & Young, D. B. (2013). Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*. *Cell Reports*, 5(4), 1121–1131. <https://doi.org/10.1016/j.celrep.2013.10.031>
- Dar, D., Shamir, M., Mellin, J. R., Koutero, M., Stern-Ginossar, N., Cossart, P., & Sorek, R. (2016). Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science*, 352(6282), aad9822. <https://doi.org/10.1126/science.aad9822>
- Dar, H. A., Zaheer, T., Ullah, N., Bakhtiar, S. M., Zhang, T., Yasir, M., Azhar, E. I., & Ali, A. (2020). Pangenome Analysis of *Mycobacterium tuberculosis* Reveals Core-Drug Targets and Screening of Promising Lead Compounds for Drug Discovery. In *Antibiotics* (Vol. 9). <https://doi.org/10.3390/antibiotics9110819>
- Dejesus, M. A., Gerrick, E. R., Xu, W., Park, S. W., Long, J. E., Boutte, C. C., Rubin, E. J., Schnappinger, D., Ehrt, S., Fortune, S. M., Sassetti, C. M., & Ioerger, T. R. (2017).

Comprehensive essentiality analysis of the *Mycobacterium tuberculosis* genome via saturating transposon mutagenesis. *MBio*, 8(1), 1–17.

<https://doi.org/10.1128/mBio.02133-16>

Denman RB. (1993). Using RNAFOLD to predict the activity of small catalytic RNAs.

Biotechniques, 15(6), 1090–1095.

DiChiara, J. M., Contreras-Martinez, L. M., Livny, J., Smith, D., McDonough, K. A., & Belfort,

M. (2010). Multiple small RNAs identified in *Mycobacterium bovis* BCG are also expressed in *Mycobacterium tuberculosis* and *Mycobacterium smegmatis*. *Nucleic*

Acids Research, 38(12), 4067–4078. <https://doi.org/10.1093/nar/gkq101>

Dinan, A. M., Tong, P., Lohan, A. J., Conlon, K. M., Miranda-casoluengo, A. A., & Malone, K.

M. (2014). *Relaxed Selection Drives a Noisy Noncoding Transcriptome in Members of the*. 5(4), 1–9. <https://doi.org/10.1128/mBio.01169-14>. Editor

Frohman, M. A., Dush, M. K., & Martin, G. R. (1988). Rapid production of full-length cDNAs

from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. *Proceedings of the National Academy of Sciences of the United States of America*, 85(23), 8998–9002. PubMed. <https://doi.org/10.1073/pnas.85.23.8998>

Gerrick, E. R. (2018). *Discovery of Small RNAs and Characterization of Their Regulatory Roles*

in Mycobacterium Tuberculosis [PhD Thesis, Harvard University, Graduate School of Arts & Sciences]. <https://dash.harvard.edu/handle/1/41129159>

Gerrick, E. R., Barbier, T., Chase, M. R., Xu, R., François, J., Lin, V. H., Szucs, M. J., Rock, J. M.,

Ahmad, R., Tjaden, B., Livny, J., & Fortune, S. M. (2018). Small RNA profiling in *mycobacterium tuberculosis* identifies mrsi as necessary for an anticipatory iron sparing response. *Proceedings of the National Academy of Sciences of the United States of America*, 115(25), 6464–6469. <https://doi.org/10.1073/pnas.1718003115>

- 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>
- Gómez-Lozano, M., Marvig, R., Molin, S., & Long, K. (2014). Identification of Bacterial Small RNAs by RNA Sequencing. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 1149, pp. 433–456). https://doi.org/10.1007/978-1-4939-0473-0_34
- Grünberger, F., Knüppel, R., Jüttner, M., Fenk, M., Borst, A., Reichelt, R., Hausner, W., Soppa, J., Ferreira-Cerca, S., & Grohmann, D. (2020). Exploring prokaryotic transcription, operon structures, rRNA maturation and modifications using Nanopore-based native RNA sequencing. *BioRxiv*, 2019.12.18.880849. <https://doi.org/10.1101/2019.12.18.880849>
- Houghton, J., Rodgers, A., Rose, G., & Arnvig, K. B. (2020). The *Mycobacterium tuberculosis* sRNA F6 regulates expression of *groEL*/S. *BioRxiv*, 2020.07.15.204107. <https://doi.org/10.1101/2020.07.15.204107>
- Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, N.Y.)*, 324(5924), 218–223. PubMed. <https://doi.org/10.1126/science.1168978>
- Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., Rivas, E., Eddy, S. R., Finn, R. D., Bateman, A., & Petrov, A. I. (2021). Rfam 14: Expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, 49(D1), D192–D200. <https://doi.org/10.1093/nar/gkaa1047>

- Kapopoulou, A., Lew, J. M., & Cole, S. T. (2011). The MycoBrowser portal: A comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis*, 91(1), 8–13. <https://doi.org/10.1016/j.tube.2010.09.006>
- Kumar, K., Chakraborty, A., & Chakrabarti, S. (2020). PresRAT: A server for identification of bacterial small-RNA sequences and their targets with probable binding region. *RNA Biology*, 2020.04.03.024935. <https://doi.org/10.1101/2020.04.03.024935>
- Lamichhane, G., Arnvig, K. B., & McDonough, K. A. (2013). Definition and annotation of (myco)bacterial non-coding RNA. *Tuberculosis*, 93(1), 26–29. <https://doi.org/10.1016/j.tube.2012.11.010>
- Leonard, S., Meyer, S., Lacour, S., Nasser, W., Hommais, F., & Reverchon, S. (2019). APERO: a genome-wide approach for identifying bacterial small RNAs from RNA-Seq data. *Nucleic Acids Research*, 47(15), e88–e88. <https://doi.org/10.1093/nar/gkz485>
- Li, L., Huang, D., Cheung, M. K., Nong, W., Huang, Q., & Kwan, H. S. (2013). BSRD: a repository for bacterial small regulatory RNA. *Nucleic Acids Research*, 41(Database issue), D233–D238. PubMed. <https://doi.org/10.1093/nar/gks1264>
- Livny, J., Teonadi, H., Livny, M., & Waldor, M. K. (2008). High-Throughput, Kingdom-Wide Prediction and Annotation of Bacterial Non-Coding RNAs. *PLOS ONE*, 3(9), e3197.
- Mai, J., Rao, C., Watt, J., Sun, X., Lin, C., Zhang, L., & Liu, J. (2019). Mycobacterium tuberculosis 6C sRNA binds multiple mRNA targets via C-rich loops independent of RNA chaperones. *Nucleic Acids Research*, 47(8), 4292–4307. <https://doi.org/10.1093/nar/gkz149>
- Malone, K. M., Rue-Albrecht, K., Magee, D. A., Conlon, K., Schubert, O. T., Nalpas, N. C., Browne, J. A., Smyth, A., Gormley, E., Aebersold, R., MacHugh, D. E., & Gordon, S. V. (2018). Comparative 'omics analyses differentiate Mycobacterium tuberculosis and

Mycobacterium bovis and reveal distinct macrophage responses to infection with the human and bovine tubercle bacilli. *Microbial Genomics*, 4(3).

<https://doi.org/10.1099/mgen.0.000163>

Mao, X., Ma, Q., Liu, B., Chen, X., Zhang, H., & Xu, Y. (2015). Revisiting operons: An analysis of the landscape of transcriptional units in E. coli. *BMC Bioinformatics*, 16(1), 356.

<https://doi.org/10.1186/s12859-015-0805-8>

Martini, M. C., Zhou, Y., Sun, H., & Shell, S. S. (2019). Defining the Transcriptional and Post-transcriptional Landscapes of Mycobacterium smegmatis in Aerobic Growth and Hypoxia. In *Frontiers in Microbiology* (Vol. 10).

<https://www.frontiersin.org/article/10.3389/fmicb.2019.00591>

McClure, R., Balasubramanian, D., Sun, Y., Bobrovskyy, M., Sumby, P., Genco, C. A., Vanderpool, C. K., & Tjaden, B. (2013). Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Research*, 41(14), e140–e140.

<https://doi.org/10.1093/nar/gkt444>

Miotto, P., Forti, F., Ambrosi, A., Pellin, D., Veiga, D. F., Balazsi, G., Gennaro, M. L., Di Serio, C., Ghisotti, D., & Cirillo, D. M. (2012). Genome-Wide Discovery of Small RNAs in Mycobacterium tuberculosis. *PLOS ONE*, 7(12), e51950.

Moore, A., Riesco, A. B., Schwenk, S., & Arnvig, K. B. (2017). Expression, maturation and turnover of DrrS, an unusually stable, DosR regulated small RNA in Mycobacterium tuberculosis. *PLOS ONE*, 12(3), e0174079.

NCBI Resource Coordinators. (2014). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 42(D1), D7–D17.

<https://doi.org/10.1093/nar/gkt1146>

- Newton-Foot, M., & Gey van Pittius, N. C. (2013). The complex architecture of mycobacterial promoters. *Tuberculosis*, 93(1), 60–74. <https://doi.org/10.1016/j.tube.2012.08.003>
- O'Neill, M. B., Mortimer, T. D., & Pepperell, C. S. (2015). Diversity of Mycobacterium tuberculosis across Evolutionary Scales. *PLOS Pathogens*, 11(11), e1005257.
- Ozuna, A., Liberto, D., Joyce, R. M., Arnvig, K. B., & Nobeli, I. (2019). baerhunter: An R package for the discovery and analysis of expressed non-coding regions in bacterial RNA-seq data. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btz643>
- Pelly, S., Bishai, W. R., & Lamichhane, G. (2012). A screen for non-coding RNA in Mycobacterium tuberculosis reveals a cAMP-responsive RNA that is expressed during infection. *Gene*, 500(1), 85–92. <https://doi.org/10.1016/j.gene.2012.03.044>
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1), 24–26. <https://doi.org/10.1038/nbt.1754>
- Rufai, S. B., Ozer, E. A., & Singh, S. (2020). Pan-genome analysis of *Mycobacterium tuberculosis* identifies accessory genome sequences deleted in modern Beijing lineage. *BioRxiv*, 2020.12.01.407569. <https://doi.org/10.1101/2020.12.01.407569>
- Sawyer, E. B., Phelan, J. E., Clark, T. G., & Cortes, T. (2021). A snapshot of translation in Mycobacterium tuberculosis during exponential growth and nutrient starvation revealed by ribosome profiling. *Cell Reports*, 34(5). <https://doi.org/10.1016/j.celrep.2021.108695>
- Schwenk, S., & Arnvig, K. B. (2018). Regulatory RNA in Mycobacterium tuberculosis, back to basics. *Pathogens and Disease*, 76(4). <https://doi.org/10.1093/femspd/fty035>
- Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findei, S., Sittka, A., Chabas, S., Reiche, K., Hackermller, J., Reinhardt, R., Stadler, P. F., & Vogel, J. (2010). The

- primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, 464(7286), 250–255. <https://doi.org/10.1038/nature08756>
- Shell, S. S., Wang, J., Lapierre, P., Mir, M., Chase, M. R., Pyle, M. M., Gawande, R., Ahmad, R., Sarracino, D. A., Ioerger, T. R., Fortune, S. M., Derbyshire, K. M., Wade, J. T., & Gray, T. A. (2015). Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLOS Genetics*, 11(11), e1005641.
- Shockey, A. C., Dabney, J., & Pepperell, C. S. (2019). Effects of Host, Sample, and in vitro Culture on Genomic Diversity of Pathogenic Mycobacteria. In *Frontiers in Genetics* (Vol. 10). <https://www.frontiersin.org/article/10.3389/fgene.2019.00477>
- Šiková, M., Janoušková, M., Ramaniuk, O., Páleníková, P., Pospíšil, J., Bartl, P., Suder, A., Pajer, P., Kubičková, P., Pavliš, O., Hradilová, M., Vítovská, D., Šanderová, H., Převorovský, M., Hnilicová, J., & Krásný, L. (2019). Ms1 RNA increases the amount of RNA polymerase in *Mycobacterium smegmatis*. *Molecular Microbiology*, 111(2), 354–372. <https://doi.org/10.1111/mmi.14159>
- Smith, C., Canestrari, J., Wang, J., Derbyshire, K., Gray, T., & Wade, J. (2019). Pervasive Translation in *Mycobacterium tuberculosis*. *BioRxiv*, 665208. <https://doi.org/10.1101/665208>
- Solans, L., Gonzalo-Asensio, J., Sala, C., Benjak, A., Uplekar, S., Rougemont, J., Guilhot, C., Malaga, W., Martín, C., & Cole, S. T. (2014). The PhoP-Dependent ncRNA Mcr7 Modulates the TAT Secretion System in *Mycobacterium tuberculosis*. *PLOS Pathogens*, 10(5), e1004183.
- Soneson, C., Yao, Y., Bratus-Neuenschwander, A., Patrignani, A., Robinson, M. D., & Hussain, S. (2019). A comprehensive examination of Nanopore native RNA sequencing for

characterization of complex transcriptomes. *Nature Communications*, 10(1), 3359.

<https://doi.org/10.1038/s41467-019-11272-z>

Sridhar, J., Narmada, S. R., Sabarinathan, R., Ou, H. Y., Deng, Z., Sekar, K., Rafi, Z. A., &

Rajakumar, K. (2010). sRNAsScanner: A computational tool for intergenic small RNA detection in bacterial genomes. *PLoS ONE*, 5(8).

<https://doi.org/10.1371/journal.pone.0011970>

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential

expression in RNA-seq: A matter of depth. *Genome Research*, 21(12), 2213–2223.

<https://doi.org/10.1101/gr.124321.111>

Toffano-Nioche, C., Luo, Y., Kuchly, C., Wallon, C., Steinbach, D., Zytnecki, M., Jacq, A., &

Gautheret, D. (2013). Detection of non-coding RNA in bacteria and archaea using the DETR'PROK Galaxy pipeline. *Methods*, 63(1), 60–65.

<https://doi.org/10.1016/j.ymeth.2013.06.003>

Vanderpool, C. K., Balasubramanian, D., & Lloyd, C. R. (2011). Dual-function RNA regulators

in bacteria. *Biochimie*, 93(11), 1943–1949.

<https://doi.org/10.1016/j.biochi.2011.07.016>

Vernikos, G., Medini, D., Riley, D. R., & Tettelin, H. (2015). Ten years of pan-genome

analyses. *Current Opinion in Microbiology*, 23, 148–154.

<https://doi.org/10.1016/j.mib.2014.11.016>

Wang, M., Fleming, J., Li, Z., Li, C., Zhang, H., Xue, Y., Chen, M., Zhang, Z., Zhang, X. E., & Bi,

L. (2016). An automated approach for global identification of sRNA-encoding regions in RNA-Seq data from *Mycobacterium tuberculosis*. *Acta Biochimica et Biophysica Sinica*, 48(6), 544–553. <https://doi.org/10.1093/abbs/gmw037>

Yan, B., Boitano, M., Clark, T. A., & Ettwiller, L. (2018). SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nature Communications*, 9(1), 3676.

<https://doi.org/10.1038/s41467-018-05997-6>

Yu, S.-H., Vogel, J., & Förstner, K. U. (2018). ANNOgesic: A Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes. *GigaScience*, 7(9).

<https://doi.org/10.1093/gigascience/giy096>

Zimpel, C. K., Brandão, P. E., de Souza Filho, A. F., de Souza, R. F., Ikuta, C. Y., Neto, J. S. F., Soler Camargo, N. C., Heinemann, M. B., & Guimarães, A. M. S. (2017). Complete genome sequencing of *Mycobacterium bovis* SP38 and comparative genomics of *Mycobacterium bovis* and *M. tuberculosis* strains. *Frontiers in Microbiology*, 8(DEC), 1–14. <https://doi.org/10.3389/fmicb.2017.02389>

Zimpel, C. K., Patané, J. S. L., Guedes, A. C. P., de Souza, R. F., Silva-Pereira, T. T., Camargo, N. C. S., de Souza Filho, A. F., Ikuta, C. Y., Neto, J. S. F., Setubal, J. C., Heinemann, M. B., & Guimaraes, A. M. S. (2020). Global Distribution and Evolution of *Mycobacterium bovis* Lineages. In *Frontiers in Microbiology* (Vol. 11). <https://www.frontiersin.org/article/10.3389/fmicb.2020.00843>