Noncoding RNA in Mycobacteria

KRISTINE B. ARNVIG, TERESA CORTES, and DOUGLAS B. YOUNG
National Institute for Medical Research, Mycobacterial Research Division, London, NW7 1AA, United Kingdom

ABSTRACT Efforts to understand the molecular basis of mycobacterial gene regulation are dominated by a protein-centric view. However, there is a growing appreciation that noncoding RNA, i.e., RNA that is not translated, plays a role in a wide variety of molecular mechanisms. Noncoding RNA comprises rRNA, tRNA, 4.5S RNA, RnpB, and transfer-messenger RNA, as well as a vast population of regulatory RNA, often dubbed “the dark matter of gene regulation.” The regulatory RNA species comprise 5′ and 3′ untranslated regions and a rapidly expanding category of transcripts with the ability to base-pair with mRNAs or to interact with proteins. Regulatory RNA plays a central role in the bacterium’s response to changes in the environment, and in this article we review emerging information on the presence and abundance of different types of noncoding RNA in mycobacteria.

In the 15 years since the publication of the genome sequence of Mycobacterium tuberculosis, analysis of transcriptional regulation has provided a dominant framework for rendering genomic information into understanding of the functional biology of the organism. The powerful combination of molecular biology tools for characterization of individual genetic loci alongside the genome-wide perspective provided by microarray analysis has led to the use of transcription profiles as a surrogate for key phenotypic states implicated in pathogenesis, immunogenicity, and response to drug treatment. As technologies move toward definition of phenotypes by more direct measurement of metabolic status and high-resolution ultrastructural imaging, it is important to develop a rigorous understanding of the quantitative relationship between RNA transcript abundance and broader aspects of cellular physiology, both at the level of bulk populations and at the level of individual cells. The growing awareness—driven largely by the application of high-throughput sequencing technologies to the analysis of RNA (RNA-seq)—that bacteria transcribe much more RNA than is required for direct translation into proteins is likely to be important in this context.

Bacteria were largely excluded from debates surrounding the presence of “junk DNA” in eukaryotic genomes; prokaryotic coding sequences are tightly packed with apparently minimal intergenic space. It is now clear that junk DNA of eukaryotes is in fact transcribed into a vast repertoire of noncoding RNAs (ncRNAs), many of which play profound and pervasive roles in the regulation of gene expression. However, bacteria have become active participants as the debate moves from “junk DNA” to “junk RNA.” Do these novel ncRNAs in bacteria reflect some form of transcriptional “noise,” or do they provide a crucial layer of posttranscriptional regulation with a central—and perhaps manipulable—function in determining the amount of protein that is produced from an individual mRNA transcript? While “both of the above” seems intuitively likely, answering this question will require cataloguing the total noncoding transcriptome together with targeted functional studies. It is clear that computational techniques will be essential for integration of the rapidly accumulating mass of relevant genome-wide datasets.

In describing experimental and computational approaches applied to M. tuberculosis and other bacteria, this article summarizes the initial attempts to elucidate the role of ncRNA in mycobacteria.

Received: 17 August 2013, Accepted: 25 September 2013, Published: 14 March 2014
Editors: Graham F. Hatfull, University of Pittsburgh, Pittsburgh, PA, and William R. Jacobs, Jr., Howard Hughes Medical Institute, Albert Einstein College of Medicine, Bronx, NY
Correspondence: KB Arnvig, karnvig@nimr.mrc.ac.uk © 2014 American Society for Microbiology. All rights reserved.
WHAT IS ncRNA?

ncRNA refers to any transcript that does not code for a protein or peptide. In bacteria this includes a group of well-studied stable RNAs with defined cellular functions, i.e., rRNA (ribosomal RNA), tRNA (transfer RNA), 4.5S RNA, RnpB, and transfer-messenger RNA (tmRNA, although a small part of this actually is coding). Distinct from this group are the regulatory RNAs, which include 5′ and 3′ untranslated regions (UTRs) linked to coding sequences, and a rapidly expanding category of cis- or trans-encoded transcripts with an ability to base-pair with mRNAs or to interact with proteins (reviewed in, e.g., references 1, 2). Members of this last category are often loosely referred to as small RNAs (sRNAs). Bacterial sRNAs are typically in the range of 50 to 500 nucleotides and hence are comparable to eukaryotic long ncRNAs. There are a few reports of smaller transcripts, i.e., equivalent to eukaryotic micro RNAs, but it remains to be determined if these have any functional significance (3, 4).

It is important to note that the terms “noncoding” and “regulatory” are not interchangeable, although there are many overlaps (Fig. 1). For example, sRNAs can be coding or noncoding and regulatory or not. Some sRNAs have been found to encode small peptides as well as regulate on an RNA level; the best-characterized examples are the SgrS RNA (227 nucleotides) found in enteric bacteria and the unusually large (514 nucleotide) RNAIII found in Staphylococcus aureus (reviewed by Vanderpool et al. [5]).

Finally, the definitions of what is regulatory and what is not may be ambiguous. An RNA such as tmRNA is generally considered to be a housekeeping RNA (6) but could be regarded as a stress-induced regulator of translation.

In this article we focus on the potentially regulatory RNA species, i.e., leader regions, UTRs, and antisense and intergenic sRNAs.

cRNA in M. tuberculosis

The experimental identification of noncoding, potentially regulatory RNA in mycobacteria is so far limited to just over half a dozen studies, mostly focused on M. tuberculosis (7–14). The annotation of newly identified M. tuberculosis ncRNAs was inconsistent until it was addressed recently (15). In this article we use this new annotation, along with the nomenclature used in the primary publications where required for clarification. Deep sequencing of M. tuberculosis RNA (RNA-seq) demonstrates that during exponential growth about a third of nonribosomal RNA in M. tuberculosis is noncoding and hence potentially regulatory. The proportion of ncRNA varies with the growth phase, increasing to more than half in stationary phase and during starvation (7, 16). The quantitative increase in ncRNA during stationary phase is largely due to accumulation of a single, highly abundant sRNA transcript, ncRv13661/MTS2823 (see below). Genome browsers such as Artemis facilitate the visualization of high-throughput RNA-seq data and represent a powerful way of conveying the complexity and flexibility of transcriptomic data (17).

There are significant annotation issues associated with the assignment of coding RNAs versus ncRNAs in M. tuberculosis. Short open reading frames (ORFs) are more frequently annotated as coding sequences in M. tuberculosis CDC1551 (18, 19) than in M. tuberculosis H37Rv (20). For example, a transcript mapping to the intergenic region (IGR) between Rv2395 and Rv2396 in H37Rv has been described as an sRNA (mcr7) (9) but as hypothetical proteins MT2466 and MT2467 in the CDC1551 genome, and was subsequently reported to encode two small acid-inducible proteins (21).

**FIGURE 1** Venn diagram illustrating how the (nomenclature of) different types of ncRNA and regulatory RNAs overlap and, in particular, how sRNAs can be assigned to more than one category. The sRNA subcategories shown are sRNA² (purely regulatory function), sRNA⁴ (dual function, i.e., regulatory potential as well as encoding small peptide), sRNA⁶ (purely coding, i.e., no function as ribo-regulator). Thus, sRNAs can be coding or noncoding, regulatory or not regulatory. Figure modified from reference 15. doi:10.1128/microbiolspec.MGM2-0029-2013.f1
Conversely, the region between Rv3661 and Rv3662c has been annotated as encoding the hypothetical protein MT3762 on the minus strand (19) and alternatively as sRNA ncRv13661/MTS2823 (predicted as mpr4) on the plus strand (7, 9). The precise location of translation start sites determines assignment of coding versus non-coding UTR sequences and is subject to ongoing revision in M. tuberculosis (22). Application of mass spectrometry technology for proteome analysis (e.g., references 23–25) will play an important role in resolving many ambiguities.

**cis-Acting Regulatory Elements**

**cis**-Regulatory RNAs (for definitions, see Glossary) include sequences flanking either end of coding transcripts, referred to as 5′ and 3′ UTRs, or those flanking non-coding transcripts, such as tRNAs and rRNAs. These make an important contribution to the total complement of ncRNA and often play important roles in the expression, processing, and stability of their cognate effector transcript.

**The 3′ End**

The 3′ UTR often includes the site of transcription termination, thereby providing an opportunity for regulation by modification of transcriptional read-through. Moreover, a number of studies from *Escherichia coli* suggest that canonical or L-shaped intrinsic terminators (i.e., a stem-loop followed by a stretch of U residues [26]) may serve as polyadenylation signals. Unlike the situation in eukaryotes, polyadenylated 3′ UTRs in bacteria tend to facilitate RNA turnover rather than stabilize the transcripts (reviewed by Mohanty and Kushner [27]). So far, the evidence for polyadenylation in mycobacteria is very limited (28), and we do not know how this may affect RNA stability, but it is likely that the same general rules apply as in other bacteria.

A subset of *M. tuberculosis* mRNAs have well-defined 3′ ends associated with canonical intrinsic terminators or with the mycobacterium-specific terminator, TRIT (tuberculosis rho-independent terminator), which contains a highly conserved sequence but no poly-U stretch (Fig. 2) and which is present in 147 copies in *M. tuberculosis* (29). It remains to be determined if TRIT termination is dependent on additional factors, but the high degree of sequence conservations suggests that this may be the case. The paucity of canonical terminator structures complicates bioinformatic approaches to the prediction of 3′ ends in *M. tuberculosis* and is further reflected in the fact that a significant proportion of the 3′ ends observed in *M. tuberculosis* RNAs extend over more than 50 nucleotides (7). These long 3′ UTRs are particularly intriguing in the context of convergent genes, where they may provide an antisense transcript with perfect complementarity to the adjacent gene. As in the case of the internal antisense transcripts described below, understanding of the functional consequence of overlapping 3′ UTRs will require further experimental analysis, but it is intriguing that the set of genes subject to overlapping 3′ ends is relatively small.

**FIGURE 2** Transcription termination in mycobacteria. The top panel illustrates the consensus sequence and structure of the mycobacterial terminator, TRIT (tuberculosis rho-independent terminator). TRIT is a novel rho-independent terminator with high sequence conservation identified in and specific for mycobacteria (29). The bottom panel illustrates the expression of two converging genes in *M. tuberculosis*, according to RNA-seq and visualized in the Artemis genome browser; blue represents expression from the forward strand (rplA), and red represents expression from the reverse strand (mmaA4); the height of the trace represents the normalized expression level (reads) over the entire region. The traces demonstrate the termination efficiency exerted by TRIT between the two converging genes. doi:10.1128/microbiolspec.MGM2-0029-2013.f2
to this form of antisense coverage shows a nonrandom distribution that is skewed toward genes with a shared role in cell wall processes (7).

It has also been demonstrated in both *M. tuberculosis* (ncRv10243/F6/MTS194) and *Salmonella* (DapZ) that the 3′ UTRs of some mRNAs have the potential to function independently as trans-acting sRNAs (8, 30). Considering the phylogenetic distance between the two species, these are likely to be examples of a widespread phenomenon.

**The 5′ End**

In contrast to the 3′ UTR, the role of the 5′ UTR has been extensively characterized in bacteria. A prerequisite for the correct identification of a 5′ leader/UTR is a well-defined transcriptional start site (TSS). A number of methods have been applied to identify primary 5′ ends of transcripts such as RNA-ligase-mediated RACE (RLM-RACE) (31) and, more globally, exploitation of the resistance of 5′ triphosphorylated transcripts to digestion by Terminator 5′-phosphate-dependent exonuclease prior to deep sequencing (32). The 5′ UTR represents a hub for regulation; apart from the ribosome-binding site, which in itself is highly regulated by its intrinsic sequence and hence affinity for ribosomes, the 5′ UTR typically provides binding sites for regulatory proteins, metabolites, or sRNAs that modulate transcription or translation of the downstream sequence. Many 5′ UTRs have an inherent ability to alter their conformation in response to external stimuli such as changes in temperature (33), pH (34), or certain metabolites, in which case the RNA element is defined as a riboswitch (35). The changes in conformation lead in turn to changes in the expression of the downstream gene either by blocking/unblocking of the ribosome-binding site or by transcriptional termination/antitermination (36).

Protein-responsive 5′ UTRs are often associated with operons encoding proteins involved in transcription and translation. This type of element provides negative feedback regulation by free proteins encoded in the operon when these are in excess; this form of regulation is well characterized in the case of ribosomal proteins (reviewed by Lindahl and Zengel [37]). The organization of these operons is highly conserved across bacterial species, and the presence of analogous 5′ UTRs associated with corresponding genes and operons in *M. tuberculosis* suggests that similar circuits play an important role in posttranscriptional regulation (7).

A functionally related example can be seen in the 5′ leaders of rRNA transcripts. The 5′ leader (annotated as mpr7/mcr3 in reference 9) of the *M. tuberculosis* rRNA transcript contains a number of well-characterized and highly conserved regulatory elements (38, 39). In *E. coli* this region has been shown to be essential for efficient and balanced expression of the individual rRNA transcripts due to a modification of the transcription elongation complex to a so-called antitermination (AT) complex that has a higher elongation rate and which can read through rho-dependent terminators (40–42). The modification is induced by the binding of a number of AT factors to a specific region of the rRNA leader, the AT site. The sequences and proteins involved are highly conserved and are believed to function in all bacteria (39). The *M. tuberculosis* AT site has been shown to interact specifically with the NusA transcription factor, leading to an increase in RNA polymerase processivity (38, 43). Moreover, the NusA binding site overlaps with an RNase III processing half-site, which has led to the suggestion that NusA also functions as an RNA chaperone for the correct formation of the RNA duplex recognized by RNase III and hence the correct folding of the rRNA (43, 44).

**Riboswitches**

Riboswitches can regulate transcription or translation, and they are classified as “ON” or “OFF” switches depending on the outcome of ligand binding; so far, the OFF switch appears to be the most prevalent type (45). A number of highly conserved riboswitches have been identified in *M. tuberculosis* by sequence homology and covariance models (46). Genes involved in the biosynthesis of methionine provide two examples (Fig. 3). The MetC enzyme catalyzes the conversion of O-acetylhomoserine to t-methionine. In *M. tuberculosis* a SAM-IV riboswitch is located in the 5′ UTR of the metC mRNA (47, 48). The SAM-IV riboswitch represses expression of the downstream gene upon binding of S-adenosyl-methionine (i.e., an OFF switch), which is formed by the addition of ATP to methionine (catalyzed by MetK) and which is required as a cofactor for numerous methyltransferases in *M. tuberculosis* (48). The SAM-IV riboswitch is closely related to a family of seven predicted SAM-I riboswitches in *Listeria monocytogenes*. For two of these, the attenuated riboswitch moiety has been shown to act in trans as an sRNA, thereby adding further complexity to the regulation by (and the distinction between) riboswitches and sRNAs (49). Whether this mechanism applies to additional riboswitches or other attenuated transcripts in general, including those in *M. tuberculosis*, remains to be determined.

In addition to the MetC enzyme, *M. tuberculosis* encodes two isoforms of methionine synthase, MetE and
MetH. Both catalyze the conversion of homocysteine to 1-methionine. However, while the MetH enzyme requires vitamin B$_{12}$ (cobalamin) to function, the MetE enzyme does not. Instead, the metE mRNA harbors a B$_{12}$ riboswitch, which is found in two copies in M. tuberculosis. This riboswitch represses the expression of the downstream genes in the presence of B$_{12}$, presumably by occluding the ribosome-binding site, which in turn leads to destabilization of the transcript (50–52). In other words, while the MetE enzyme is only expressed when B$_{12}$ availability is low, it can also function in the absence of this cofactor, whereas the MetH enzyme requires B$_{12}$ to function and is encoded by a leaderless mRNA (16), meaning that M. tuberculosis employs at least three distinct adaptations of ribo-regulation for methionine biosynthesis.

In addition to the metE riboswitch, a second B$_{12}$ riboswitch is located upstream of an operon comprising PPE2, cobQ1, and cobU (50–52). PPE2 belongs to the family of proteins sharing proline-glutamate (PE) or proline-proline-glutamate (PPE) N-terminal motifs that were identified from the M. tuberculosis genome.
sequence and for the most part lack known biological function (reviewed in Akhter et al. [53]). However, in agreement with previous predictions (50, 51), the PPE2 gene was recently shown to encode an ABC-transporter responsible for B12 uptake in *M. tuberculosis* (54).

Another riboswitch that is also represented twice in the genome of *M. tuberculosis* is the ykok leader or Mbox. This element is often found in the context of magnesium transporters, and like the cobalamin riboswitch, it represses the expression of downstream genes upon binding of its ligand (in this case Mg2+) (55). However, unlike the B12 riboswitch, the *B. subtilis* Mbox has been shown to repress gene expression by translational rather than transcriptional attenuation upon Mg2+ binding, downregulating Mg2+ transport and ensuring homeostasis (56). In *M. tuberculosis* one highly expressed Mbox is found upstream of Rv1535, a conserved hypothetical with some homology to nucleoid-associated proteins, which is induced during Mg2+-starvation (7, 57). Rv1535 is cotranscribed with a downstream Tbox (a tRNA-responsive element [58]) that forms the 5′ UTR of the essential isoleucine tRNA synthase mRNA (IleS, Rv1536) (7). The second Mbox is located upstream of PE20, followed by a series of genes that are all induced by magnesium starvation (57) (Fig. 4). However, qRT-PCR and TSS mapping indicate that only PE20 and PPE31 are cotranscribed with the Mbox (16; K. B. Arnvig, unpublished). Transcription of PPE32 and PPE33 initiates approximately 26 base pairs upstream of the PPE32 ORF, while expression of Rv1810 and the virulence factor MgtC are directed by individual promoters; the transcription start for Rv1810 maps to an AUG codon 12 codons into the annotated ORF, suggesting that this may be the correct (leaderless) start (16). This suggests that not only is the observed magnesium responsiveness of genes downstream of PPE31 an indirect effect that does not directly involve the Mbox, but also that the two Mboxes in *M. tuberculosis* regulate genes of unknown function (Fig. 4).

A final example is the YdaO riboswitch, which is located upstream of *rpfA* (Rv0867c) in *M. tuberculosis*. The *rpfA* gene product is a member of a family of “resuscitation-promoting factors” that share structural homology with lysozyme and may play a particularly important role in reinitiation of cell division after periods of nonreplication (59–61). The YdaO riboswitch is widespread in Gram-positive bacteria and is most often found in the context of genes associated with cell wall metabolism and remodeling (62, 63). However, the ligand for this regulatory element remained elusive until

---

**FIGURE 4** Magnesium-sensing riboswitches in *M. tuberculosis*. The figure illustrates the genomic context of the two identified Mboxes (magnesium-sensing riboswitches) in *M. tuberculosis*. Genes shown in green are conserved hypotheticals, those in orange are information pathways, gray represents PE-PPE genes, and genes in blue are cell wall associated. Black arrows indicate relevant transcription start sites. doi:10.1128/MicrobiolSpectrum.MGM2-0029-2013.f4
recently, when it was demonstrated that it provides an OFF switch responsive to, and highly selective for, the second messenger cyclic di-AMP (63).

These examples demonstrate an extensive and significant role for riboswitches in regulating *M. tuberculosis* gene expression and growth. It is likely that there are many more as yet undiscovered riboswitches in *M. tuberculosis*; however, due to lack of conservation these will be more difficult to identify (64). Inspection of *M. tuberculosis* RNA-seq profiles identifies numerous truncated/attenuated 5′ transcripts that may be considered candidate riboswitches (7, 16). Care has to be exercised in distinguishing truncated 5′ UTRs from stand-alone intergenic transcripts (sRNAs), a distinction that is not necessarily trivial.

**trans-ACTING RNAs**

*trans*-Acting sRNAs (i.e., stand-alone transcripts) include a wealth of different species that either bind to proteins or base-pair with mRNA; the latter category can be further divided into *cis*- and *trans*-encoded (corresponding to antisense or intergenic location). The majority of *trans*-acting RNAs regulate gene expression by base-pairing to one or more target mRNAs, mostly around the translation start site, competing for ribosome binding, and uncoupling transcription from translation, thereby modifying mRNA stability (reviewed in references 6, 65). Protein-binding RNAs in other bacteria include CsrB, CsrC, and 6S RNA. CsrB and CsrC both bind to the carbon storage regulatory protein, CsrA, reviewed by Babitzke and Romeo (66), while 6S RNA binds specifically to the σ^70^-RNA polymerase holoenzyme, thereby downregulating a large number of housekeeping genes in stationary phase (e.g., 67, 68). To date, there are no known homologues of either 6S RNA or the CsrBC systems in mycobacteria.

**cis-ENCODED BASE-PAIRING RNAs**

cis-Encoded RNAs are encoded opposite their target RNAs; hence, they have perfect target complementarity forming duplexes between mRNAs and cognate antisense partners representing ideal substrates for RNase III. The interaction between antisense RNAs (asRNAs) and their sense targets does not require the RNA chaperone Hfq (69) (discussed below). asRNAs were initially discovered in the context of plasmid copy number control and foreign DNA elements (69), but the steadily increasing amount of data from studies using tiling arrays, RNA-seq, and TSS mapping has revealed an unexpected extent and heterologous nature of antisense transcriptomes. asRNAs vary in size from ~75 nucleotides to several kilobases, and they play a wide variety of roles in bacterial gene expression (69, 70). In some cases, asRNAs have been found to mediate transcriptional attenuation (e.g., 71, 72, and references therein), but in most cases asRNA is involved in modifying translation and stability of its mRNA targets. Thus, asRNA can increase mRNA stability as seen in the *E. coli* GadY, where base-pairing between asRNA and mRNA leads to an RNase III-mediated cleavage of an unstable polycistronic mRNA into its more stable monocistronic derivatives (73, 74). Another example of asRNA-mediated mRNA stabilization is seen in *Prochlorococcus*, where duplex formation between asRNA and mRNA masks an RNase E cleavage site on the mRNA (75).

More often, however, asRNAs are involved in mRNA destabilization and degradation. Thus, asRNAs can block translation initiation, as seen in the “classical” antisense scenario where the asRNA covers the translation initiation region of the target mRNA, often followed by destabilization of the mRNA in a manner similar to many *trans*-encoded sRNAs (69). More recently, asRNA has been shown to play a significant role in genome-wide RNase III-mediated processing of mRNA in Gram-positive bacteria (76). Consistent with this mechanism is the observation that in *M. tuberculosis* there is an inverse correlation in the abundance of sense and antisense transcripts of individual genes/gene classes (7). The fraction of asRNA remains relatively constant around 10% of total RNA, although individual transcripts show variation between growth phases.

Mycobacterial antisense transcripts have been identified by cDNA cloning followed by sequencing, by computational predictions, by expressing *E. coli* Hfq followed by pull-downs (in *Mycobacterium smegmatis*), by RNA-seq, and by TSS mapping. The sizes of these asRNAs as well as their locations relative to their cognate ORFs vary significantly: some are encoded at the 5′ end of the ORF, some in the center, and some at the 3′ end, and a few cover an entire ORF or more (7–9). asRNAs can be expressed from SNP-generated promoters in clinical isolates of *M. tuberculosis* (77). The SNPs can be synonymous or nonsynonymous with respect to the mRNA but in most cases result in formation of a −10 region with the consensus TANNNT on the antisense strand, which appears to be sufficient for expression of the asRNA in the absence of any obvious −35 motif (16). An interesting example is a mid-ORF asRNA opposite the *ino1* gene (Rv0046c), expressed from a TANNNT promoter generated by an identical SNP acquired independently in sub-branches of *M. tubercu-
loss lineages 1 and 4 (including the common lineage 4 laboratory model strain H37Rv). The asino1 is highly expressed during exponential growth and significantly downregulated in stationary phase (7) (Fig. 5). ino1 encodes an enzyme that catalyzes the first step in inositol synthesis, and this gene has been shown to be essential for virulence (78). It is currently unknown whether the regulated expression of asino1 has an effect on the synthesis of inositol. Overall, the degree of conservation of the antisense transcriptome among clinical isolates of M. tuberculosis is similar to that observed for the coding transcriptome (77).

The general conception has been that asRNAs target only their cognate mRNA. However, it was recently demonstrated that a cis-encoded sRNA from the archaean Methanosarcina mazei had a second trans-encoded mRNA target (79). At least two M. tuberculosis cis-encoded RNAs, both associated with lipid metabolism, have the potential to base-pair with multiple mRNAs generated from duplicated gene sequences (8). One of these, ASpks, is located opposite of pks12, which encodes one of multiple polyketide synthases required for biosynthesis of complex lipid molecules involved in pathogenesis (80–82). ASpks is a 75-nucleotide transcript that maps to a duplicated module in the 12.5-kb pks12 gene (Rv2048). Moreover, it has significant sequence identity with the equivalent modules in pks7, pks8, and pks15 and hence could potentially base-pair with these mRNAs as well (Fig. 6). During oxidative stress, a larger 200-nucleotide ASpks transcript is

---

**FIGURE 5** The asino1 RNA is highly expressed during exponential growth and significantly downregulated in stationary phase in H37Rv. The asino1 RNA is expressed in the majority of lineage 4 strains (here represented by H37Rv) and in some lineage 1 strains (represented by N0157), but not in other lineages (N0052—lineage 2—has been shown for comparison). RNA-seq data is visualized in Artemis. All reads are normalized to total reads and adjusted to the same scale. doi:10.1128/Microbiolspec.MGM2-0029-2013.f5
induced (8), while oxidative stress also results in a downregulation of potential target pks12, pks8, and pks15 mRNAs (83). Further analysis should clarify whether this is due to the increased ASpks levels and whether this has an effect on Pks12-specific host immune recognition (84).

The extent of the antisense transcriptome suggests that it may represent a common component of gene regulation in M. tuberculosis, with the potential in some cases for cis-regulation of their cognate mRNA partner as well as coordinated trans-regulation of related genes.

**trans-Encoded Base-Pairing RNAs**

**trans**-Encoded base-pairing RNAs are encoded in IGRs, which in most cases are distinct from the locations of their targets. These sRNAs share the common feature of imperfect complementarity to their mRNA targets, posing a challenge for target prediction (85, 86). Many sRNAs are regulated by stress stimuli and may therefore be associated with host adaptation and virulence in pathogens (reviewed in references 1, 87–89).

Most trans-encoded sRNAs have multiple targets and act by modifying the accessibility to the ribosome-binding site of their target mRNAs; hence, they are classified as posttranscriptional regulators of gene expression. However, blocking ribosome entry leads to an uncoupling of transcription from translation, associated with rho-dependent termination of transcription and, in the case of polycistronic mRNAs, polarity. This scenario has been demonstrated for the bicistronic ChiPQ mRNA regulated by the ChiX sRNA but is likely to be of significance in the regulation of other polycistronic transcripts (90).

sRNAs can regulate any gene within any regulon and tend to target genes from different regulons, thus adding complexity to regulation rather than following the same paths as their own transcriptional regulators. An emerging theme is that many sRNAs regulate transcription factors, whereby their regulatory potential is significantly enhanced (65, 91, 92). Examples of this mechanism are the E. coli/Salmonella RpoS, targeted by DsrA, ArcZ, and RprA (reviewed by Battesti et al. [93]), and the S. aureus Rot, targeted by RNA III (94). Alternatively, an sRNA can repress the expression of one or more regulators that drive the transcription of the same sRNA, thereby ensuring a faster return to homeostasis. Two coexisting examples of this type of mechanism can be found in the Vibrio cholerae quorum-sensing circuit in which LuxO and HapR both positively regulate the expression of four sRNAs, Qrr1 to 4, which in turn repress the expression of both LuxO and HapR (95).
Many intergenic sRNA candidates have been identified in mycobacteria, though only a fraction of these transcripts have been verified by Northern blotting, TSS mapping, and/or qRT-PCR (7–14). As noted above, it is important to distinguish intergenic sRNAs from truncated 5’ UTRs and from misannotated short protein ORFs. Table 1 provides a list of partially characterized \textit{M. tuberculosis} intergenic sRNAs. \textit{M. tuberculosis} sRNAs display varying degrees of conservation: some are restricted to closely related members of the \textit{M. tuberculosis} complex; others are present in multiple pathogenic mycobacteria, with \textit{Mycobacterium leprae} being the most distant relative; and finally, some are conserved in all mycobacteria, with a few being identified in other actinomycetes (8, 96).

In Gram-negative bacteria, including many pathogens, the RNA chaperone Hfq plays a central role by facilitating the interaction between \textit{trans}-encoded sRNAs and their targets, with deletion of the \textit{hfq} gene often leading to loss of virulence (reviewed in references 89, 97). However, the role of Hfq in the Gram-positive \textit{S. aureus} is more controversial (98, 99), and the protein is completely absent from several pathogens, ranging from the AT-rich (39% GC) \textit{Helicobacter pylori} to the GC-rich (67% GC) \textit{M. tuberculosis} (97). An obvious question is whether \textit{M. tuberculosis} sRNAs interact with their targets unaided or whether in \textit{M. tuberculosis} there are one or more alternative, hitherto unidentified chaperones. The high GC content in \textit{M. tuberculosis} limits the frequency of AU-rich stretches implicated in conventional intrinsic terminators (100, 101). \textit{M. tuberculosis} may have a chaperone that has functional but not sequence homology with Hfq, or regulon-specific chaperones, each of which could interact with a subset of sRNAs and their targets in a manner similar to the FbpABC proteins in \textit{B. subtilis} (102); or it may be that the high GC content facilitates interactions between sRNAs and their mRNA targets independently of chaperone mediation (103). Future investigations, in which tagged sRNAs are used as bait for the isolation of RNA-binding proteins and hence putative chaperones (104), will shed more light on these questions in the near future.

**STRESS RESPONSES OF INTERGENIC sRNAs**

The abundance of many intergenic sRNAs changes in response to different environmental stresses. Conditions such as starvation in phosphate-buffered saline, stationary phase, or infection have been shown to enhance the expression of numerous sRNAs (7–14). However, all of these conditions represent a mixture of environmental stresses, and the application of specific stresses may tell us more about the regulatory pathways associated with individual sRNAs. The following section provides some examples of changes in \textit{M. tuberculosis} sRNA expression upon the application of separable stresses encountered in the host environment.

**Oxidative Stress**

Three of the sRNAs listed in Table 1—ncRv10243, ncRv10609A, and ncRv13660c—show increased expression in response to \textit{H2O2}, which mimics the oxidative stress encountered inside the host macrophage; hence, these sRNAs are likely to be induced during the early stages of infection and may play a role in the first stages of host adaptation.

ncRv10243 is encoded between two genes involved in lipid degradation: Rv0243 and Rv0244c. The ncRv10243 promoter has been shown to have the highest occupancy of SigF, suggesting that expression of ncRv10243 has high priority under SigF-inducing conditions (105). Consistent with this prediction, expression of ncRv10243 is virtually abolished in a SigF knockout strain (J. Houghton and K. B. Arnvig, unpublished). Overexpression of ncRv10243 leads to reduced growth of \textit{M. tuberculosis}, but although an ncRv10243 homologue is present in \textit{M. smegmatis}, overexpression of ncRv10243 in \textit{M. smegmatis} has no obvious growth phenotype (8). Overexpression of the \textit{M. smegmatis} homologue has not been tested in either background.

ncRv13660c is ~95 nucleotides in size and is located between Rv3660c, which encodes the septum site determining protein Ssd (106), and Rv3661, which is predicted to be involved in cellular differentiation. Downstream of Rv3661 is another sRNA, ncRv13661 (see below), and the entire locus is highly conserved among mycobacteria, with the notable exception of \textit{M. leprae}. The predicted structure of ncRv13660c contains a so-called 6C motif consisting of two C-rich loops found to be widespread among \textit{Actinobacteria}, and this structure has led to the suggestion that this sRNA may be a structural/protein binding RNA rather than a baseparing regulatory sRNA (47). On the other hand, C-rich loops have been shown to interact specifically with the purine-rich Shine-Dalgarno sequence in \textit{S. aureus}, thereby blocking translation initiation (107). Additional \textit{M. tuberculosis} sRNAs are predicted to form similar C-rich loops (8), and due to the high GC content of the bacterium, one can imagine this to be a relatively common feature of \textit{M. tuberculosis} sRNAs. In addition, it is worth noting that while \textit{M. tuberculosis} has no Hfq, the
C-rich sRNAs in *S. aureus* are independent of Hfq for function, suggesting the possibility that this feature may be a signature for Hfq-independent sRNAs. *M. smegmatis* harbors an sRNA that is >90% identical to ncRv13660c. Although (strong) overexpression of ncRv13660c in *M. tuberculosis* is lethal, the same level of overexpression of ncRv13660c in *M. smegmatis* is possible but results in a dramatic phenotype: growth is very poor on solid and in liquid media, and the cells display aberrant morphology including irregular and elongated shape, suggesting association with cell wall synthesis and/or cell division (8).

**Hypoxia**

The master regulator of the hypoxic response in *M. tuberculosis* is DosR (108). The DosR regulon includes genes required for remodeling of protein, lipid, and energy metabolism together with members of the universal stress protein family and multiple proteins of unknown function (83). The DosR genes are highly expressed during infection in mice and in sputum samples from patients, and their expression has been linked to the generation of drug-tolerant persister cells (109, 110). A number of *M. tuberculosis* intergenic sRNAs accumulate to high levels during the transition from exponential to stationary phase, which represents nutrient starvation as well as some degree of hypoxia depending on the method of cultivation (7–9, 12–14). Some of these accumulate to even higher levels during infection, again suggesting a role in host adaptation (7, 14). An example is ncRv11733, which is almost undetectable during exponential growth. Interestingly, the number of ncRv11733 molecules relative to the number of ribosomes, estimated to be around 4,000 per cell during fast growth (111), implies that less than 10% of exponentially growing cells express ncRv11733. It therefore seems likely that the few cells expressing this sRNA are in a different metabolic state than the remaining population, and therefore these could represent putative persister cells. Moreover, the induction of ncRv11733 seen in stationary phase is almost eliminated by deletion of the dormancy regulator, DosR (7). It remains to be determined if ncRv11733 is directly involved in the generation of persister cells or is simply a marker for this subpopulation.

**Acid Stress**

*M. tuberculosis* is well known for its ability to inhibit phagosome maturation (reviewed by Pieters [112]). Nevertheless, acid stress and acid resistance still play major roles in *M. tuberculosis* intracellular survival (21, 113). A number of *M. tuberculosis* sRNAs are induced by low pH, including ncRv10243 (8) and ncRv13661 (K. B. Arnvig, unpublished). ncRv13661 is the most highly expressed nonribosomal RNA during all growth conditions investigated, with levels increasing ~10-fold between exponential and stationary phase and abundance in infected mice approaching 1:1 stoichiometry with rRNA (7). Overexpression of ncRv13661 in exponential phase results in a phenotype that resembles stationary phase and models for persistence in terms of a significant downregulation of active growth functions (7, 108, 114, 115). Genes that are most significantly downregulated by ncRv13661 overexpression are associated with the methyl citrate cycle, which may explain the reduced growth on propionate observed in the overexpression strain (7; Houghton and Arnvig, unpublished).

Downregulation of active growth functions during stationary phase transition in bacterial cultures is commonly associated with increased expression of the structurally conserved 6S RNA that binds to RNA polymerase and inhibits transcription of genes controlled only by σ70 (67, 68, 116). However, there is to date no evidence of an immediately recognizable 6S homologue, although ncRv13661, or more specifically its *M. smegmatis* homologue, have been shown to have some structural relationship to 6S (96). It remains to be determined if ncRv13661 is a conventional sRNA that acts via base-pairing with specific mRNA or protein targets or whether its mode of action differs entirely from other sRNAs.

So far, the regulatory targets of all *M. tuberculosis* sRNAs remain elusive. Target prediction web servers such as TargetRNA (86), the enhanced version TargetRNA2 (http://cs.wellesley.edu/~btjad/en/TargetRNA2/), and RNApredator (http://rna.tbi.univie.ac.at/RNApredator2/target_search.cgi [85]) provide some suggestions; however, these are often not identical and require experimental verification just like other computational predictions.

**RNA Expression and Degradation**

Increased abundance of sRNAs under different growth conditions may be the result of increased expression and/or decreased degradation. An understanding of the dynamics and mechanisms of RNA degradation in *M. tuberculosis* will be an essential component of the effort to understand the function of ncRNAs. RNase E is the principal catalytic component of the *E. coli* degradosome. This multicomponent complex is a key player in RNA processing and turnover, including sRNA-mediated mRNA degradation (20). Very little is known about RNase E and RNA metabolism in general in

---

**Noncoding RNA in Mycobacteria**

ASMscience.org/MicrobiolSpectrum
### TABLE 1 Intergenic *M. tuberculosis* sRNAs

<table>
<thead>
<tr>
<th>New annotation&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Other&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Tuberculist&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Rfam&lt;sup&gt;e&lt;/sup&gt;</th>
<th>Strand</th>
<th>Flanking CDSSs</th>
<th>Verified by</th>
<th>5'&lt;sup&gt;i&lt;/sup&gt; (RACE)&lt;sup&gt;j&lt;/sup&gt;</th>
<th>5'&lt;sup&gt;i&lt;/sup&gt; (TSS)&lt;sup&gt;j&lt;/sup&gt;</th>
<th>3'&lt;sup&gt;k&lt;/sup&gt; &lt;sup&gt;n&lt;/sup&gt;</th>
<th>Expression&lt;sup&gt;i&lt;/sup&gt;</th>
<th>Conservation&lt;sup&gt;i&lt;/sup&gt;</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ncRv10243</td>
<td>F6, mcr14, mpr13</td>
<td>MTB000051</td>
<td>RF01791</td>
<td>F</td>
<td>Rv0243; Rv0244c</td>
<td>Northern, RACE, TSS</td>
<td>293604</td>
<td>293602</td>
<td>293705</td>
<td>SigF; H2O2, low pH</td>
<td>NPMB</td>
<td>8, 9</td>
</tr>
<tr>
<td>ncRv10609A</td>
<td>B55</td>
<td>MTB000052</td>
<td>RF01783</td>
<td>F</td>
<td>Rv0609A; Rv0610; Rv0685; Rv0686</td>
<td>Northern, TSS, Rv0244c, Rv0244c, Rv0609A, Rv0610, Rv0685, Rv0686</td>
<td>704187</td>
<td>704187</td>
<td>704247</td>
<td>SigA; H2O2</td>
<td>MTBC</td>
<td>8</td>
</tr>
<tr>
<td>ncRv10685c</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>R</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>786164</td>
<td>786038; 786090</td>
<td>—</td>
<td>MTBC</td>
</tr>
<tr>
<td>ncRv11051</td>
<td>mpr5</td>
<td>MTB000061</td>
<td>—</td>
<td>F</td>
<td>Rv1051c; Rv1052; Rv1144; Rv1145</td>
<td>Northern, TSS, Rv1051c, Rv1052, Rv1144, Rv1145</td>
<td>—</td>
<td>1174124; 1175178; 1272168; 1271907</td>
<td>—</td>
<td>—</td>
<td>PMB</td>
<td></td>
</tr>
<tr>
<td>ncRv11144c</td>
<td>MT50900</td>
<td>—</td>
<td>—</td>
<td>R</td>
<td>Rv1147; Rv1148c</td>
<td>Northern, TSS, Rv1147, Rv1148c</td>
<td>—</td>
<td>1275957; 1275993</td>
<td>—</td>
<td>—</td>
<td>NPM</td>
<td>162 and Gaudion et al., unpublished</td>
</tr>
<tr>
<td>ncRv11147c</td>
<td>MT50930</td>
<td>—</td>
<td>—</td>
<td>R</td>
<td>Rv1264; Rv1265</td>
<td>Northern, RACE, TSS, Rv1264, Rv1265</td>
<td>1413224</td>
<td>1413225</td>
<td>1413107</td>
<td>SigA; CRP, stationary phase, low pH, infection</td>
<td>MTBC</td>
<td>7</td>
</tr>
<tr>
<td>ncRv11264c</td>
<td>mcr11, MT50997, MT1502</td>
<td>MTB000063</td>
<td>—</td>
<td>R</td>
<td>Rv1373; Rv1374c</td>
<td>Northern, TSS, Rv1373, Rv1374c</td>
<td>—</td>
<td>1547129; 1547350**</td>
<td>—</td>
<td>—</td>
<td>PMB</td>
<td>112 and Gaudion et al., unpublished</td>
</tr>
<tr>
<td>ncRv11373</td>
<td>MT51082**</td>
<td>MTB000076</td>
<td>—</td>
<td>F</td>
<td>Rv1435c; Rv1436; Rv1689; Rv1690</td>
<td>Northern, TSS, Rv1435c, Rv1436, Rv1689, Rv1690</td>
<td>—</td>
<td>1915190; 1915028; 1915028; 1915028; 1915028; 1915028; 1914962</td>
<td>—</td>
<td>—</td>
<td>NPMB</td>
<td>8</td>
</tr>
<tr>
<td>ncRv11435</td>
<td>RNA1</td>
<td>—</td>
<td>—</td>
<td>F</td>
<td>Rv1436</td>
<td>Northern, TSS, Rv1436</td>
<td>—</td>
<td>1915028; 1915028; 1915028; 1915028; 1915028; 1915028; 1915028; 1915028; 1914962</td>
<td>—</td>
<td>—</td>
<td>MTBC</td>
<td>7</td>
</tr>
<tr>
<td>ncRv11689c</td>
<td>G2</td>
<td>MTB000054</td>
<td>RF01798</td>
<td>R</td>
<td>Rv1733c; Rv1734c</td>
<td>Northern, RACE, TSS, Rv1733c, Rv1734c</td>
<td>1960667</td>
<td>1960667</td>
<td>1960783</td>
<td>exponential growth, DosR, stationary phase, NO, infection</td>
<td>MTBC</td>
<td>7</td>
</tr>
<tr>
<td>ncRv11733</td>
<td>MT51338*</td>
<td>MTB000077</td>
<td>—</td>
<td>F</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>MTBC</td>
<td>7</td>
</tr>
<tr>
<td>ncRv1789c</td>
<td>mcr1</td>
<td>—</td>
<td>—</td>
<td>R</td>
<td>Rv1179; Rv1180 Northern, TSS — 2028215 — MTBC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRv12659</td>
<td>MTS2048*</td>
<td>—</td>
<td>—</td>
<td>F</td>
<td>Rv2659c; Rv2660c Northern, RACE, TSS 2980911 2980911 2981083 SigA; starvation MTBC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRv13059</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>F</td>
<td>Rv3059; Rv3060c Northern, TSS — 3421223 MTBC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRv13660c</td>
<td>B11,mpr19</td>
<td>MTB000058</td>
<td>RF01066</td>
<td>R</td>
<td>Rv3660c; Rv3661 Northern, RACE, TSS 4099478 4099480 4099386 SigA; H2O2 stationary phase, low pH, infection NPM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRv13661</td>
<td>mcr8, mpr4, MTS2823</td>
<td>MTB000078</td>
<td>—</td>
<td>F</td>
<td>Rv3661; Rv3662c Northern, RACE, TSS 4100683 4100682 4100987 NPM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRv13943</td>
<td>MTS2975</td>
<td>MTB000080</td>
<td>—</td>
<td>F</td>
<td>Rv3943c; Rv3844 Northern, TSS — 4317071 PMB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRv13596</td>
<td>mpr17</td>
<td>MTB000070</td>
<td>—</td>
<td>F</td>
<td>Rv3596c; Rv3597c Northern, TSS — 4040805 NPMB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Most of the listed sRNAs have been verified by at least two methods; transcripts listed as sRNAs/sRNA candidates in primary publications that are not listed here include those that are likely to be 5′ UTRs, fragments of mRNAs, those in which the ends are not mapped and cannot be deduced by our TSS mapping, and those that are antisense to annotated ORFs. Exceptions include ncRv11373, ncRv11733, and ncRv12659, for which we have found no evidence for expression from the opposite strand and which have all been further described elsewhere (7, 190, 191).

Refers to annotation rules in Lamichhane et al. (15).

Refers to the names given in previous publications.

Refers to annotation on Tuberculist.

Rfam identifier if present.

Identified by RNA-seq TSS mapping, all from reference 16.

Approximate 3′ end identified by RACE or canonical terminator structure (first U) or calculated from 5′ end and size; according to primary publication.

Conservation (BLAST only): MTBC, only identified in M. tuberculosis complex; PMB, also in other pathogenic mycobacteria beyond MTBC; NPMB, also in nonpathogenic mycobacteria.

Indicates that sRNA is antisense to annotated ORF but that we have seen little (weak TSS, and no total RNA) or no (no TSS, no total RNA) evidence of expression from ORF annotation strand.

As suggested by RNAseq and terminator structure; Northern suggests shorter (processed) transcript.
M. tuberculosis, although factors such as hypoxia and cold shock have been shown to have a stabilizing effect on mRNA (117). M. tuberculosis RNase E appears to interact only with a subset of the degradosome components found in enteric bacteria (118). On the other hand, the Gram-positive B. subtilis does not encode an RNase E homologue, and hence the B. subtilis degradosome has a significantly different composition (119). This raises a number of interesting questions with respect to mRNA-sRNA interactions, M. tuberculosis chaperones, and the M. tuberculosis degradosome including (i) does M. tuberculosis form an RNA degradosome-(like) structure, and (ii) what is its role in RNA metabolism and, in particular, sRNA regulation? The stable secondary structures predicted for many intergenic sRNAs may inhibit their degradation and contribute observed increases in abundance.

**OTHER M. TUBERCULOSIS ncRNAs**

**CRISPR**

The clustered regularly interspaced short palindromic repeat (CRISPR) locus provides a specialized source for generation of M. tuberculosis sRNAs. The CRISPR locus incorporates sequences from phage and other invading genetic elements and repackages them as sRNA defense molecules that confer resistance to reinfection (reviewed by Deveau et al. [120]). The CRISPR loci vary substantially between species of mycobacteria, and the composition of the CRISPR in Mycobacterium canetti suggests acquisition by horizontal transfer prior to final branching of the M. tuberculosis complex (121). Sporadic deletion of integrated foreign sequence elements from the CRISPR domain has been exploited in a widely used system for differentiation of M. tuberculosis strains referred to as “spoligotyping” (122). Phylogenetically, the M. tuberculosis CRISPR system belongs to the type III-A family (123 and references therein) and is related to a well-characterized CRISPR system in Staphylococcus epidermidis (124). Inspection of RNA-seq profiles shows that the CRISPR domain is transcribed and processed into a series of 50 to 60 nucleotide CRISPR-RNAs (crRNAs) that are presumably incorporated into a cascade effector complex. The transcription start site driving crRNA expression overlaps with a divergent start site that generates a long antisense transcript covering Rv2816c and Rv2817c, which code for the Cas2 and Cas1 proteins, respectively, involved in incorporation of novel sequences into the CRISPR locus (Fig. 7) (2).

This profile is consistent with silencing of the genes required to capture novel crRNAs, but functional expression of Cas6 and Cas10 enzymes (Rv2824c and Rv2823c) is required for processing of pre-existing crRNAs. An exception is the sublineage of M. tuberculosis known as the Beijing family, which is undergoing a current global expansion in marked association with multiple drug resistance (125). A genomic deletion has removed most of the CRISPR locus and many of the associated Cas genes, eliminating expression of crRNAs in these strains (126). A similar deletion has been found in an unrelated sublineage referred to as “pseudo-Beijing” (127). It is open to speculation whether the presence of a functional CRISPR system in the majority of M. tuberculosis strains has current biological relevance or reflects fortuitous retention of a defense system appropriate to an ancestral lifestyle.

**HOW TO IDENTIFY ncRNA**

Until 2001, only 10 sRNAs had been identified in the model organism E. coli through genetic screens or radiolabeling of total RNA and subsequent isolation from gels (128). Over the last decade, a combination of computational prediction and experimental approaches has allowed identification of a wider set of sRNAs in E. coli and other closely related bacteria like Salmonella typhimurium (129). Attempts to predict sRNAs in bacteria involve computational prediction methods that employ RNA sequence homology, comparative genomics, thermodynamically favorable secondary structures, detection of transcription signals, and ab initio methods using sRNA-specific features. However, the use of computational methods for the detection of sRNAs in bacterial genomes is not easy because sRNAs are diverse in length (ranging from 50 to 500 nucleotides), they do not share a common secondary structure, they do not exhibit any clear nucleotide biases, and generally they are not well conserved among distantly related bacteria. In combination with the recent development of powerful bio-computing algorithms for the prediction of ncRNA, the advent of high-throughput sequencing technologies, including RNA-sequencing and 5′ transcription start site mapping, has allowed the identification of sRNAs at a genome-wide level followed by in vivo validations. This section reviews the past and current methods used for the identification of ncRNAs in bacteria, with a special focus on the identification in mycobacteria.

**Computational sRNA Prediction**

One of the major challenges in sRNA prediction is that although there is a common acceptance of some shared characteristics such as small size (usually less than 200
nucleotides) and lack of coding capacity, there is a general lack of accepted identifiers for bacterial sRNAs. Consensus structure prediction is a good start for sRNA prediction when a given sRNA secondary structure is known. Tools like RNAMotif (130) can be used to scan defined secondary structures of RNA throughout the genome, but they are limited by the known secondary structure of a specific sRNA family.

Comparative genomics has been extensively applied for de novo sRNA prediction in bacterial genomes (131-135). Briefly, shared conserved sequences are first identified in IGRs, followed by clustering and comparison by multiple alignments that are scored based on predicted RNA structural features. Programs like eQRNA (136), ERPIN (137), ISI (138), INFERNAL (139), MSARI (140), and evofold (141) have been used for the detection of bacterial sRNAs by comparative genomics. RNAz is a program that combines comparative genomics with thermodynamic stability values (prediction of conserved stable RNA secondary structures) for predicting structural noncoding RNAs in bacteria (142, 143).

Noncomparative approaches have been based on systematic scanning for transcriptional signals, including conserved promoter sequences, rho-independent/intrinsic terminators, transcription factor binding sites and orphan transcriptional signals. sRNAPredict (144) is an algorithm for the prediction of putative sRNAs in bacteria, using databases like TRANSTERMHP (145) or TRANSFAC (146). sRNAPredict3 and SIPHT are recent versions of this program that have been used for the prediction of sRNAs in bacteria (147). sRNAscanner (148) has recently been developed to overcome the limitation of the lack of transcription signals predicted in all the available

**FIGURE 7** RNA-seq data (visualized in Artemis) of the *M. tuberculosis* CRISPR locus. The upper trace records TSS mapping (i.e., enriched for primary transcripts) from the right-hand side of the CRISPR locus, showing overlapping start sites for the Rv2816c antisense transcript (forward direction in blue) and the single CRISPR-RNA (crRNA) transcript (reverse, in red). The lower trace records sequencing of total RNA, showing the antisense transcript covering Rv2816c and Rv2817c (blue) and illustrating how the single crRNA transcript is processed into a series of mature, smaller crRNAs (red). A similar crRNA profile is seen on the left-hand side of the CRISPR locus, upstream of the Rv2614c/Rv2615c IS6110 insertion sequence (not shown). doi:10.1128/microbiolspec.MGM2-0029-2013.f7
genome sequences and has been proved to be an efficient platform for the prediction of sRNAs in any bacterial genome.

*Ab initio* sRNA finders are used for the identification of sRNAs based on specific RNA structural elements such as known RNA sequence motifs, di/trinucleotide preferences, uridine loops, or atypical GC content. Programs like RNAGENiE ([149]), Atypical GC ([150]), and NAPP ([151]) can be used.

**Low-Throughput Experimental Methods and Computational Predictions**

The first 10 ncRNAs discovered were found incidentally following genetic screens in *E. coli*. The first group identified included 4.5S, 6S, tmRNA, RNase P, and Spot 42 detected by metabolic radiolabeling ([152–154]). OxyS and CsrB sRNAs were detected while mapping the transcription start site of the adjacent regulator gene OxyR ([155]) and after copurification with overproduced CsrA protein ([156]), respectively. Finally, some sRNAs (e.g., MicF and DsrA) were identified when genomic fragments harboring these genes were shown to affect gene regulation in *E. coli* ([157, 158]).

Subsequent discovery of sRNAs has been based mainly on systematic genome-wide screens combining computational and experimental approaches, including microarrays and shotgun cloning ([159] and references therein). Identification of ncRNAs through computational approaches has relied mainly on sequence and structural analysis, conservation across species, and position within IGRs. The first computational-based genome-wide screens predicted several hundred new sRNA candidates in *E. coli* ([31, 136, 160, 161]). Genes encoding novel putative sRNAs were identified by Argaman et al. ([31]) following a computational strategy centered basically on three main parameters: length (search restricted from 50 to 400 nucleotides), sequence conservation, and presence of transcription signals (search for promoter sequences recognized by the *E. coli* principal RNA polymerase sigma factor σ^70^ and intrinsic terminators). Wassarman et al. ([161]) employed an algorithm that was also focused on sequence conservation but further extended the predictions of transcription signals, incorporating sequence elements for potential promoters, terminators, and inverted repeat regions, and a method developed by Chen et al. ([160]) restricted the detection of σ^70^ conserved promoter sequences within a short distance of a rho-independent terminator. While these methods were mainly based on primary sequence conservation of the identified sRNAs, the QRNA tool, developed by Rivas et al. ([136]) focused mainly on searching for conserved RNA secondary structures in combination with comparative genomics. Bioinformatic predictions of sRNA candidates identified by these algorithms were verified by Northern blot analyses and mapping of 5′ and 3′ ends, leading to the identification of several new sRNAs in *E. coli*.

The first computational predictions applied for the identification of ncRNAs in the *M. tuberculosis* genome were the result of a comparative genomics analysis carried out using the sRNAPredict2 program ([162]). The IGRs of 11 diverse pathogens were sought on the basis of sequence conservation and transcription signals (intrinsic terminators and putative promoter sequences) allowing the detection of 56 putative intergenic sRNAs in *M. tuberculosis*. Of the 56 predicted sRNA candidates, 1 (mpr8) was subsequently predicted by a different algorithm, tested by Northern blotting, and annotated as the 5′ UTR of *infC* ([7, 9]). Of the remaining sRNA candidates, a few have been verified in our laboratory by RNA-seq and/or Northern blotting (Table 1; A. Gaudion, K. B. Arnvig, and D. B. Young, unpublished).

The first experimental evidence for the existence of sRNAs in *M. tuberculosis* came from Arnvig and Young in 2009 and was based on screening of cDNA libraries generated from low-molecular-weight RNA ([8]). A set of nine ncRNAs was identified, five located within IGRs and four located antisense to annotated coding genes. Northern blot analyses demonstrated variations in expression between exponential and stationary growth phases and in response to environmental stresses. All of the sRNAs displayed predicted stable secondary structures, and some had recognizable intrinsic terminators, but none of them were predicted by the computational method discussed above ([162]); this may be due to a mixture of the heterologous nature of mycobacterial promoters and the lack of canonical terminator structure.

A mixed approach using a cloning-based screen together with a computational approach was used by DiChiara et al. for the identification of ncRNAs in *M. bovis* BCG ([9]). They described 37 sRNA candidates in BCG, 34 of which were novel. The computational search was performed using the SIPHT program, in which identification of ncRNAs is based on the presence of intergenic sequence conservation upstream of a putative intrinsic terminator. *In silico* identification of putative sRNAs yielded 144 candidates that were further reduced to 67 candidates partially conserved in mycobacterial species outside of the MTB complex. All 67 candidates were tested by Northern blot analyses and 21 sRNAs were confirmed, of which 5 were intergenic sRNAs. The cloning-based experimental approach
identified 60 candidates, 13 located in IGRs and 47 located inside annotated coding genes. Northern blot analysis confirmed 19, of which 4 were intergenic sRNAs. Among the overall set of intergenic sRNAs, only two, ncRv10243 and ncRv13661, were identified by both methodologies. The sRNAs were predicted to be present in a wide range of mycobacterial species, highlighting their potential regulatory role in highly conserved cellular functions. Sequence conservation identified similar sequences for all of them within the *M. tuberculosis* genome, although only 20% were found to be expressed in *M. tuberculosis* under the conditions tested, leading to the identification of 17 novel sRNA candidates.

Recently, Pelly et al. used a cDNA cloning strategy to identify sRNA species in *M. tuberculosis* CDC1551 in the size range of 70 to 200 nucleotides from RNA isolated during late exponential growth (14). After removal of tRNA and rRNA clones, 12 clones representing intergenic sequences were further investigated. The previously described sRNA ncrMT1302 (named mcr11 in reference 9 and MTS0997 in reference 7), located in a locus involved in cAMP metabolism, was further investigated; the study demonstrated that the expression of ncRv11264c responded to changes in pH and cAMP concentration and confirmed that the sRNA is expressed during infection.

### High-Throughput Experimental Methods

While low-throughput and computational methods have been mainly useful for the detection of sRNAs in IGRs, high-throughput methods such as tiling arrays and RNA sequencing (RNA-seq) have been very useful for the genome-wide detection of sRNAs and especially for the detection of asRNAs, unraveling a whole new level of antisense regulation in bacterial transcriptomes including *L. monocytogenes* (163), *S. aureus* (164), *B. subtilis* (165), and *Streptococcus pyogenes* (166). A tiling array consisting of overlapping 60-mer probes covering the entire genome was designed for the identification of ncRNAs in *M. leprae* (167).

The application of high-throughput sequencing technologies has revolutionized our understanding of ncRNAs and antisense regulation in bacterial genomes. RNA-seq has allowed the sequencing of whole transcriptomes in a strand-specific manner from total RNA or mRNA-enriched samples. Recent studies generating RNA-seq transcriptomes cover pathogenic bacteria as diverse as *L. monocytogenes* (168), *Bacillus anthracis* (169), *Burkholderia cenocepacia* (170), *Mycoplama pneumoniae* (171), *Salmonella* spp. (172, 173), *H. pylori* (32), and *Campylobacter jejuni* (174); these are all generating important results that are reshaping our understanding of bacterial gene regulation and adaptation.

The first *M. tuberculosis* transcriptome study using RNA-seq was performed by Arnvig et al. (7), who investigated whole-transcriptome expression during exponential growth and stationary phase. During exponential growth, ~79% of the annotated coding transcriptome had significant levels of expression (defined as reads mapped to that particular genome region), while during stationary phase the transcriptome expression decreased dramatically (only 11% of the genome expressed) and was mainly dominated by expression of DosR regulon genes. Regarding the non-coding transcriptome, this study revealed an extensive presence of ncRNA in the *M. tuberculosis* genome, including long 5′ and 3′ UTRs, antisense transcripts, and intergenic sRNA molecules (7). A substantial portion of antisense transcripts were derived from long 3′ UTRs of converging genes, whereas others were independent of 3′ UTRs and located antisense within annotated coding genes. The dominant sources of intergenic reads corresponded to 5′ UTRs and sRNA molecules. Twenty-one intergenic sRNA candidates were identified, some of which had been identified previously (8, 9); three of these (ncRv11264c, ncRv11733, and ncRv13661) were further characterized and shown to be highly upregulated in a mouse model of infection.

A comparative genomics approach combined with RNA-seq was conducted by Pellin et al. (13) to identify sRNA candidates in the *M. tuberculosis* genome. They constructed the *M. tuberculosis* effective target genome for ncRNAs (excluding regions annotated as coding genes and stable RNAs, hence only keeping IGRs and antisense sequences as possible regions for ncRNA discovery). By combining RNA-seq and IGR conservation analysis they predicted ~2,000 sRNA candidates in *M. tuberculosis*. Although this study has greatly increased the putative number of ncRNAs in the *M. tuberculosis* genome, only 52% of the previously described and verified sRNAs were identified by means of this approach. A subset of these potential sRNAs was further investigated by Miotto et al. (12) through microarray expression analysis. Approximately 19% of the previously predicted transcripts were validated by microarray expression profiling including intergenic sRNA candidates, 5′/3′ UTRs, and antisense sRNAs. These showed enriched higher secondary structure stability, and functional enrichment analyses suggested that antisense sRNAs might target genes involved in two-component systems and membrane activity. Analysis of sequences
upstream of the transcription start sites identified a SigA promoter consensus sequence (TANNNT) in the majority of cases and a canonical terminator structure in about one fifth. Twenty transcripts were verified by Northern blot analyses and/or primer extension. Furthermore, some putative riboswitches were described.

Copurification with Proteins

Copurification of sRNAs with proteins has been a successful experimental approach for the detection of sRNAs in bacteria (175). Coimmunoprecipitation (coIP) with Hfq antibodies followed by hybridization to tiling arrays, conventional RNomics, or RNA-seq has been a common technique for sRNA discovery in Gram-negative bacteria including *E. coli* (176), *Salmonella* (177), and *Pseudomonas aeruginosa* (178).

As discussed above, the role of the Hfq protein in Gram-positive bacteria is less clear, and no Hfq orthologs have been described in important bacterial pathogens like *H. pylori* and *M. tuberculosis*. Despite this, the RNA-binding potential of Hfq can still be exploited, and using coIP, Hfq-binding sRNAs have been identified in *L. monocytogenes* (179) and *B. subtilis* (180). Furthermore, by using heterologous Hfq, which presumably has different sequence specificity than the native protein, sRNAs that are not detected under standard assay conditions have been detected in *Salmonella* (181). In addition, Hfq has been expressed in *M. smegmatis*, which does not have Hfq at all (10). The latter example involved the expression of FLAG-tagged Hfq from *E. coli* in *M. smegmatis*, followed by coIP and RNA-seq of Hfq-binding transcripts. This led to the identification of 24 novel sRNAs including both cis-acting and trans-acting sRNAs targeting genes involved in metabolic pathways, as well as two of the previously described *M. tuberculosis* sRNAs (8). The 24 novel sRNAs have been verified by Northern blot analysis showing differential expression in exponential and stationary phases (10). Prediction of secondary structures using RNAfold (182–184) and centroidFold (185, 186) identified a typical I-shaped terminator or a stem-loop without a long poly(U) stretch for most of the identified sRNAs. Phylogenetic conservation among mycobacteria revealed that 13 out of 24 were exclusively present in *M. smegmatis*, and only 2 were found in *M. tuberculosis*.

The application of coIP to isolate RNA binding to proteins other than Hfq includes CsrA (66, 187), σ70 RNA polymerase (188), and RNase III (189). However, the approach could in theory be applied to a larger number of RNA-binding proteins, some of which have been described by Pichon and Felden (175). Additional insights into the interactions between regulatory RNAs and proteins as well as regulatory RNAs and other RNAs will take us one step further into the world of mycobacterial regulatory RNA that we have only just entered.

GLOSSARY

- **3′ UTR** 3′ Untranslated region. Part of an mRNA downstream of a stop codon, located between the stop codon and a transcriptional terminator.
- **5′ UTR** 5′ Untranslated region. The 5′ part of an mRNA that is not translated, i.e., located between a transcription start site and translation start site.
- **Antisense** RNA that can base-pair with sense RNA; the pairing can be imperfect or perfect (“true” antisense).
- **cis-Encoded** “True” antisense. RNA encoded in the same genomic location but opposite its target. Can cover coding sequence as well as UTRs.
- **cis-Regulatory** A segment of RNA that regulates expression of another segment of the same transcript, e.g., UTRs, riboswitches, etc.
- **IGR** Intergenic region. A sequence between open reading frames.
- **Leader region** Similar to 5′ UTR but part of ncRNAs such as ribosomal RNA (rRNA) or tRNA.
- **Leaderless mRNA** mRNA that lacks a 5′ UTR and hence the Shine-Dalgarno sequence involved in ribosome binding. In leaderless mRNAs the transcription start site and the translation start site are the same.
- **ncRNA** Any RNA that does not code for a protein or peptide.
- **Primary transcript** Unprocessed transcript. Characterized by a 5′ triphosphate, unlike processed transcripts that have a 5′ monophosphate or 5′ OH.
- **Regulatory RNA** RNA that has regulatory potential in the nucleotide sequence. Usually UTRs and sRNAs that respond to changes in growth conditions, but in essence most RNA harbors regulatory potential in the form of codon bias, Shine-Dalgarno sequences, etc.
- **Riboswitch** cis-regulatory 5′ end of mRNA that changes conformation in response to changes in concentration of small molecule ligands; can be on or off switches and can act on either a transcriptional or translational level.
Shine-Dalgarno (SD) sequence  AG-rich hexameric sequence located approximately 10 to 20 nucleotides upstream of the start codon. The SD sequence has partial complementarity to 16S rRNA and thus facilitates binding of the small ribosomal subunit to the mRNA.

sRNA  Abbreviation of small RNA, mostly refers to small regulatory RNA.

trans-Encoded sRNA encoded in a different genomic location than its target.

Translation initiation region  Part of mRNA where translation initiates. Includes Shine-Dalgarno (SD) sequence and initiation/start codon.

TSS  Transcription start site. The nucleotide position(s) where RNA polymerase initiates transcription.

REFERENCES


Noncoding RNA in Mycobacteria


Noncoding RNA in Mycobacteria


