Term-seq reveals an abundance of conditional, Rho-dependent termination in *Mycobacterium tuberculosis*.

Alexandre D'Halluin[1], Peter Polgar[1], Terry Kipkorir[1], Zaynah Patel[1], Teresa Cortes[2], Kristine B. Arnvig[1]*

[1] Structural and Molecular Biology, University College London, London WC1E 6BT, UK

[2] Instituto de Biomedicina de Valencia (IBV), CSIC, Valencia 46010, Spain

* To whom correspondence should be addressed. Tel: +44 2076792242 Email: k.arnvig@ucl.ac.uk

ABSTRACT

Accurately controlled termination of transcription is critical for gene expression and adaptation to changing environments. The aim of this study was to define transcription termination sites in *Mycobacterium tuberculosis* at nucleotide resolution. To this end, we applied Term-seq to cultures of *M. tuberculosis* and established that by far the most abundant mechanism of transcription termination is Rho-dependent. Moreover, we find that conditional or premature termination of transcription regulates a large proportion of genes and that this is often associated with translated upstream open reading frames (uORFs). Finally, our results show that overlapping ORFs are abundant in many cases in the absence of a recognisable Shine-Dalgarno sequence, indicating tightly coordinated/coupled translation. Our results provide a detailed image and comprehensive catalogue of novel *M. tuberculosis cis*-regulatory elements, where Rho-dependent conditional termination of transcription and translational coupling together play major roles in gene expression control.

INTRODUCTION

The control of gene expression plays a fundamental role in the pathogenesis of *Mycobacterium tuberculosis,* the cause of tuberculosis (TB) e.g. (1-4). Transcription is the first step of gene expression, and while the process of transcription initiation and the role of transcription factors are relatively well characterised in *M. tuberculosis*, reviewed in (5), the role and the molecular mechanisms underlying transcription elongation and termination are poorly understood. Few canonical intrinsic terminators, i. e. a stem-loop followed by a poly-U tract ('L-shaped') have been identified, let alone investigated in *M. tuberculosis* and results are conflicting. The search for alternative mechanisms behind *M. tuberculosis* intrinsic termination has led to the suggestion that 'I-shaped' terminators, i.e. a stem-loop without poly-U tail may be sufficient (6), but more recent findings indicate that, as a minimum, a partial or interrupted poly-U stretch needs to be in place for unidirectional, factor-independent termination to take place (7-10).

A second mechanism of transcription termination involves head-on collisions between transcription elongation complexes on converging genes. This is often associated with hairpins and accounts for a significant proportion of termination events in *Escherichia coli* (11).

A third mechanism of transcription termination is mediated by Rho, a hexameric helicase that binds to pyrimidine-rich stretches (*RUT* sites) on nascent RNA with a preference for

C-residues and a C:G ratio > 1 (12,13). The sequence bias other than favouring pyrimidines is limited, and the transcription termination site (TTS) can be located up to 100 nucleotides downstream of the primary *RUT* site (13). This and the immediate and rapid post-termination 3' processing makes identification of actual *RUT* sites challenging (14).

M. tuberculosis rho is essential for cell viability, suggesting that Rho-dependent termination of transcription plays an important role in the control of gene expression (15). Bicyclomycin is generally considered a potent inhibitor of Rho, however, due to a leucine to methionine substitution in *M. tuberculosis* Rho, bicyclomycin binding is obstructed, rendering *M. tuberculosis* resistant to the antibiotic (16). Instead, a Rho depletion strain (RhoDUC), in which the addition of Anhydrotetracyclin (ATC) simultaneously represses transcription of *rho* and induces degradation of the Rho protein has been used to characterise its role in *M. tuberculosis* (15). RNA-seq analysis of this strain indicated significant changes in gene expression after six hours of depletion eventually leading to cell death, corroborating the importance of Rho-dependent termination of transcription in *M. tuberculosis* (15).

Correct termination of transcription at the 3' end of genes or operons is an integral part of the transcription cycle, while premature termination of transcription in the 5' untranslated region (UTR) upstream of or early within an open reading frame (ORF), will repress the expression of downstream gene(s). This process, referred to as conditional termination offers an additional and widespread mechanism of regulation, and potential intervention (17). Conditional termination of transcription is a mode of post-transcriptional control, whereby specific physiological signals or ligands modulate the potential of a terminator to promote transcriptional readthrough or termination seen for example in the context of riboswitches and certain antibiotic resistance genes (18-20). The application of Term-seq to map RNA 3' ends has revealed an unexpected abundance of conditional terminators and associated *cis*-regulatory elements in other bacterial systems (19,21). However, this type of control remains largely unexplored in *M. tuberculosis* (22,23).

To better understand fundamental aspects of *M. tuberculosis* transcription termination and to identify conditional terminators associated with known and new 5' leaders and/or riboswitches, we have applied a combination of Term-seq (19) and tagRNA-seq (24) to map *M. tuberculosis* RNA 3' ends. We found that few *M. tuberculosis* transcription termination sites (TTS) were associated with canonical intrinsic terminators, while the vast majority of termination events were Rho-dependent. Moreover, almost half of all the

mapped TTS were associated with conditional termination of transcription and therefore likely play a role in the regulation of gene expression. We also found that conditional termination in many cases was associated with small upstream open reading frames (uORFs), located in regions previously considered as 5' UTRs, indicating that translating ribosomes provide a significant contribution to conditional termination of transcription. Our study provides novel and detailed insights into post-transcriptional control of gene expression in *M. tuberculosis,* with emphasis on conditional, Rho-dependent termination of transcription associated with translational coupling. Moreover, it comprises a rich catalogue of novel *cis*-regulatory RNA leaders for further molecular investigations.

RESULTS

TTS identification and classification in *Mycobacterium tuberculosis*

To identify *M. tuberculosis* TTS on a genome-wide scale, RNA was isolated from triplicate, log-phase cultures of *M. tuberculosis* H37Rv, using previously published methods (7,25); the RNA was subsequently analysed using Term-seq and tagRNA-seq to simultaneously map RNA 3' ends and 5' tri- and mono-phosphates, respectively, Fig. 1 (19,24). To catalogue the complete number of transcription start sites (TSS) used in our analyses, we combined our TSS mapping, obtained with tagRNA-seq with previously identified TSS from (26); newly identified TSS used in this study are listed in Supplementary Table 1.

To filter out low-confidence, non-specific peaks, a minimum threshold for 3' Term-seq peaks was determined (see materials and methods). A similar approach was adopted to identify potential processing sites (PS) from 5' tagRNA-seq. After calculating background noise levels associated with intergenic regions, a median coverage value of 1.27 normalised reads was set as threshold for detection for 5' mono-phosphate peaks (see materials and methods). We detected a total of 57,755 likely PS (Supplementary Table 2). Of these, 1,235 (2.1%) coincided with TSS, suggesting that they were conversions from 5' tri-phosphates to mono-phosphates. We observed no significant difference in the conversion frequency of individual nucleotides.

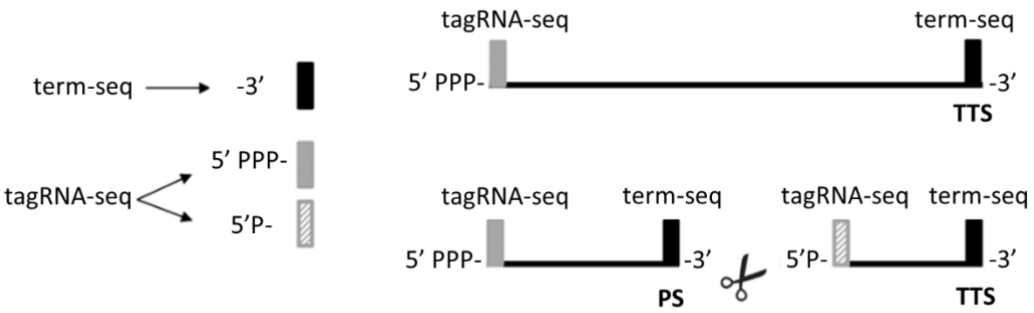

Fig. 1: Definitions of Transcription Termination Sites and Processing Sites. Term-seq enables the identification of RNA 3' ends (3', Black bar), while tagRNA-seq enables the distinction between tri-phosphorylated (5' PPP-, grey bar) and mono-phosphorylated (5' P-, hatched bar) transcripts. A 3' signal followed by a 5' P signal indicates a likely processed site (PS), while the remaining 3' signals are classified as true transcription termination sites (TTS).

To distinguish genuine TTS from PS, Term-seq signals followed within 50 nucleotides (to allow for some 3' trimming, see materials and methods) by a 5' tagRNA-seq signal (i.e. a 5' mono-phosphate) was defined as a PS, while Term-seq signals without PS downstream signals were defined as TTS (Fig. 1). Following this approach, we mapped a total of 2,567 TTS in the *M. tuberculosis* transcriptome (Supplementary Table 3).

After filtering 3' ends against PS as described above, we obtained 2,567 *bona fide* TTS, which at a glance could be separated into four different TTS profiles: a single, sharp peak (e.g. *sthA*, *eccA3*, *ilvB1*, see Fig. 2A), a cluster of peaks within a defined region, sometimes with one dominant peak (e.g. *glyA2*, *pe20*, *rpfA*, see Fig. 2B), multiple, low-intensity peaks throughout entire genes and/or operons (e.g. *Rv0298-0299*, *fixA*, *rplS*, see Fig. 2C) and finally overlapping sense and antisense peaks (e.g. *rpsT*, *gabD1*, *Rv3212*, see Fig.2D). Additional profiles can be found in Supplementary Fig. 1.

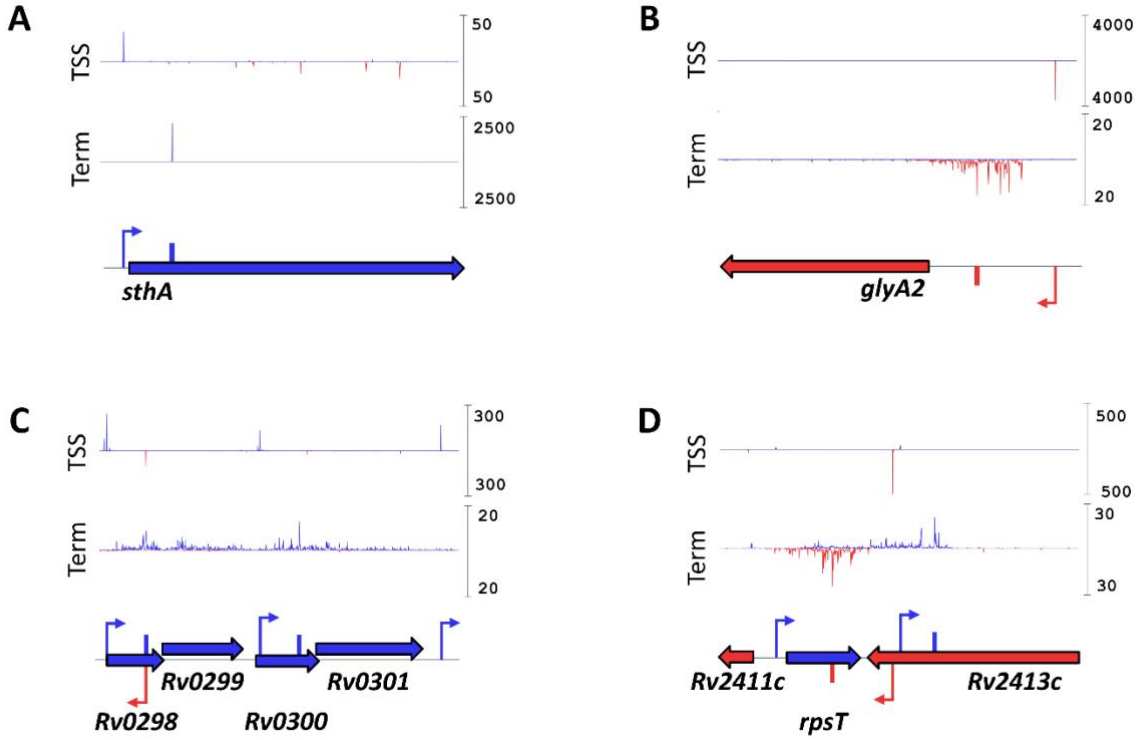

Fig. 2: Transcription termination site (TTS) profiles. TTS profiles followed four main patterns: individual, sharp peaks (A); a cluster of peaks within a defined region (B), multiple, low-intensity peaks covering entire genes or operons (C) and overlapping peaks (D). Each plot shows one representative region. Blue traces: Coverage on plus strand. Red traces: Coverage on minus strand. Horizontal Blue/Red arrows: ORF. Vertical Blue/Red bars: Mapped (dominant) TTS. Vertical arrows: Mapped TSS.

All TTS were classified according to their position relative to TSS and annotated open reading frames (ORFs), using the definitions from Adams *et al*. with minor modifications (Fig. 3)(21). Thus, Internal TTS are located within annotated Open Reading Frames (ORFs); Antisense TTS are found antisense to annotated ORFs (From TSS to Stop Codon of the ORF); Omega refers to the dominant TTS located within 500 nucleotides downstream of ORFs, (unless the region was interrupted by another TSS or ORF); Orphan are TTS located in intergenic regions, including those located in 5' UTRs.

According to these definitions, Orphan TTS represented the smallest fraction at only 7%, followed by Omega at 10% and Antisense at 18%. Two thirds of TTS, were classified as Internal, in other words, associated with translated regions. Furthermore, by mapping each Internal TTS relative to its cognate ORF size, we observed a significant enrichment of TTS within the first quarter of ORFs compared to the remaining three quarters, suggesting the presence of specific *cis*-regulatory elements (Supplementary Fig. 2, Supplementary Table 3).

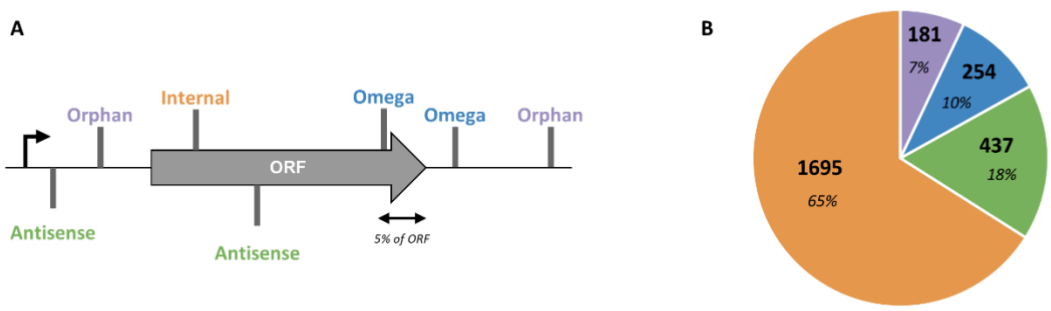

Fig. 3: TTS classification and distribution. (A) TTS were classified according to nearest TSS and annotated ORFs. Grey vertical bars indicate TTS. Black vertical arrow: TSS. (B) shows the number and proportion of each TTS class

We have previously reported that *M. tuberculosis* 3' UTRs can be several hundred nucleotides in length (7). To determine the distribution of 3' UTR lengths based on Term-seq data in *M. tuberculosis,* we calculated the distance from the stop codon to Omega TTS and plotted the frequency of each length. We observed that some of these 3' UTRs had a negative value, i.e. the TTS mapped upstream of stop codons. This could be due to internal termination sites, but more likely, this is due to trimming of the 3' end, as has been reported for *E. coli* (14).

The results could thus be divided into four populations: likely trimmed (3' UTR < 0; n=70), short (0 < 3' UTR < 50 nucleotides; n=60) long (50 ≥ 3' UTR < 200 nucleotides; n=58) and very long (3' UTR ≥ 200; n=21; Fig. 4). This indicates that *M. tuberculosis* does indeed have several long 3' UTRs with a potential for post-transcriptional regulation.

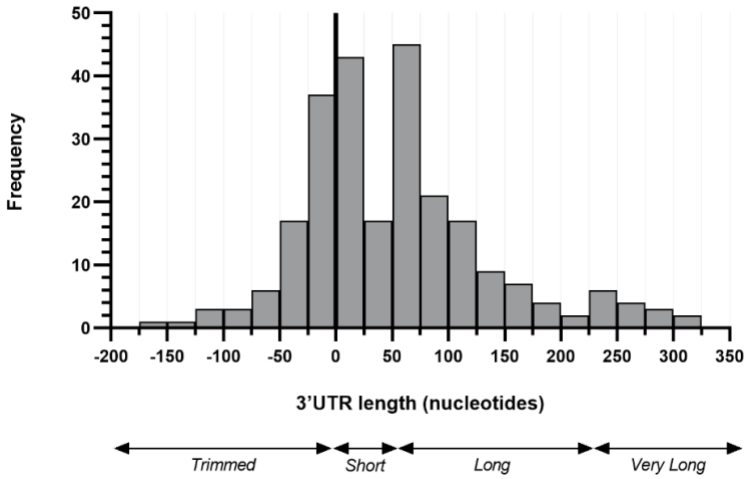

Fig. 4: Distribution of 3' UTR lengths. The length of 3' UTRs in nucleotides was plotted after calculating the distance from the Stop Codon of ORF to their Omega TTS. Grey bars: Number of 3' UTRs assigned. Black vertical line (0): End of Stop codon.

Intrinsic terminators are rare in *Mycobacterium tuberculosis*

M. tuberculosis is known for its paucity in canonical intrinsic terminators (ITs)(7,8,22). Instead, alternative motifs and/or mechanisms of termination such as Rho-dependent terminators or Rho-independent terminators with imperfect or no poly-U tails have been suggested to play a role (6,8,9,15). To gauge the relative contribution of individual mechanisms to transcription termination in *M. tuberculosis*, we investigated how our mapped TTS correlated with different types of predicted terminators.

TransTermHP (27), WebGeSTer DB (28) and RNIE (9) are algorithms used to predict canonical or 'L-shaped' intrinsic terminators, 'I-shaped' intrinsic terminators or Tuberculosis Rho-independent Terminators, respectively, and these all showed minimal overlap with *M. tuberculosis* TTS (2%-4%; Supplementary Fig. 3).

The most prominent overlap between prediction and our experimental data was found using RhoTermPredict (RTP), which uses a two-step approach to predict Rho-dependent terminators (12). The first step is identification of a 78-nt *RUT* site based on C:G ratio and regularly spaced C residues, while the second step searches for potential RNAP pause sites in the form of hairpin structures within 150 nucleotides downstream of the *RUT* site(12). After applying RTP with default parameters to the *M. tuberculosis* genome, we found that 1,879, i.e. 73% of our TTS mapped within this 228-nt window. However, as Rho-dependent termination rarely happens beyond 100 nucleotides downstream of *RUT* sites (13), we repeated the prediction where the downstream window was reduced to 100 nucleotides instead of the default 150 nucleotides (Supplementary Table 3). This nevertheless resulted in 1,480 TTS (58%) overlapping with predicted Rho-dependent terminators (Supplementary Fig. 3, Supplementary Table 3).

These results suggest that Rho-dependent termination of transcription is by far the dominant mechanism in *M. tuberculosis*.

Depletion of Rho affects the majority of *M. tuberculosis* TTS

The *M. tuberculosis rho* gene is essential, eliminating the option of a deletion strain. Moreover, *M. tuberculosis* Rho is resistant to bicyclomycin, which is often used to define the extent and role of Rho-dependent termination. To circumvent these issues, Botella *et al*. generated the RhoDUC strain, in which the addition of ATC leads to rapid depletion of Rho with significant changes in RNA pools after six hours (15). We wanted to understand where in the genome and to what extent individual Rho-dependent terminators contributed to these changes, and we hypothesised that such contributions

might be evident before six hours. Hence, we grew cultures of the *M. tuberculosis* RhoDUC strain (kindly provided by D. Schnappinger) and isolated RNA at times 0, 3 H, 4.5 H and 6 H after ATC addition to apply Term-seq and RNA-seq.

For each time point, the normalised coverage of the 2,567 TTS mapped in H37Rv was extracted in a window of 5 nucleotides (i.e. 2 nucleotides either side of the TTS). To ensure that termination in the RhoDUC strain at time 0 was a valid representation of termination in H37Rv, we first compared the TTS windows of the two and found that the main profiles were similar with only 80 TTS out of 2567 (3%) showing a $\log_2$ variance of more than 2 (Supplementary Fig. 4A). Conversely, a comparison of the Term-seq profiles at time 0 and 6 H showed a widespread variance occurring across the genome (Supplementary Fig. 4B).

Next, we calculated the fold change of mapped TTS from the uninduced state (time 0) to times 3 H, 4.5 H and 6 H for the 2,567 H37Rv-derived TTS (Fig. 5A). To our surprise, we found that while overall, the coverage of 41% of TTS was reduced ≥50% after 4.5 H, some TTS appeared more sensitive to Rho depletion than others (Fig. 5B). In other words, 13% of TTS displayed a ≥50% reduction after 3 H (leading), another 28% changed between 3.5 H and 4.5 H (trailing) and 35% changed between 4.5 and 6 H (final), totalling 76% of all TTS (Supplementary Table 4). This step-wise/gradual change might to some extent be explained by the nature of depletion leading to more noisy data than instantaneous inhibition (obtained by the addition of bicyclomycin). To better understand the basis for this time-dependent difference, we investigated Rho-sensitive TTS by a complementary approach, namely transcriptional readthrough (RT). RhoDUC RNA from time 0, 3 H, 4.5 H and 6 H was subjected to RNA-seq and a RT-score was calculated for each time point based on the ratio of reads upstream and downstream of each TTS (Fig. 5A and 5C). By this method, we observed fewer changes over time with only 18% TTS displaying increased readthrough after 4.5 H (Supplementary Table 5). However, the apparent difference in Rho-sensitivity between time points remained, suggesting a limited or selective effect of Rho depletion after 3 H.

Conditional termination of transcription in *M. tuberculosis*

Our results show that 626 (37%) of Internal TTS and 138 (76%) of Orphan TTS are essentially preventing transcription of complete or full-length mRNA. These likely represent conditional terminators of transcription often seen in the context of *cis*-regulatory elements such as RNA leaders and riboswitches(21,22,29). To investigate

conditional termination sites (CondTTS) further, we focused on all Orphan TTS as well as Internal TTS that fell within the first quarter of annotated ORFs, as we had found that this region was enriched for TTS (Supplementary Fig. 2). Moreover, we only included the dominant TTS for each region (see materials and methods for details). A total of 506 RNA leaders (123 associated with Orphan TTS and 383 with Internal TTS) fell into this category, representing 23% of all *M. tuberculosis* TTS and regulating 12% of annotated *M. tuberculosis* genes (Supplementary Table 6).

We extracted the time-dependent changes in CondTTS coverage listed in Supplementary Table 6, and calculated the fraction of leading, trailing and late TTS as before (Fig. 5E, Supplementary Table 4). The results indicated that 84% of CondTTS were Rho-sensitive across the entire time course, compared to 76% of all TTS (Fig. 5B and E). We also calculated the readthrough with the Readthrough method for the CondTTS. However, in the case of CondTTS, we took the leader length into consideration, and instead of comparing a fixed region of 100 nucleotides on either side of a TTS, we compared coverage within the actual RNA leader (i.e. from TSS to TTS) to the coverage of a downstream region of identical size following the CondTTS for each timepoint (Fig. 5D, Supplementary Table 5; details in materials and methods). By this method, 73% of CondTTS were Rho-sensitive across the time course compared to 42% of all TTS (compare Fig. 5C and Fig. 5F). These results indicate that Rho-dependent termination of transcription is even more pronounced for conditional termination than for the average *M. tuberculosis* termination event.

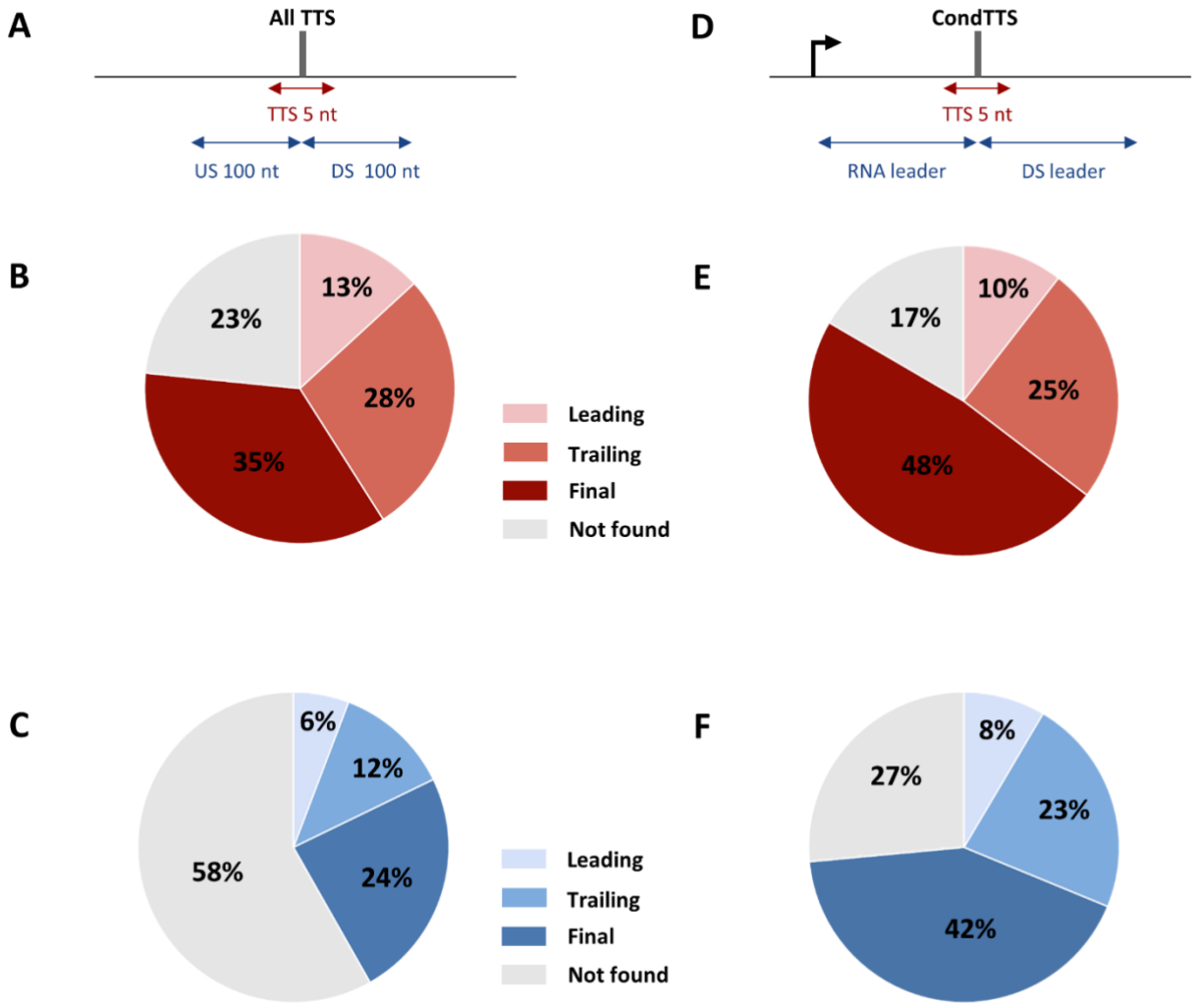

Fig. 5. Changes in TTS coverage and readthrough (RT) over time. A window of 5 nucleotides, represented by a red arrow, was used to defined a TTS score (A and D). A minimum of 50% decrease in normalised coverage from time 0 to times 3 H, 4.5 H and 6 H was calculated and classified as leading (pink), trailing (red) and final (burgundy), respectively for all TTS (B) and CondTTS (E). Two different approaches, using a 100 nucleotide window (A) or a TSS to CondTTS window size (D) were used to calculate a readthrough (RT) score represented with blue arrows. Statistically significant differences in RT, cumulated with an RT-score higher than 1.1 was calculated using Fisher's exact test between each time point. A significant score from time 0 to times 3 H, 4.5 H and 6 H was calculated and classified as leading (light blue), trailing (blue) and final (dark blue), respectively for all TTS (C) and CondTTS (F). Percentages are indicated for all TTS (B and C) and CondTTS (E and F). Grey bar: Mapped TTS. Black arrow: TSS.

Presumably RNA leaders require a certain length to accommodate regulatory elements including those that lead to conditional termination of transcription. To test this notion, we extracted the length of *M. tuberculosis* leaders from TSS to CondTTS and plotted these against their frequencies. The results indicated a median length of 171 nucleotides, with the majority (70%) having a length between 50 and 250 nucleotides and <4% being

shorter than 50 nucleotides (Fig. 6). This supports the notion that CondTTS tend to be located at a substantial distance from their cognate TSS within a leader that provides plenty of scope for post-transcriptional regulation (Fig. 6). Next, we investigated the distribution of RNA leaders ≥50 nucleotides in length across different functional gene categories in *M. tuberculosis* and found a significant enrichment within Cell Wall & Cell Processes, Information Pathways and PE/PPE genes (p-value<0.05) (Supplementary Fig. 5).

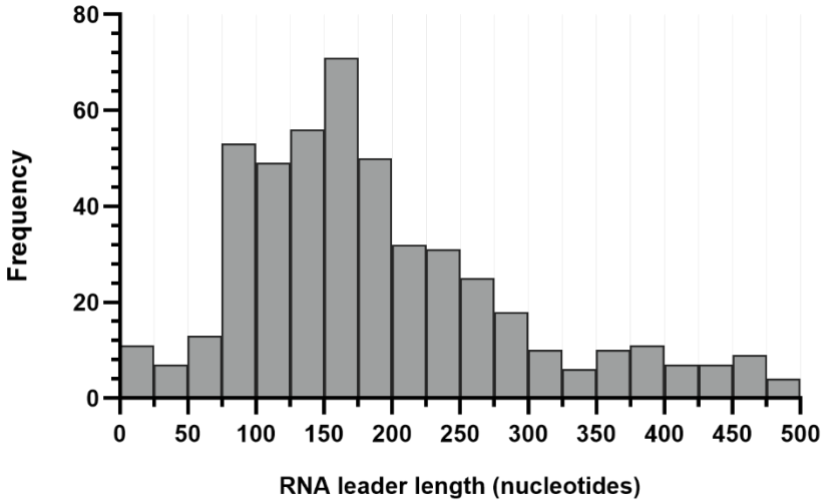

Fig. 6: Variation and distribution of RNA leader lengths. The length of the RNA leaders from the TSS to their respective conditional TTS was extracted and plotted. Each bar indicates the number of leaders within the specified 25-nucleotide window.

M. tuberculosis CondTTS associated with translated uORFs

We noticed that several Orphan TTS fell within small to medium ORFs located upstream of annotated ORFs. To gauge whether these uORFs might be translated thereby shifting Orphan TTS to Internal TTS, we mined data from Sawyer *et al*. to extract potentially translated regions (30). We focussed on Orphan (TTS associated) regions between TSS and annotated ORFs ('5' UTRs') that were potentially translated based on Ribo-seq coverage (≥250 reads), which were either leaderless or had a canonical Shine-Dalgarno (SD) (a minimum of 5 consecutive purines, 5-15 nucleotides upstream of a start codon, see materials and methods for details). Of the 123 such '5' UTRs', 67 had one or more likely translated uORFs, including some associated with known riboswitches (Supplementary Table 7). These results suggested that at least 35 out of the 123 Orphan CondTTS were in fact located Internal to uORFs.

This search also revealed that several uORFs overlapped with either an annotated ORF or with another uORF downstream, indicating translational coupling (31,32). For comparison, we performed a systematic search for overlaps across the *M. tuberculosis* genome. The results revealed that 776 (19%) of annotated ORFs overlap with another annotated ORF upstream, indicating that ORF overlaps are widespread in *M. tuberculosis*, as previously suggested (31). The most common configuration involved direct overlaps between stop and start codons, likely associated with a Termination-Reinitiation (TeRe) mechanism of translation (31). Of these, NUGA (where N denotes any nucleotide) four-nucleotide overlaps were almost five-fold more abundant than URAUG (where R denotes a purine) one-nucleotide overlaps (515 or 66% *versus* 110 or 14%, Fig. 7). Applying the same search to overlaps between uORFs and annotated ORFs or uORF-uORF pairs resulted in the same trend, i.e. four-nucleotide and one-nucleotide overlaps being the most and second-most abundant type, respectively (Supplementary Table 8). This indicates that likely functional ORFs sharing the active site within a ribosome are a common feature in *M. tuberculosis*.

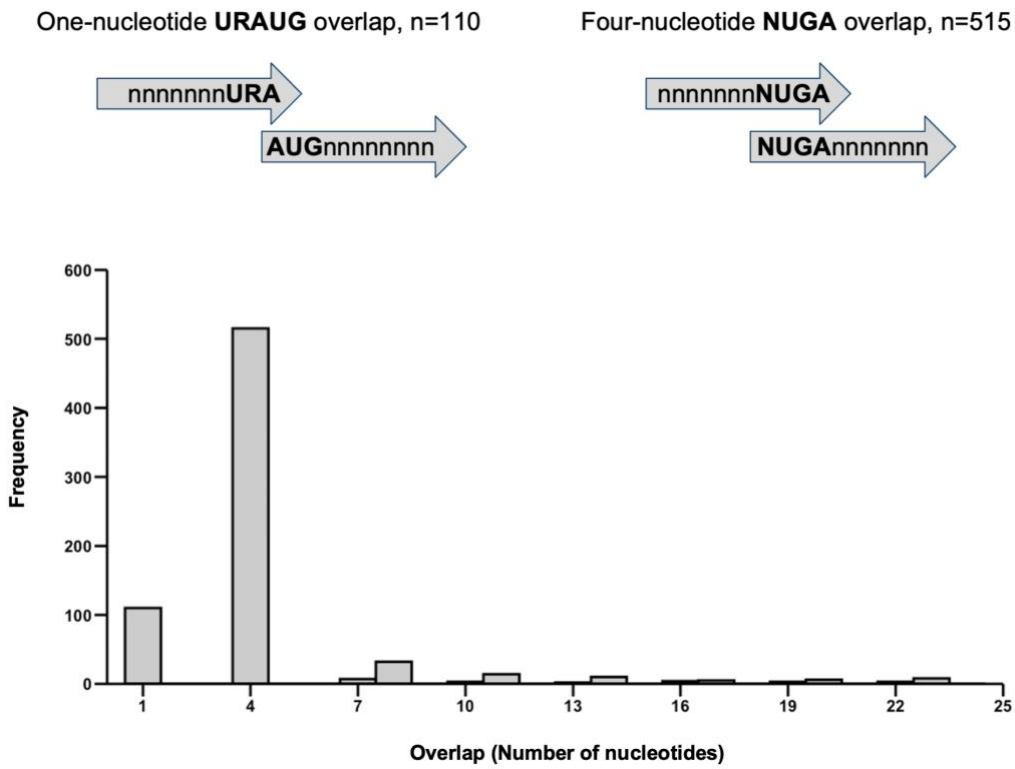

Fig. 7: *M. tuberculosis* overlapping ORFs. A search for overlaps between all *M. tuberculosis* ORFs annotated in Mycobrowser(33) indicated that 19% of all ORFs overlap between 1 and 25 nucleotides with an upstream ORF. The most abundant constellation was the four-nucleotide NUGA (N=any nucleotide) overlap followed by a one-nucleotide URAUG (R=purine) overlap; two- and five-nucleotide overlaps are not possible with conventional stop and start codons.

Like all *M. tuberculosis* ORFs, the initiating nucleotide of the second ORF was dominated by purines with only a minor fraction of UUG and no CUG starts. Curiously, almost two thirds (64%) of ORFs that started with a NUGA or URAUG overlap did not have a canonical SD (as defined above), although a weaker or non-canonical SD might be present. This is somewhat higher than suggested for Actinobacteria in general (31). Finally, we found that overlapping ORFs were particularly abundant in genes associated with virulence, detoxification and adaptation including several Toxin-Antitoxin (TA) modules; e.g. *relBE* and *relJK*, 28 of 48 *vapBC* modules, 2 of 9 *mazEF* modules, both *parDE* modules and *higAC* had NUGA overlaps (Supplementary Fig. 6, Supplementary Table 8).

Translation of *M. tuberculosis* uORFs and overlapping ORFs

Naturally, the presence of a Ribosome footprint does not necessarily mean that a given uORF is translated. Therefore, to probe whether selected uORFs were translated, we made in-frame *lacZ* fusions and expressed these in *Mycobacterium smegmatis*. We selected leaderless (*rne*), isolated (i.e. *pe20*, *Rv1535*, *ilvB1*) and overlapping (i.e. *dnaA*, *glyA2*, *rpfA*) uORFs. The results demonstrate that all of the selected uORF-*lacZ* fusions are translated in *M. smegmatis*, albeit to different extents but nevertheless indicating that the actual uORFs are translated in *M. tuberculosis* (Fig. 8).

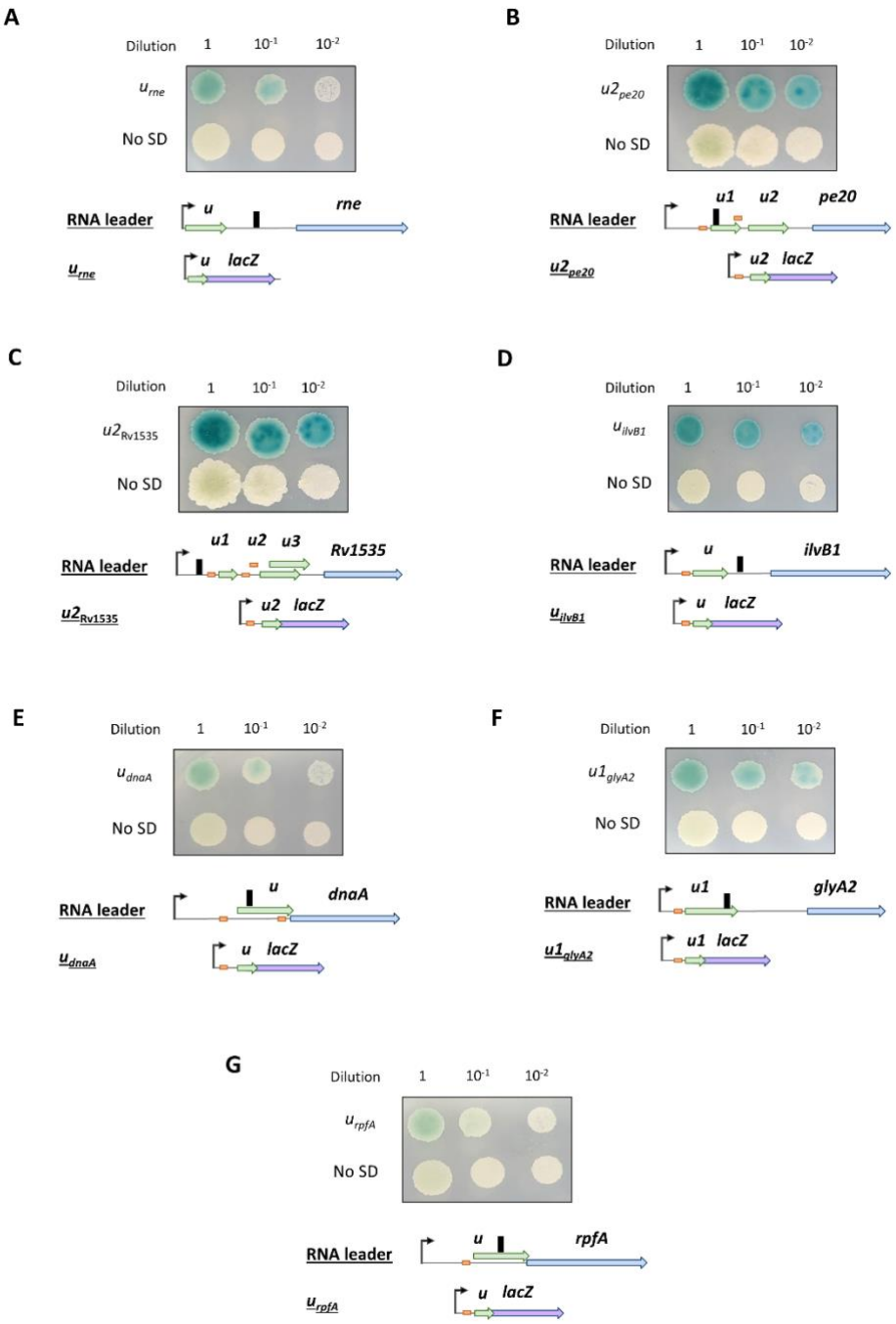

Fig. 8: Validation of uORF translation using in-frame *lacZ* fusions. Translation of selected uORF was validated by fusing the region spanning 20 nucleotides upstream of the identified SD to four codons into the uORF in-frame to *lacZ*. Black arrow: TSS. Blue arrows: Annotated ORF. Green arrow: Newly identified uORFs. Purple arrow: *lacZ*

This in turn indicates that the associated CondTTS were Internal rather than Orphan.

Interestingly, some of the identified uORFs resided within the aptamers of known riboswitches including the Glycine, Cobalamin and YdaO riboswitch aptamers,

respectively (Supplementary Fig. 7). We are currently investigating how translation affects riboswitch function and *vice versa*.

DISCUSSION

The aim of this study was to define at nucleotide resolution, where in the *M. tuberculosis* genome transcription termination happens and by which mechanism(s). To this end, we applied Term-seq to cultures of *M. tuberculosis* and found that i) Rho-dependent termination of transcription is the dominant mechanism across the genome; ii) conditional or premature termination of transcription regulates a large proportion of genes; iii) conditional terminators are often associated with translated uORFs; iv) closely overlapping ORFs are abundant and in some cases devoid of a recognisable SD sequence, indicating tightly coordinated/coupled translational control.

We observed four main types of TTS profiles of which the individual, sharp peak was the least common. This is possibly due to the nature of Rho-dependent termination, which is generally associated with more a more diffuse termination pattern than e.g. intrinsic terminators (14). Conversely, we observed an abundance of the low-level multiple peaks covering large regions (Fig. 2C). These are likely associated with the exposure of multiple *RUT* sites throughout the ORFs in question, and the profile was particularly obvious in large operons such as those encoding r-proteins. All TTS, but particularly CondTTS, likely associated with specific *cis*-regulatory elements such as riboswitches or protein-binding leaders, appeared to be dominated by Rho-dependent termination. Using TTS coverage and RT for all TTS and CondTTS we found in three out of four cases that Rho depletion affected between 73% and 83% of TTS, with estimates based on TTS coverage providing higher estimates for Rho-dependence than RT values (Fig. 5). These values indicate that the vast majority of *M. tuberculosis* TTS are Rho-dependent, further supported by the finding that 58%-73% of the TTS overlap with predicted Rho-dependent terminators (Supplementary Fig. 3). This fraction is significantly higher than for *E. coli* (>30%), although the study by Dar and Sorek refers to genes rather than TTS (14).

Our results indicate that >10% of all genes are controlled by *cis*-regulatory elements, i.e. 5' leaders with robust CondTTS. This is based on a conservative definition of leaders reaching no further than the first 25% of an ORF. However, we did observe several CondTTS further into ORFs e.g. within multi-*cistronic* operons, where they likely regulate expression of the downstream ORF(s).

Comparing *M. tuberculosis* TTS locations to those of *E. coli* TTS (21), highlights one major difference between the two species. Although Internal TTS (i.e. within ORFs) was by far the largest class in both organisms, we note a more than 2-fold difference in their relative abundance (68% in *M. tuberculosis versus* 29% in *E. coli*). At the same time, the TTS flanking either end of ORFs, i.e. Orphan (5') and Omega (3') represented a much lower fraction in *M. tuberculosis* than in *E. coli* (7% *versus* >20%). Together, this suggests that *M. tuberculosis* TTS tend to be more associated with coding regions than *E. coli* TTS. This could to some extent be ascribed to differences in ORF annotations and data analysis, but the fact that the proportion of Antisense TTS was almost identical in the two species indicates that the difference reflects actual, biological variation. The prominent role of Rho combined with a lack of intrinsic terminators in *M. tuberculosis* likely contribute to this disparity, as Rho-dependent termination is generally considered to be linked to translation. Moreover, the reduced motor function of *M. tuberculosis* Rho (16), could potentially lead to some degree of readthrough from untranslated (Orphan) into coding (Internal) regions. Finally, although we did re-assign some Internal TTS to Omega TTS, it is possible that post-termination 3' trimming varies due to variation in 3' structures and the complement of nucleases in the two species (14,34).

Our results further suggested that many of the identified leaders were coding, in line with previous findings (35,36), and indicating that conditional termination of transcription is in many cases linked to translation. While we did find some commonality between our proposed leader peptides and those recently reported by Smith *et al*. (36) (i.e. 9 exact matches and 28 isoforms, Supplementary Table 7), it appears from the number of peptides that did not match, that the methods of Ribo-seq/Ribo-ret and Term-seq are complementary in identifying potentially translated, and perhaps in particular, regulatory uORFs. It remains unclear at this stage whether translated uORFs are regulatory or encoding peptides with *trans* functions or both. It also seems likely that regulatory *versus trans* acting would be represented differently by the two approaches, but at least in some cases, the levels of expression and amino acid conservation indicate functional peptides (unpublished).

We found that overlapping ORFs, including uORFs, are common in *M. tuberculosis* and that many of the downstream ORFs are devoid of canonical signals for translation initiation, suggesting translational coupling (Kipkorir *et al*. manuscript in preparation). It has been suggested that such coupling can be via 70S scanning, which requires a SD (32) or via TeRe, which may involve 30S as well as 70S ribosomes and not necessarily

requiring an SD (31). A quarter of *M. tuberculosis* ORFs are leaderless (26) i.e. initiating translation with 70S ribosomes, indicating a propensity towards maintaining ribosomes in their assembled state. We therefore imagine that the 'handover' from one ORF to a downstream ORF is likely to involve 70S ribosomes, which may be a means of saving energy in addition to coordinating gene expression.

Similar to transcription units or operons, *M. tuberculosis* appears to have translational units where two or even more ORFs overlap, and where expression of an ORF may to some degree depend on the translation of its upstream ORF. Such prominence of translational control is in line with previous findings of an extensive leaderless transcriptome, pervasive translation and uORF-dependent gene expression (26,30,35-37). However, it remains to be firmly established whether transcription and translation are coupled, as in *E.coli* (38) or uncoupled as in *Bacillus subtilis* (39). The picture that emerges is, that a significant fraction of *M. tuberculosis* gene expression control is post-transcriptional in the form of Rho-dependent conditional termination of transcription, translation units or a combination of these. Although there is still much to be learnt, this study adds to our understanding of a pathogen that does not conform to the images of model organisms.

MATERIALS AND METHODS

Cultures plasmids and clonings

Strains and plasmids used in this study are listed in Supplementary Table 9.

pIRATE2020 was made by replacing the region in pIRATE (25) between Xho I and Hind III with a region harbouring a shorter polylinker, slightly modified promoters driving the expression in both directions and an intrinsic terminator between the two promoters (See Supplementary Table 10 for the inserted sequence)

M. tuberculosis H37Rv and *M. smegmatis* MC^2 155 were cultured on Middlebrook agar 7H11 supplemented with 10% OADC (Sigma), 0.5% Glycerol and 50 µg/ml hygromycin if appropriate and in liquid Middlebrook 7H9 supplemented with 10% ADC (Sigma), 0.5% Glycerol, 0.05% Tween 80 and 50 µg/ml hygromycin where appropriate. Cultures were harvested at an OD_{600nm} ~0.6 for exponential phase.

M. tuberculosis RhoDUC was grown as previously described (15) with 50 µg/ml Hygromycin, 20 µg/ml Kanamycin and 50 µg/ml Zeocin. When the cultures reached an

OD_{600nm}~0.6, depletion of Rho was induced using 500 ng/ml Anhydrous Tetracycline (ATC) and cells were harvested after 0 H, 1.5 H, 3 H, 4.5 H and 6 H.

Escherichia coli DH5α was used for all clonings and were cultured on solid LB 1.5% agar or in liquid LB supplemented with 250 µg/ml Hygromycin. pIRATE plasmids, listed in table 10, were constructed using Gibson assemblies with oligos (Sigma) or geneBlocks (IDT) listed in Supplementary Table 10. Generally, inserts spanned the region from 20 basepairs upstream of the presumed SD sequence and 20 basepairs upstream to the first four-six amino acids of the tested uORF. Once plasmids with the desired sequence had been identified, these were transformed into *M. smegmatis by* electroporation.

RNA isolation and purification

M. tuberculosis and *M. smegmatis* RNA was extracted as previously described (7,40). Briefly, cells were cold shocked by adding 30% ice directly to cultures, followed by centrifugation at 5000 rpm for 10 minutes at 4°C and finally RNA extraction using the FastRNA Pro Blue Kit (MP Biomedicals) following the instruction of the supplier. RNA concentration and purity was assessed using Nanodrop 2000 (ThermoFisher). Residual genomic DNA was removed using Turbo DNase (ThermoFisher) according to manufacturer's instructions, followed by extraction with phenol-chloroform and ethanol precipitation. The RNA was subsequently checked for DNA by PCR using RedTaq readymix (Sigma). RNA integrity was assessed using 2100 Bioanalyzer (Agilent) before library preparation and RNA sequencing.

Library preparation and RNA sequencing

Library construction and sequencing were handled by Vertis Biotechnology AD (https://www.vertis-biotech.com/home), where all details are available.

Libraries were sequenced on an Illumina NextSeq 500 using 1x75 basepair read length. RNA sequencing quality control was assessed using FastQC. Sequences were mapped to the genome of *M. tuberculosis* AL123456.3 using Bowtie2 (41) and reads mapped more than one were discarded using Samtools (42). Coverage was extracted using Bedtools (43). For each dataset, the coverage was normalised to counts per million. All RNA-seq traces were visualised using Artemis software.

Data are accessible on ArrayExpression with the accession number E-MTAB-11753.

Transcription Start Site (TSS) and Processed Site (PS) mapping

Three biological replicates of *M. tuberculosis* H37Rv were harvested at OD_{600nm}~0.6, and total RNA sequenced with tagRNA-seq (24). The TSS and PS were extracted as below. For the TSS mapping, the coverage of the 5' end of reads was extracted using Bedtools for 5' tri-phosphate RNA fractions. The coverage at each position was normalised to counts per million and a geometric mean calculated from the three biological replicates. A threshold of 50 normalised reads was used to extract peaks at a single nucleotide level and were merged in a window of 3 nucleotides. Each TSS was assessed to the first downstream feature in a window of 500 nucleotides and compared to data from Cortes *et al.* (26). This resulted in 59 new TSS that were added to the data from Cortes *et al*. and used in the screening of RNA leaders described below.

For PS mapping, coverage at the 5' end of reads from the 5' mono-phosphate RNA fractions was extracted using Bedtools and normalised to counts per million. The geometric mean across all three replicates? was calculated at every position and a background noise threshold, determined using regions thought to be underrepresented by processed signal, was applied. The coverage in intergenic regions was extracted for each position, the geometric mean calculated and a median value of 1.27 was used as cut-off for background noise. 5' peaks higher or equal to this value were assigned as PS. This resulted in the identification of 57,755 PS that were used to filter termination sites as described below.

Transcription Termination Site (TTS) mapping

M. tuberculosis RNA used for tagRNA-seq were also sequenced by Term-seq and transcription TTS extracted as below.

Read coverage was extracted using Bedtools, normalised to counts per million, and the geometric mean calculated for each genomic position across all three replicates. A defined threshold was calculated using a similar approach to that of PS. We assumed that within an ORF, TTS will be primarily associated with either the start (for conditional termination) or end (for RNA 3' trimming) of the ORF, except for multiple low-intensity TTS peaks covering entire ORFs (see Fig. 2C). Following this rationale, each annotated ORF was divided in three equal length and the middle third of each ORF used to screen for background coverage levels. Following this analysis, a median of 4.80 was used as background cut-off for TTS mapping.

Bedgraph files were generated using deeptools (44) on each strand separately for each biological replicates. The three biological replicates were used together to run the package termseq_peaks (21) to extract peaks from our Term-seq data with default parameters. The coverage at the 3' end of reads, normalised to counts per million, was used to define the strongest nucleotide position corresponding to the identified peak. The presence of PS sites within a 50-nucleotide window downstream of the newly identified 3'ends was checked using Bedtools. A signal falling within this window was classified as processed site, while absence of clear PS signal within this window allowed the identification of a TTS.

TTS were classified according to their genome position and localisation compared to features using bedtools. A TTS located within an annotated ORF was classified as Internal. A TTS located in antisense of a gene, from its assigned TSS or start codon to the stop codon, was classified as Antisense. The strongest TTS located within 500 nucleotides downstream of a gene was classified as Omega. Internal TTS located within the last 5% of an annotated ORF were considered as trimmed TTS and re-assigned as Omega. The other TTS located in intergenic regions and between a TSS and an ORF were classified as Orphan. In case a TTS was located between two annotated ORF, and a TSS associated to the downstream ORF is located within the upstream gene, the TTS was classified as Orphan (see Figure 3).

Computational prediction of transcriptional terminators

Experimentally identified TTS genomic positions were compared to predicted terminators using software based on different parameters. TransTermHP (http://transterm.cbcb.umd.edu) predicts L-shaped intrinsic terminators, based on a stem-loop, flanked by an upstream A-tail and a downstream T-tail within 15 nucleotides.(27) WebGeSTer DB (http://pallab.serc.iisc.ernet.in/gester/index.html) predicts L-, I-, U-, X- and V-shaped intrinsic terminators based on the best stable structure downstream of a stop codon.(28) RNIE is based on covariance using the Infernal package, which allows the discovery of the only conserved motif of terminators in *M. tuberculosis*, TRIT (Tuberculosis Rho-Independent Terminator) (9). RhoTermPredict (RTP) searches for consensus *RUT* sites with a high C:G ratio followed by a palindromic sequence (stem-loop) for RNA polymerase pausing (12). RTP was run with default parameters and subsequently with reduced downstream window as described with all predictions listed in Supplementary Table 3.

TTS score and Readthrough score after Rho depletion.

Two biological replicates of *M. tuberculosis* RhoDUC strain were harvested before and 3 H, 4.5 H and 6 H after the onset of Rho-depletion. Total RNA was sequenced with RNA-seq and Term-seq methods.

Using Term-seq alignment, the 3' end coverage was extracted for each biological replicate, normalised to counts per million and the geometric mean calculated at each genomic position. The TTS profiles were first compared between *M. tuberculosis* H37Rv and RhoDUC strain. The reads coverage of the TTS was extracted in a window of 5 nucleotides (2 nucleotides on either side of the extracted TTS position from H37Rv), and the values compared between the two strains. The $\log_2$ variation between RhoDUC and H37Rv was calculated and MA plotted. The TTS were considered highly variable between the two strains if the $\log_2 \geq 2$ or $\log_2 \leq -2$.

TTS scores at time 0 and 3 H, 4.5 H and 6 H after Rho depletion were calculated in a similar manner. The extracted coverage in a 5-nucleotide window was directly compared from time 0 to times 3 H, 4.5 H and 6 H after Rho-depletion. The TTS ratio was considered significant if the value was $\geq$ 1.5 (Representing a 50% decrease in expression). TTS were classified as "leading" if TTS_0:$TTS_3 \geq 1.5$; "trailling" if TTS_0:$TTS_3 < 1.5$, while TTS_0:$TTS_{4.5} \geq 1.5$ and "final" if TTS_0:$TTS_{4.5} < 1.5$, while TTS_0:$TTS_6 \geq 1.5$.

A readthrough (RT) score (RT-score) was calculated for each time point from RNA-seq data by extracting read counts in a window of 100 nucleotides upstream (US) and downstream (DS) of each identified TTS using htseq-counts and the "union" parameter (21). A ratio of DS over US was used for each time point and in replicate to normalise the RT. Finally, a ratio of ratio was calculated to get an accurate estimation of the increase in RT at a specific time (X) over previous time points (P). The formula used is as follows:

$$RT - score(X) = \frac{DS\ (X)/US\ (X)}{DS\ (P)/US\ (P)}$$

Using the average reads counts extracted at each time point, a Fisher's exact test was performed for each RT-score, and RT-scores > 1.1 and with an associated p-value < 0.05 were considered Rho-dependent.

A more tailored approach was conducted for CondTTS. The 100-nucleotide window was replaced by a specific window for each identified RNA leader, which spanned the region from the TSS to the CondTTS, and a same length nucleotide window downstream of the

TTS. Using these RNA leader-specific windows, RT-scores and Fisher's exact tests were performed as described above.

Identification of CondTTS and RNA leaders

A total of 764 CondTTS were identified, which 138 were classified as Orphan (TTS falling outside annotated ORFs, but after the assigned TSS) and 626 as Internal (TTS located within the first 25% of an annotated ORF). We specifically searched for CondTTS where a TSS was associated to the downstream annotated gene. In total, 588 (77%) of them were found associated with a TSS from our data. In case of multiple TTS associated within one RNA leader, the strongest signal was used in the rest of the analysis, and the others (82) were reclassified as secondary CondTTS. Proximity of TTS, TSS and Annotated ORFs was calculated using Bedtools.For functional gene category enrichment analysis associated with RNA leaders, a hypergeometric test was used and functional gene categories with p-values < 0.05 were considered significant.

uORF identification and re-classification of Orphan TTS

To identify potentially translated upstream ORFs (uORFs), we mined Ribo-seq data from Sawyer *et al*. (30) with focus on RNA leaders. Adapter sequences were trimmed using cutadapt, and reads aligned to the AL123456.3 genome using Bowtie2 (41). Coverage at each nucleotide position was assessed using Bedtools, and normalised to counts per million. Ribosome footprints within our identified RNA leaders were extracted using SciPy (45). To filter out low-confidence peaks, an arbitrary minimum threshold of 250 reads in each biological replicate was applied. Of the 123 Orphan TTS, 109 had a significant Ribo-seq signal either upstream or downstream of the CondTTS (Supplementary Table 7).

To identify Shine-Dalgarno(SD)-associated uORFs, for each of the RNA leaders-extracted windows, the find-boxes function from the segmentation-fold tool available on the Galaxy Webserver (https://usegalaxy.eu/), was used to search for potential SD sequences (SDs) and associated start codons.

For SDs, all regions of 5 consecutives purines found in RNA leaders that were associated with an Orphan TTS were extracted and marked as potential SDs. Next, NTG start codons were identified and the distance to the potential SD determined. An ORF was annotated as likely being translated if the distance between the 3' edge of the SD-sequence and the start codon was between 5 and 15 nucleotides. Finally, the window

covering the SD and the start codon was compared to the extracted Ribo-seq peaks and the uORF of interest identified.

A similar approach was designed to identify potential leaderless uORFs in RNA leaders associated with an Orphan TTS. ORFs located no more than 5 nucleotides downstream of TSS were considered likely leaderless. Ribo-seq peaks within these 5-nucleotide windows were assigned as leaderless uORFs. Predictions were filtered with already annotated ORF using Bedtools (43), and with a minimal ORF length of 5 amino acids.

From the 123 Orphan TTS, a new search was conducted using Bedtools (43) to find TTS localised within uORFs. The Orphan TTS found in those windows were re-assigned as uORF-Internal.

Identification of overlapping ORFs in *M. tuberculosis*:

Data analysis for identification of overlapping annotated ORFs was performed in R. The code used to generate the results and the corresponding figures are available at https://github.com/ppolg/nuga_overlaps. Briefly, overlapping ORFs were identified by checking the annotated start and stop nucleotides of *M. tuberculosis* H37Rv ORFs based on the latest release of the MycoBrowser annotation (Release 4 (2021-03-23)) (33). ORFs were identified as overlapping if the 3' nucleotide of the ORF is both downstream to the following ORF's 5' end and upstream to the following ORF's 3' end. Overlap length for picking 4-nucleotide (NUGA) overlaps was calculated using genomic coordinates. SDs for the purpose of picking 4-nucleotide overlaps without an SD were identified using the same criteria employed above. Genomic coordinates of the overlaps were used to obtain nucleotide sequences of the overlap, using the H37Rv reference genome FASTA file from NCBI. Overlapping ORFs were grouped by functional category based on the Mycobrowser annotation. Plots were drawn using ggplot2 (46).

Translational reporter gene fusions

In-frame *lacZ* fusions were made in pIRATE2020 (Supplementary Table 10 and 11) by cloning inserts between the Hind III and Nco I sites, and all were expressed in *M. smegmatis*. Spotting assays were done with 5 µl *M. smegmatis* at OD_{600}~0.6 at the dilutions indicated. Beta-gal assays were performed in triplicate on cultures as described previously (47). Briefly, cultures were cooled on ice prior to centrifugation, pellets were washed with cold Z-buffer before disrupting cells in a FastPrep instrument (MP Bio). Extracts were cleared by centrifugation and an aliquot of each was used to determine the total protein concentration using a BCA protein assay kit (Pierce).

Beta-mercaptoethanol was added to the remainder and beta-galactosidase activity was determined and normalised to total protein concentration.

FUNDING

KBA is funded by the UK Medical Research Council grant number [MR/S009647/1], TC by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No. 637730) and TK by a Newton International Fellowship NIF\R5A\0035 and by Wellcome Institutional Strategic Support Fund 204841/Z/16/Z

ACKNOWLEDGEMENTS

The authors are grateful to Dirk Schnappinger at Weill Cornell for providing the RhoDUC strain, and to Finn Werner at UCL for critical reading of the manuscript

REFERENCES

1. Baker, J.J., Dechow, S.J. and Abramovitch, R.B. (2019) Acid Fasting: Modulation of Mycobacterium tuberculosis Metabolism at Acidic pH. *Trends Microbiol*, **27**, 942-953.
2. Ehrt, S., Schnappinger, D. and Rhee, K.Y. (2018) Metabolic principles of persistence and pathogenicity in Mycobacterium tuberculosis. *Nat Rev Microbiol*, **16**, 496-507.
3. Mashabela, G.T., de Wet, T.J. and Warner, D.F. (2019) Mycobacterium tuberculosis Metabolism. *Microbiol Spectr*, **7**.
4. Schnappinger, D., Ehrt, S., Voskuil, M.I., Liu, Y., Mangan, J.A., Monahan, I.M., Dolganov, G., Efron, B., Butcher, P.D., Nathan, C. *et al.* (2003) Transcriptional Adaptation of Mycobacterium tuberculosis within Macrophages: Insights into the Phagosomal Environment. *J Exp Med*, **198**, 693-704.
5. Flentie, K., Garner, A.L. and Stallings, C.L. (2016) Mycobacterium tuberculosis Transcription Machinery: Ready To Respond to Host Attacks. *J Bacteriol*, **198**, 1360-1373.
6. Unniraman, S., Prakash, R. and Nagaraja, V. (2001) Alternate paradigm for intrinsic transcription termination in eubacteria. *J Biol Chem*, **276**, 41850-41855.
7. 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. *et al.* (2011) Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of Mycobacterium tuberculosis. *PLoS Pathog*, **7**, e1002342.
8. Czyz, A., Mooney, R.A., Iaconi, A. and Landick, R. (2014) Mycobacterial RNA polymerase requires a U-tract at intrinsic terminators and is aided by NusG at suboptimal terminators. *mBio*, **5**, e00931.
9. Gardner, P.P., Barquist, L., Bateman, A., Nawrocki, E.P. and Weinberg, Z. (2011) RNIE: genome-wide prediction of bacterial intrinsic terminators. *Nucleic Acids Res*, **39**, 5845-5852.
10. Schwenk, S., Moores, A., Nobeli, I., McHugh, T.D. and Arnvig, K.B. (2018) Cell-wall synthesis and ribosome maturation are co-regulated by an RNA switch in Mycobacterium tuberculosis. *Nucleic Acids Res*, **46**, 5837-5849.
11. Ju, X., Li, D. and Liu, S. (2019) Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. *Nat Microbiol*, **4**, 1907-1918.

12. Di Salvo, M., Puccio, S., Peano, C., Lacour, S. and Alifano, P. (2019) RhoTermPredict: an algorithm for predicting Rho-dependent transcription terminators based on Escherichia coli, Bacillus subtilis and Salmonella enterica databases. *BMC Bioinformatics*, **20**, 117.
13. Ray-Soni, A., Bellecourt, M.J. and Landick, R. (2016) Mechanisms of Bacterial Transcription Termination: All Good Things Must End. *Annu Rev Biochem*, **85**, 319-347.
14. Dar, D. and Sorek, R. (2018) High-resolution RNA 3'-ends mapping of bacterial Rho-dependent transcripts. *Nucleic Acids Res*, **46**, 6797-6805.
15. Botella, L., Vaubourgeix, J., Livny, J. and Schnappinger, D. (2017) Depleting Mycobacterium tuberculosis of the transcription termination factor Rho causes pervasive transcription and rapid death. *Nat Commun*, **8**, 14731.
16. Saridakis, E., Vishwakarma, R., Lai-Kee-Him, J., Martin, K., Simon, I., Cohen-Gonsaud, M., Coste, F., Bron, P., Margeat, E. and Boudvillain, M. (2022) Cryo-EM structure of transcription termination factor Rho from Mycobacterium tuberculosis reveals bicyclomycin resistance mechanism. *Commun Biol*, **5**, 120.
17. Dar, D. and Sorek, R. (2017) Regulation of antibiotic-resistance by non-coding RNAs in bacteria. *Curr Opin Microbiol*, **36**, 111-117.
18. Breaker, R.R. (2012) Riboswitches and the RNA world. *Cold Spring Harb Perspect Biol*, **4**.
19. Dar, D., Shamir, M., Mellin, J.R., Koutero, M., Stern-Ginossar, N., Cossart, P. and Sorek, R. (2016) Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science*, **352**, aad9822.
20. Lee, J.H., Lee, E.J. and Roe, J.H. (2022) uORF-mediated riboregulation controls transcription of whiB7/wblC antibiotic resistance gene. *Mol Microbiol*, **117**, 179-192.
21. Adams, P.P., Baniulyte, G., Esnault, C., Chegireddy, K., Singh, N., Monge, M., Dale, R.K., Storz, G. and Wade, J.T. (2021) Regulatory roles of Escherichia coli 5' UTR and ORF-internal RNAs detected by 3' end mapping. *Elife*, **10**.
22. Arnvig, K.B. (2019) Riboswitches: choosing the best platform. *Biochem Soc Trans*, **47**, 1091-1099.
23. Schwenk, S. and Arnvig, K.B. (2018) Regulatory RNA in Mycobacterium tuberculosis, back to basics. *Pathog Dis*, **76**.
24. Innocenti, N., Golumbeanu, M., Fouquier d'Herouel, A., Lacoux, C., Bonnin, R.A., Kennedy, S.P., Wessner, F., Serror, P., Bouloc, P., Repoila, F. *et al.* (2015) Whole-genome mapping of 5' RNA ends in bacteria by tagged sequencing: a comprehensive view in Enterococcus faecalis. *RNA*, **21**, 1018-1030.
25. Moores, A., Riesco, A.B., Schwenk, S. and Arnvig, K.B. (2017) Expression, maturation and turnover of DrrS, an unusually stable, DosR regulated small RNA in Mycobacterium tuberculosis. *PLoS One*, **12**, e0174079.
26. Cortes, T., Schubert, O.T., Rose, G., Arnvig, K.B., Comas, I., Aebersold, R. and Young, D.B. (2013) Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in Mycobacterium tuberculosis. *Cell Rep*, **5**, 1121-1131.
27. Kingsford, C.L., Ayanbule, K. and Salzberg, S.L. (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol*, **8**, R22.
28. Mitra, A., Kesarwani, A.K., Pal, D. and Nagaraja, V. (2011) WebGeSTer DB--a transcription terminator database. *Nucleic Acids Res*, **39**, D129-135.
29. Millman, A., Dar, D., Shamir, M. and Sorek, R. (2017) Computational prediction of regulatory, premature transcription termination in bacteria. *Nucleic Acids Res*, **45**, 886-893.
30. Sawyer, E.B., Phelan, J.E., Clark, T.G. and Cortes, T. (2021) A snapshot of translation in Mycobacterium tuberculosis during exponential growth and nutrient starvation revealed by ribosome profiling. *Cell Rep*, **34**, 108695.
31. Huber, M., Faure, G., Laass, S., Kolbe, E., Seitz, K., Wehrheim, C., Wolf, Y.I., Koonin, E.V. and Soppa, J. (2019) Translational coupling via termination-reinitiation in archaea and bacteria. *Nat Commun*, **10**, 4006.

32. Yamamoto, H., Wittek, D., Gupta, R., Qin, B., Ueda, T., Krause, R., Yamamoto, K., Albrecht, R., Pech, M. and Nierhaus, K.H. (2016) 70S-scanning initiation is a novel and frequent initiation mode of ribosomal translation in bacteria. *Proc Natl Acad Sci U S A*, **113**, E1180-1189.
33. Kapopoulou, A., Lew, J.M. and Cole, S.T. (2011) The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis (Edinb)*, **91**, 8-13.
34. Durand, S., Tomasini, A., Braun, F., Condon, C. and Romby, P. (2015) sRNA and mRNA turnover in Gram-positive bacteria. *FEMS Microbiol Rev*, **39**, 316-330.
35. 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. *et al.* (2015) Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLoS Genet*, **11**, e1005641.
36. Smith, C., Canestrari, J.G., Wang, A.J., Champion, M.M., Derbyshire, K.M., Gray, T.A. and Wade, J.T. (2022) Pervasive translation in Mycobacterium tuberculosis. *Elife*, **11**.
37. Canestrari, J.G., Lasek-Nesselquist, E., Upadhyay, A., Rofaeil, M., Champion, M.M., Wade, J.T., Derbyshire, K.M. and Gray, T.A. (2020) Polycysteine-encoding leaderless short ORFs function as cysteine-responsive attenuators of operonic gene expression in mycobacteria. *Mol Microbiol*, **114**, 93-108.
38. Wang, C., Molodtsov, V., Firlar, E., Kaelber, J.T., Blaha, G., Su, M. and Ebright, R.H. (2020) Structural basis of transcription-translation coupling. *Science*, **369**, 1359-1365.
39. Johnson, G.E., Lalanne, J.B., Peters, M.L. and Li, G.W. (2020) Functionally uncoupled transcription-translation in Bacillus subtilis. *Nature*, **585**, 124-128.
40. Arnvig, K.B. and Young, D.B. (2009) Identification of small RNAs in Mycobacterium tuberculosis. *Mol Microbiol*, **73**, 397-408.
41. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **9**, 357-359.
42. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.
43. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
44. Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A. and Manke, T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res*, **42**, W187-191.
45. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*, **17**, 261-272.
46. Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.
47. Arnvig, K.B., Gopal, B., Papavinasasundaram, K.G., Cox, R.A. and Colston, M.J. (2005) The mechanism of upstream activation in the rrnB operon of Mycobacterium smegmatis is different from the Escherichia coli paradigm. *Microbiology (Reading)*, **151**, 467-473.