

Harnessing toxin-mediated ribosome stalling as a complementary tool to annotate bacterial ORFs

Eduardo A Troian^{1,†}, Valdir C Barth^{1,2,†}, Unnati Chauhan¹, Haiyan Zheng³, Caifeng Zhao³, Jumei Zeng², Robert N Husson², Nancy A Woychik^{1,*}

¹Department of Biochemistry and Molecular Biology, Rutgers University, Robert Wood Johnson Medical School, Piscataway, NJ 08854, United States

²Division of Infectious Diseases, Department of Pediatrics, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, United States

³Biological Mass Spectrometry Facility of Robert Wood Johnson Medical School and Rutgers, Rutgers University, Piscataway, NJ 08854, United States

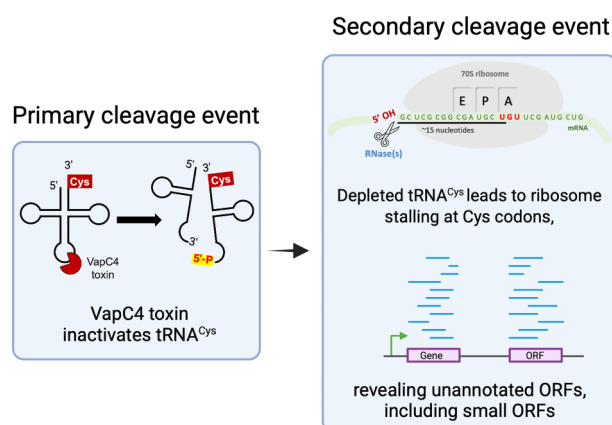
*To whom correspondence should be addressed. Email: nancy.woychik@rutgers.edu

† The first two authors should be regarded as Joint First Authors.

Abstract

The *Mycobacterium tuberculosis* (Mtb) VapC4 endoribonuclease toxin exclusively cleaves and inactivates tRNA^{Cys}, which leads to extensive ribosome stalling at Cys codons. Serendipitously, the precise position of stalled ribosomes is revealed within our 5' RNA-seq datasets used to identify and validate the tRNA target of the toxin, precluding the need for Ribo-seq. Here we show how mapping of stalled ribosomes can be harnessed as an innovative tool for reliable detection of new Cys-containing Mtb open reading frames (ORFs). Using proteogenomics we unmasked 96 unannotated ORFs; of which 54% are small ORFs ≤ 50 amino acids. We validated 69% of the 96 ORFs by mass spectrometry, including four whose spectra was matched to synthetic controls. Also, 25% of these unannotated ORFs were identified by previously published Ribo-RET. Some of the 96 ORFs are Cys-responsive attenuators or encode stable Cys-containing proteins that map immediately before, or within, genes in the opposite, or same, orientation. These ORF sequences can also reveal functional clues, e.g. zinc-binding motifs or encode novel EsxB-like proteins. Our findings demonstrate that toxin-mediated ribosome stalling can serve as a robust genome annotation tool that is applicable to mycobacteria and other bacteria, with unique advantages that complement existing genome annotation methods.

Graphical abstract



Introduction

The burst in the release of bacterial whole genome sequences has required increasingly sophisticated genome annotation programs, such as the Reference Sequence (RefSeq) project at the National Center for Biotechnology Information (NCBI) [1], the first tool enlisted to define gene numbers and bound-

aries. Bioinformatic annotation is especially difficult for detection of small open reading frames (sORFs; here defined as ≤ 50 amino acids) and their “small protein” products. For example, the RefSeq annotation pipeline uses a 40 amino acid cut-off for ORF consideration [1]. Consequently, specialized programs have been developed for sORF detection [2–5]. These

Received: May 14, 2025. Revised: February 27, 2026. Accepted: March 4, 2026

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License

(<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other

permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

difficulties are exacerbated in mycobacteria [6, 7], whose messenger RNAs (mRNAs) can use one of several alternate translation start codons, AUG, GUG, UUG, and AUU for leadered, versus AUG or GUG for leaderless transcripts [7]. Thus, bioinformatic identification of bacterial genes and the translation start sites of their transcripts has limitations, most profoundly for prediction of sORFs in mycobacteria.

Earlier genetic and biochemical analyses, sometimes assisted by bioinformatic approaches, also uncovered irrefutable evidence for the existence of previously overlooked small proteins in bacteria (reviewed in [8, 9]). Conserved sORFs with possible regulatory functions were then reported in mycobacteria [7]. Others used targeted sORF searches coupled with identification of orthologs in other bacteria to strengthen the idea that these sORFs were functionally relevant [10]. The prevalence of sORFs in bacteria was further solidified upon computational identification of >4000 conserved, yet predominantly novel, sORFs from human microbiome sequences derived from multiple anatomical sites [11].

The existence of abundant sORFs in bacteria (reviewed in [12]) can also be supported by genome-scale approaches coupled with much improved proteomic methods. First, RNA-seq can be used to confirm active transcription of the predicted sORF. For example, Sberro *et al.* compared subsets of the human microbiome sORFs with orthologs in metatranscriptomic datasets to assess whether they were actively transcribed [11]. Second, ribosome profiling by Ribo-seq enables detection of ribosome footprints that suggest that the sORF is being translated (reviewed in [13–15]). Third, antibiotics or antibacterial peptides that trap ribosomes at start codons enable even finer mapping of translation start sites. Ribo-seq with rapamulin (Ribo-RET) or the antimicrobial peptide Onc112 are used to trap initiating bacterial ribosomes at start codons [16–18]. Finally, mass spectrometry is typically used to establish translation of unannotated open reading frames (ORFs) to a protein (reviewed in [5, 19]).

In this study, we report that an Mtb transfer RNA-specific ribonuclease (tRNase) that inactivates tRNA^{Cys} leads to widespread Cys-codon specific ribosome stalling. This feature can be enlisted as a tool to uncover ORFs that is mechanistically distinct from other methods used to date. The complementary features of toxin-mediated ribosome stalling can be exploited in conjunction with existing technologies to identify new ORFs and improve the accuracy of Mtb genome annotation. This general method is applicable to other mycobacteria, e.g. *Mycobacterium smegmatis* and *Mycobacterium abscessus*; it can also be adapted for *Escherichia coli*, other model organisms, and other clinically important bacterial pathogens.

Materials and methods

Strains, plasmids, and reagents

All experiments were performed using Mtb mc² 6206 (Δ panCD Δ leuCD), generously provided by the Williams Jacobs laboratory, Albert Einstein College of Medicine, and Mtb strain H37Rv (ATCC 25618). VapC4 (Rv0595c locus) was amplified by PCR from genomic DNA and cloned into the low-expressing vector pMC1s under the control of an anhydrotetracycline (ATc)-inducible promoter. Induction of VapC4 was carried out by the addition of ATc to a final concentration of 200 µg/ml every 48 h.

Cells were grown under constant shaking at 170 revolutions per minute (rpm) at 37°C in 7H9 Middlebrook media containing 1× OADC (Sigma), 0.05% tyloxapol, and 25 µg/ml kanamycin (for plasmid selection). For the growth of the attenuated strain Mtb mc² 6206, the media was supplemented with additional 50 µg/ml of pantothenic acid and 100 µg/ml of L-leucine.

RNA isolation

Total RNA was harvested from cells grown in the presence or absence of ATc for 24 or 72 h. Cells were centrifuged at 2000 × *g* for 10 min at 4°C, and the pellets were resuspended in 1 ml of Trizol and transferred to 2 ml lysing kits tubes (Bertin Corp.) containing 0.1 mm glass beads. Cells were lysed with a Precellys Evolution homogenizer (Bertin Corp.) using four cycles of 30 s with the agitation set to 9000 rpm with 1 min cooling between each cycle. Lysate was centrifuged at 16 000 × *g* for 5 min at 4°C, and total RNA was isolated using the Direct-zolTM RNA MiniPrep Plus kit (Zymo Research). Residual genomic DNA was removed by treating RNA with 1 U of TURBOTM DNase (Thermo Fisher) for 30 min at 37°C twice. RNA was re-purified using the RNA Clean and Concentrator kit (Zymo Research), and concentration was determined by spectrophotometry using a µCuvette in a BioSpectrometer (Eppendorf).

5'-OH RNA-seq

Preparation and analysis of 5'-OH libraries was performed as described in Barth *et al.* [20]. Libraries were sequenced in an Illumina HiSeq 2500/4000 platform at Genewiz Corp or New York University's Genome Technology Center. For data analysis, only reads with at least 1 Read Per Million (RPM) for mRNAs and 5 RPM for transfer RNAs (tRNAs) in the induced sample were considered. Frequency logos were generated with WebLogo [21].

Unannotated ORF search

The search of unannotated ORFs was performed on 5'-OH datasets obtained from Barth *et al.* [22], Troian *et al.* [23] (5'-P dataset; Supplemental Table 4), and new 5'-OH RNA-seq datasets for (i) VapC4 expressed in H37Rv for 24 and 72 h and (ii) VapC4 expressed in mc² 6206 for 24 and 72 h (Supplemental Table 1). We searched for unannotated ORFs with either one of the two Cys codons (UGC or UGU) 14, 15, or 16 nts from the RNA cleavage site (corresponding to the stalled ribosome at the “hungry” codon). We then determined if the Cys-containing sequence was in a translation frame comprising at least 10 amino acids with a start (AUG, GUG, UUG, and AUU) and a stop (UGA, UAA, or UAG) codon (according to Shell *et al.* [7]). We selected the longest possible ORF since transcription start sites (TSS) for these ORFs are not known except for instances where shorter start sites were clear by Ribo-RET and/or Ribo-seq and mapped in the Wadsworth interactive genome browser <http://mtb.wadsworth.org> [7, 24] or based on functional data [22, 25].

Inferring TSS of unannotated ORFs using Ribo-RET

To validate unannotated Mtb ORFs identified by 5'-OH RNA-seq, we reannotated the TSS based on the Ribo-RET data from Smith *et al.* 2022 [24]. Using an in-house script, we mapped the stop codon of the unannotated ORFs to the ORFs

identified through Ribo-RET. In case of a positive match, the start codon was reannotated to match the data from Smith *et al.*, otherwise, we maintained the predicted ORF based of 5'OH RNA-seq.

Unannotated ORFs false discovery rate calculation

To estimate the false discovery rate (FDR) for the 96 ORFs, we used a similar approach by Smith *et al.* (2022) [24]. First, we determined the probability of having the first nt of a cysteine codon at position 14, 15, or 16 nt downstream of a random genomic coordinate by randomly picking 100 000 genome coordinates and determining the fraction associated with a Cys codon, the variable 'R'. Next, we determined the number of true positives (the variable 'O'). These are defined as the array of sequences identified by 5' RNA-seq that contain a Cys codon 14, 15, or 16 nt downstream of the RNase cut site 5' of the stalled ribosome of all annotated plus unannotated ORFs. Note that our O is higher than the number of distinct transcripts identified by 5'-OH RNA-seq because here O represents the number of stalling events that fit our parameters; some transcripts have multiple stalled ribosomes. Factoring in the total number of sequences 14, 15, or 16 nt downstream of an RNase cut site regardless of whether there is a Cys codon there or not (including *any* RNase cleavage site from random RNase activity) from our 5'-OH RNA-seq datasets (the variable 'I'). We then applied the equation below to calculate the FDR:

$$\frac{(100 \times (I - O) \times (\frac{R}{1-R}))}{O}$$

Mass spectrometry analysis

Mtb *mc2 6206* and Δ vapBC4 Mtb *mc2 6206* were labeled with L-azidohomoalanine (AHA) for detection of newly synthesized proteins and grown to an OD₆₀₀ of 0.1 before induction for 48 and 72 h. Mtb H37Rv was grown without labeling and induced for 72 h prior to harvesting. For all three strains, fifty ml cultures were centrifuged at 2000 × g at 4°C for 10 min and washed with 1× phosphate buffered saline twice to remove traces of the albumin-containing 7H9 media. Pellets were then resuspended in lysis buffer (2% CHAPS, 8 M urea) and lysed using Precellys Evolution homogenizer (described in the 'RNA isolation' section). Lysates were pelleted at 12 000 × g at 4°C for 10 min. AHA-labeled proteins were selectively captured using alkyne-coated agarose beads from the Click-iT™ Protein Enrichment Kit (Thermo Fisher), following manufacturer's protocol. Unlabeled proteins were harvested using a trizol chloroform approach adapted from Hummon *et al.* [26].

Tryptic digests were analyzed using an Orbitrap Eclipse Tribrid Mass Spectrometer and nano LC system (Thermo Fisher), as described in Barth *et al.* [20]. Raw LC-MS data were converted to mzML format using MSConvert (v3.0.25343) and searched using a modified Mtb database (Uniprot Taxon ID: 83332) containing common laboratory contaminants from the CRAP (Common Repository of Adventitious Proteins) database (<https://www.thegpm.org/crap/>) and the novel ORFs identified in this study. Database searches were performed using a local implementation of SearchGUI (v4.3.17) [27] and PeptideShaker (v3.0.12) [28]. Identification settings were as follows: Trypsin, Specific, with a maximum of two missed cleavages, 10.0 ppm as MS1 and 20.0 ppm as MS2 tolerances; fixed modifications: Carbamidomethyla-

tion of C (+57.021464 Da), variable modifications: Acetylation of protein N-term (+42.010565 Da), Deamidation of N (+0.984016 Da), Deamidation of Q (+0.984016 Da), Dioxidation of M (+31.989829 Da), Oxidation of M (+15.994915 Da), Phosphorylation of S (+79.966331 Da), Phosphorylation of T (+79.966331 Da), Phosphorylation of Y (+79.966331 Da), variable modifications during refinement procedure: Pyrolydione from E (−18.010565 Da), Pyrolydione from Q (−17.026549 Da), Pyrolydione from carbamidomethylated C (−17.026549 Da). An FDR was set to 1% for Peptide Spectrum Matches, peptide and protein, and matches were validated using the target-decoy hit distribution.

Small protein validation using mass spectrometry

To validate the unannotated Mtb proteins identified by 5' OH RNA-seq, we searched for the estimated masses of their trypsin digestion products in publicly available Mtb proteomic datasets (retrieved from the PRIDE Archive, www.ebi.ac.uk/pride). The public dataset accession numbers used for this purpose were: PXD003842, PXD004165, PXD005290, PXD006039, PXD006389, PXD008555, PXD009239, PXD010929, PXD011466, PXD012584, PXD013677, PXD015680, PXD017004, PXD018957, PXD022644, PXD035082, PXD039671, PXD050258, PXD050763, PXD052312, PXD052936, PXD055726, PXD060534, PXD064766, PXD065910, PXD068926, PXD025047, PXD040615, PXD057673, PXD025774 along with two in-house proteomic datasets PXD071737 and PXD071729. Peak lists obtained from tandem mass spectrometry (MS/MS) spectra were identified using both X! Tandem VENGEANCE 2015.12.15.2 [29] and MS-GF+ 2024.03.26 [30]. The search was conducted using SearchGUI version 4.3.17 [27] and PeptideShaker 3.0.12 [28] for datasets acquired in data-dependent acquisition mode, and DIA-NN 1.9 [31] for datasets acquired in data-independent acquisition mode.

Protein identification was conducted against a concatenated target/decoy version of the Mtb H37Rv database (Uniprot Taxon ID: 83332), complemented with all unannotated proteins found in this study and a list of common contaminants (<https://www.thegpm.org/crap/>). Decoy sequences were created by reversing the target sequences in SearchGUI [27]. The identification settings used were the same as described in the section above. For data analyzed using DIA-NN, precursor- and protein-level identifications were filtered at a q-value threshold of 0.01 (corresponding to a 1% estimated FDR) using the built-in target-decoy approach. The validation strength increases when peptides are identified multiple times and/or in more than one dataset (Supplemental Table 2). Four novel ORF-derived peptides were chemically synthesized (Alan Scientific). Synthetic peptides were subjected to liquid chromatography-tandem mass spectrometry (LC-MS/MS) as described above. MS/MS spectra from synthetic peptides were compared with theoretical fragment ions and with experimentally observed peptide spectra from database searches. A maximum mass tolerance of 20 ppm was applied when matching spectra to the theoretical fragment ions.

RNA-seq and Ribo-seq data acquisition and handling

The FASTQ files from previous studies [24, 32] were obtained from NCBI Sequence Read Archive (SRA) number

SRP063670 for *Mycobacterium abscessus* and the European Nucleotide Archive numbers E-MTAB-8039 and E-MTAB-10695 for Mtb. Adapters were trimmed using Cutadapt (v5.1 [33]; Mtb dataset) or Trim Galore! (v0.6.10; *M. abscessus* datasets). Reads were then mapped to the *Mycobacterium abscessus* ATCC 19977 genome or *Mycobacterium tuberculosis* H37Rv using Bowtie2 [34]. Genome-wide coverage for both RNA-seq and ribosome profiling was computed using DeepTools (v3.5.6) [35] and normalized to RPM. For Ribo-seq, coverage was calculated based on the extracted 3'-end positions of mapped reads.

Pairwise Ribo-seq and RNA-seq coverage analysis

Feature-level RNA-seq and Ribo-seq abundance was quantified as reads per kilobase per million (RPKM) using FPKM Count from the RSeQC package (v5.0.4, [36]). For both RNA-seq and Ribo-seq, RPKM values were averaged from two replicates. Data were plotted using a custom python script.

Results

Patterns indicating the presence of a stalled ribosome are revealed within 5'-OH RNA-seq datasets

Toxin-antitoxin (TA) systems represent one class of molecular switches thought to aid in Mtb stress survival [37, 38]. These TA system switches are triggered when the toxin is freed from the antitoxin, leaving the toxin free to act on its target [39]. Through the study of Mtb TA system toxins we serendipitously discovered that ribosome stalling can be detected at single nt resolution within specialized RNA-seq datasets when a single tRNA isoacceptor is depleted by a TA toxin [22, 20]. For example, the VapC4 TA toxin is an endoribonuclease that exclusively targets tRNA^{Cys} by recognition of a defined consensus sequence within the proper structural context [22, 40, 41]. Using a battery of genome-scale approaches, we previously demonstrated that this VapC4 tRNase activity leads to reprogramming of Mtb metabolism to defend against stress pathways essential for viability of this pathogen during infection [22]. One of these approaches was a specialized 5'-end-dependent RNA-seq method developed in our laboratory that enables us to create selective libraries of Mtb RNAs with either 5'-hydroxyl (5'-OH) or 5'-monophosphate (5'-P) ends [42]. VapC4, as with all VapC endoribonuclease toxins, generates an RNA product with a 5'-P upon cleavage [43]. Consequently, 5'-P RNA-seq libraries prepared from control versus VapC4-expressing Mtb cells enabled us to identify the VapC4 target as the sole tRNA^{Cys} (servicing both UGU and UGC Cys codons; Fig. 1A) [22]. As a control, we routinely prepare parallel 5'-OH RNA-seq libraries with the same input RNA (with or without toxin expression) to ensure the toxin target is evident only in the 5'-P datasets.

However, VapC4 5'-OH RNA-seq datasets serve as more than a control for target identification; they also fortuitously validate the identity of the tRNA target [22]. This is because they also contain signatures of transcriptome-wide ribosome stalling at Cys codons, an expected downstream consequence of tRNA^{Cys} depletion (Fig. 1A) [22]. This stalling signature was revealed by the consistent, conspicuous presence of an in-frame Cys UGU or UGC codon (red) ~15 nt (we considered all Cys codons starting exactly 14, 15, or 16 nt downstream) from the 5'-OH cleavage site (dotted line in front of capi-

talized green nt, Fig. 1B) in transcripts preferentially cleaved upon VapC4 expression (Fig. 1B,C).

We used the following rationale to explain, then confirm, the pattern displayed in our 5'-OH RNA-seq datasets. Since the bacterial ribosome footprint is ~28 nt [44] and the distance from the P-site to the 3' end of this footprint is 15 nts [44], we posited that this ~15 nt gap represents the ribosome footprint spanning its 5' edge to the "hungry" Cys codon at the A-site (illustrated in Fig. 1D). This prediction was confirmed by Ribo-seq; we documented preferential transcriptome-wide stalling at the codon serviced by the depleted tRNA [20]. Therefore, 5'-OH RNA-seq both validates the tRNA targeted by the specific Mtb VapC toxin (identified in the 5'-P RNA-seq dataset) and provides high resolution mapping of stalled ribosomes without the extra effort and caveats associated with Ribo-seq methods.

Note that the endoribonuclease activity of VapC4 is not responsible for these 5'-OH cleavages (dotted line, Fig. 1B) but instead the result of an unknown Mtb RNase that generates a 5'-OH upon cleavage. VapC4 leaves a 5'-P and requires a precise cleavage consensus sequence in proper structural context [22, 40]. Instead, there is no sequence preference at the 5'-OH cleavage site (Supplemental Fig. 1). Finally, we have not been able to identify the enzyme(s) responsible for this cleavage on the 5' side of the stalled ribosome in mycobacteria. While we could detect stalling in *E. coli* by ectopic expression of the mycobacterial gene encoding RNase J (Rv2752c), its deletion in mycobacteria did not disrupt stalling in 5'-OH datasets [20]. If a functionally redundant partner to RNase J exists in mycobacteria, it does not share clear significant sequence similarity. The *E. coli* enzyme SmrB cleaves mRNAs upstream of stalled ribosomes [45]. However, the existence of an SmrB ortholog in mycobacteria remains elusive since no proteins in *M. smegmatis* or Mtb share clear sequence similarity to SmrB.

Ribosome stalling at Cys codons uncovers unannotated sORFs/ORFs

Our ability to definitively map the presence and precise position of stalled ribosomes on mRNAs within the VapC4 5'-OH RNA-seq dataset can also be exploited as a powerful tool to identify transcripts actively undergoing translation. As expected, we detected hundreds of stalled ribosomes on annotated mRNAs, corresponding to 341 distinct transcripts within the VapC4 5'-OH RNA-seq dataset [22]. The ribosomes in these 341 unique transcripts were stalled 14, 15, or 16 nt downstream at either of the two Cys codons, UGC and the UGU, serviced by the single Cys tRNA (tRNA^{CysGCA}) in Mtb. While there were no examples of ribosomes stalled at a Cys codon beginning 17 nt from the 5' RNase cleavage site, ribosome stalling at Cys codons 12 nt from the cleavage site occurred 15% of the time.

Most bacterial sORFs are abundant in intergenic regions. In fact, recent bioinformatic analysis of 5668 genomes in the family Enterobacteriaceae led to the identification of 67 297 clusters of intergenic sORFs [46]. Within our datasets we also noted that many ribosomes stalled at Cys codons did not map to an annotated ORF (Fig. 2B). We mapped the position of the stalled ribosomes relative to annotated Mtb H37Rv genes and determined that of the 96 "unannotated" transcripts harboring a stalled ribosome at Cys codons UGC and the UGU 91 were not annotated and five were recently annotated with an Rv number followed by an "A"—the sORF Cys-response

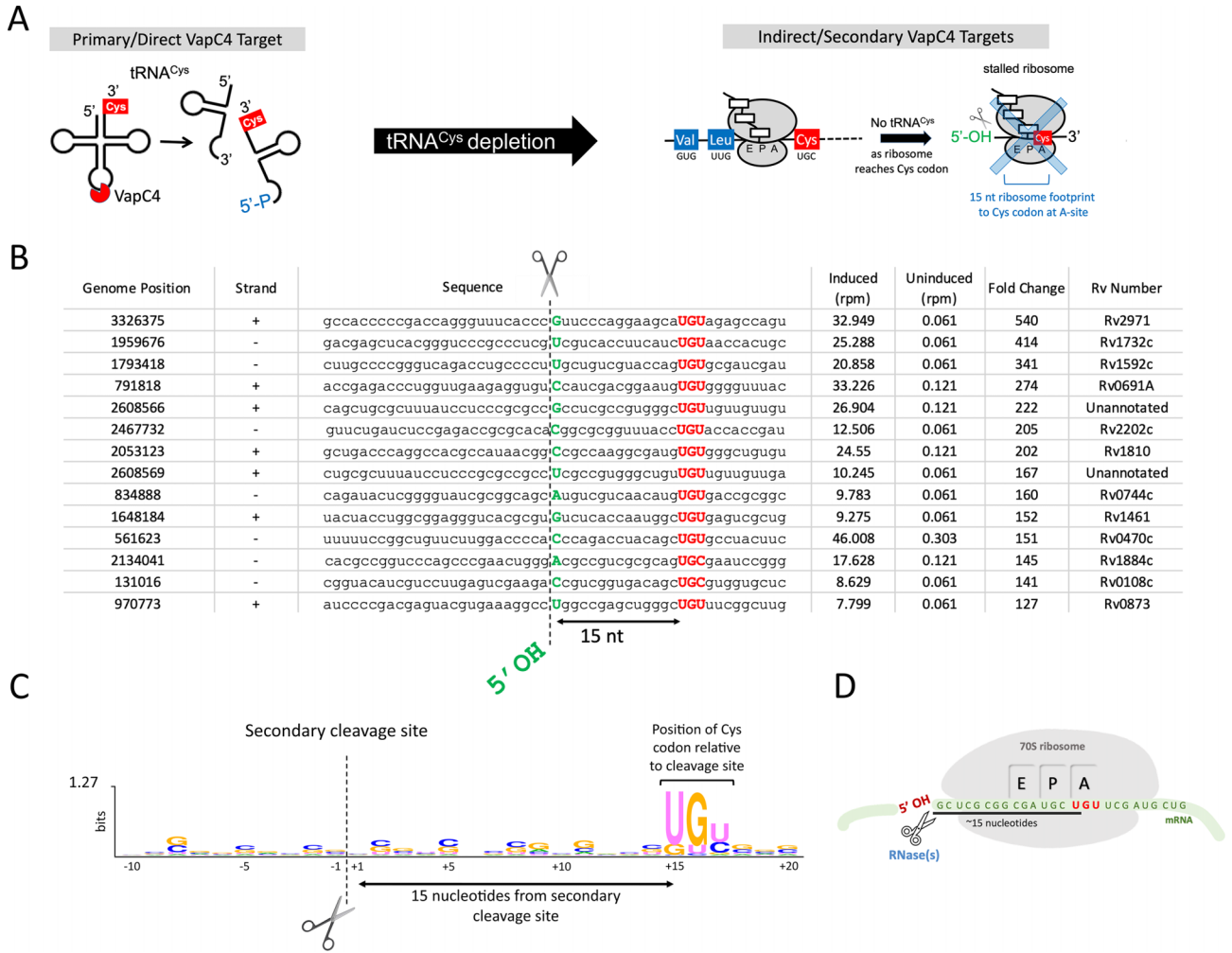


Figure 1. Depletion of tRNA^{Cys} by VapC4 leads to ribosome stalling at Cys UGU and UCG codons. **(A)** VapC4-mediated cleavage of tRNA^{Cys} leads to transcriptome-wide ribosome stalling at cysteine codons. **(B)** mRNA hits in descending order of fold-change relative to the control in the 5'-OH RNA-seq libraries constructed from Mtb mc² 6206 RNA extracted after 24 h of VapC4 induction. Cysteine codons are in bold red, ~15 nucleotides downstream of the 5'-OH cleavage site (5' of the green capitalized letter). Genome position and strand shown along with transcript Rv number. **(C)** WebLogo [21] from the top 100 mRNAs identified by 5'-OH RNA-seq. Positions are numbered relative to the 5'-OH cleavage sites. **(D)** Detailed illustration of mRNA stalled at hungry Cys codon at the A-site showing cleavage by an unspecified recycling RNase ~15 nts upstream.



Figure 2. VapC4-mediated ribosome stalling uncovers nearly 100 unannotated ORFs. **(A)** Pie chart of the 437 Cys codon stalled hits from the 5'-OH RNA-seq dataset from Fig. 1B displayed as annotated versus unannotated. The FDR was 15.2%, essentially identical to the 15% FDR reported for the 555 unannotated ORFs reported by Smith *et al.* [24]. **(B)** Unannotated mRNA hits in descending order of fold-change relative to the control in the 5'-OH RNA-seq from Fig. 1B. Cysteine codons are in bold red, ~15 nucleotides downstream of the 5'-OH cleavage site (5' of the green capitalized letter). Genome position and strand shown. Unannotated Cys codon-containing ORFs were identified as follows: (i) Cys codons with a stalled ribosome by 5'-OH RNA-seq were considered in-frame, (ii) minimum ORF length was set to ≥ 10 amino acids, (iii) start codons AUG, GUG, UUG, or AUU [7] and a stop UGA, UAA, or UAG codon were required to define an ORF. We selected the longest possible ORF since TSS for these ORFs are not known except for instances where shorter start sites were clear by Ribo-RET and/or Ribo-seq and mapped in the Wadsworth interactive genome browser <http://mtb.wadsworth.org> [7, 24] or based on functional data [22, 25].

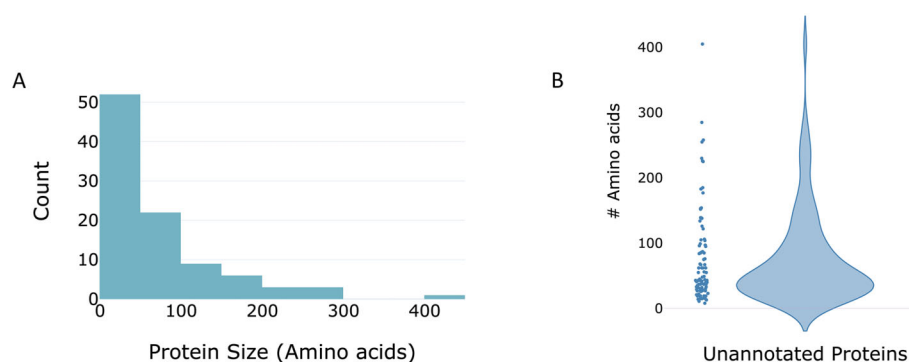


Figure 3. Most unannotated ORFs are ≤ 100 amino acids. Histogram, A, and violin plot, B, showing size distribution of the 96 unannotated ORFs.

attenuators Rv2334A, Rv0815A, Rv0485A, and Rv2391A [7, 22, 25], and the 134 amino acid Rv2742A validated by mass spectrometry [47] (Fig. 2A and Supplemental Table 1). Many (77%) of these 96 transcripts encode proteins < 100 amino acids (Fig. 3A and B). Of the unannotated transcripts 54% (52 of 96) were derived from sORFs encoding transcripts ≤ 150 nts and proteins ≤ 50 amino acids (Fig. 3A and B), following the convention for small proteins [48] and consistent with the < 150 nt cutoff used for mycobacteria sORFs [25]. The FDR of the 96 ORFs was calculated as 15.2% (using the equation in the ‘Materials and methods’ section), comparable to the 15% FDR reported for the 555 unannotated ORFs reported by Smith *et al.* [24].

Finally, using the Wadsworth interactive genome browser (<http://mtb.wadsworth.org>) constructed from published Mtb RNA-seq, Ribo-seq, and Ribo-RET data [6, 7, 24] we identified 12 unannotated leaderless ORFs; most of these leaderless transcripts encode small proteins ≤ 50 amino acids. Four of these leaderless intergenic sORFs (Supplemental Table 1) control Cys attenuation-responsive regulons [22]; one of these four is illustrated in Fig. 4A. In addition, we detected signatures of stalled ribosomes on leaderless mRNAs in a variety of orientations that were distinct from Cys attenuation-responsive regulons (Fig. 4B and C). One unannotated leaderless Cys-containing ORF maps immediately upstream of another ORF, with no intervening region, as illustrated in Fig. 4B. Another novel leaderless ORF overlaps with two other annotated genes on the opposite strand (Fig. 4C).

We could not clearly establish if many of the other unannotated ORFs were leadered or leaderless based on the data posted on the Wadsworth interactive genome browser [7, 24]. Nevertheless, Fig. 4D–F illustrates three examples of typical mapped locations of unannotated ORFs: upstream of a gene in the opposite orientation (Fig. 4D), overlapping a gene in the opposite orientation (Fig. 4E) and overlapping a gene in the same orientation (Fig. 4F). There are multiple examples of each configuration in Fig. 4A–F among the 96 unannotated transcripts harboring a stalled ribosome at Cys codons. The legitimacy of these unusual ORF configurations illustrated in Fig. 4 is supported by Smith *et al.* [24]. Using Ribo-RET and Ribo-seq without antibiotics, the authors detected similar ORF locations/orientations among higher-confidence Mtb ORFs [24]. Overall, our ability to detect ribosome stalling within actively translated transcripts from annotated genes (in agreement with the findings of Smith *et al.* [24]) suggests that the Mtb genome has adapted to expand its protein coding potential by engaging an abundance of unconventional ORFs.

Ribo-RET and/or mass spectrometry validation of a subset of ORFs harboring ribosomes stalled at Cys codons

While RNA-seq data indicate whether a possible ORF is transcribed, actual ribosome binding provides more reliable evidence that the putative ORF is translated into a protein. Although our 5'-OH RNA-seq data enables us to map stalled ribosomes to Cys codons—which strongly suggests translation—the selection of the translation initiation codon is less clear. We typically look upstream from the earliest ribosome stalled at a Cys codon within a transcript for possible start codons, choose the longest possible protein, and then modify as needed based on Ribo-seq and Ribo-RET [7, 24] and other published functional data [22, 25]. The choice of start codon was based on genome-wide trends reported by Shell *et al.* with translation initiation from leadered transcripts occurring from AUG, GUG, UUG, and AUU codons, and from AUG or GUG codons in leaderless transcripts [7]. We then used published Mtb Ribo-RET data to not only validate translation of our Cys-containing ORFs but also confirm or clarify their translation initiation site [24]. We also considered the mapped TSS included in the Wadsworth Interactive Browser [6, 7]. We chose the first start codon following the mapped TSS when there was no definitive Ribo-RET data or clean Ribo-seq to guide start codon assignment. Finally, of the 388 cleavage sites in unannotated ORFs that we mapped to Cys codons (mined from all five 5'-OH RNA-seq datasets), only 38 lacked a start codon and thus, not associated with a putative ORF.

The illustration in Fig. 5A contrasts the mechanics of toxin-mediated ribosome stalling versus Mtb Ribo-RET. Both approaches rely on stalled ribosomes, one due to scarcity of a single tRNA species, and the other caused by the antibiotic retapamulin that binds to the peptidyl transferase center and perturbs peptide bond formation to reveal translation start sites. Both transcriptome-wide approaches exploit the ~ 15 nt distance measured from inside the ribosome to mRNA cleavage either 5' (Fig. 5A, left) or 3' (Fig. 5A, right) of the stalled ribosome. We found that 24 of our 96 ORFs were present in the Mtb Ribo-RET dataset (Supplemental Table 1, ‘Found in Smith *et al.*’ column; Fig. 5B, red). Therefore, toxin-mediated ribosome stalling and Ribo-RET data are complementary approaches to strengthen the quality of genome annotation (especially for genes encoding sORFs) and uncover new biological and regulatory mechanisms.

Next, we sought to validate putative new ORFs identified through ribosome stalling at Cys codons with evidence that

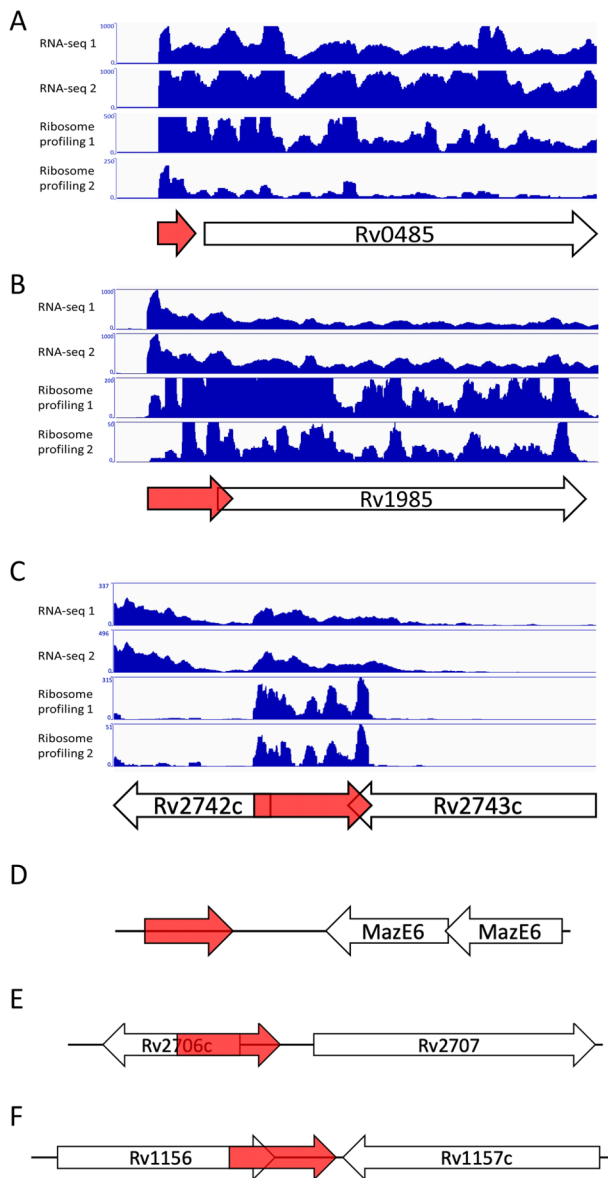


Figure 4. Unannotated ORFs can occur in diverse contexts. RNA-seq and Ribo-seq coverage data in blue (retrieved from <http://mtb.wadsworth.org>) support the transcription and translation of unannotated ORFs (red arrows) identified by 5'-OH RNA-seq occurring upstream as a Cys-responsive attenuator (A) or of unknown function (B), in the opposite direction (C) of annotated genes. X-axis shows the aligned genes to the RNA-seq and Ribo-seq data; Y-axis, individual sequence reads. Other examples of unannotated ORFs were found without any overlap (D) or with partial overlap in the reverse (E) or the same (F) direction to annotated genes.

they are translated into a stable protein. To this end, we mined publicly available Mtb mass spectrometry datasets for evidence that the putative unannotated Cys codon-containing proteins we uncovered are present in the Mtb proteome. We identified one or more tryptic peptides mapping to 69% (66 of the 96) of ORFs with ribosome stalling at Cys codons (Fig. 5B and Supplemental Table 2), providing strong, or moderate, supporting evidence (depending on how often they were detected) that these sORFs are translated in Mtb. Overall, we were able to validate 74% (71/96) of the ORFs identified by VapC4 toxin-mediated ribosome stalling

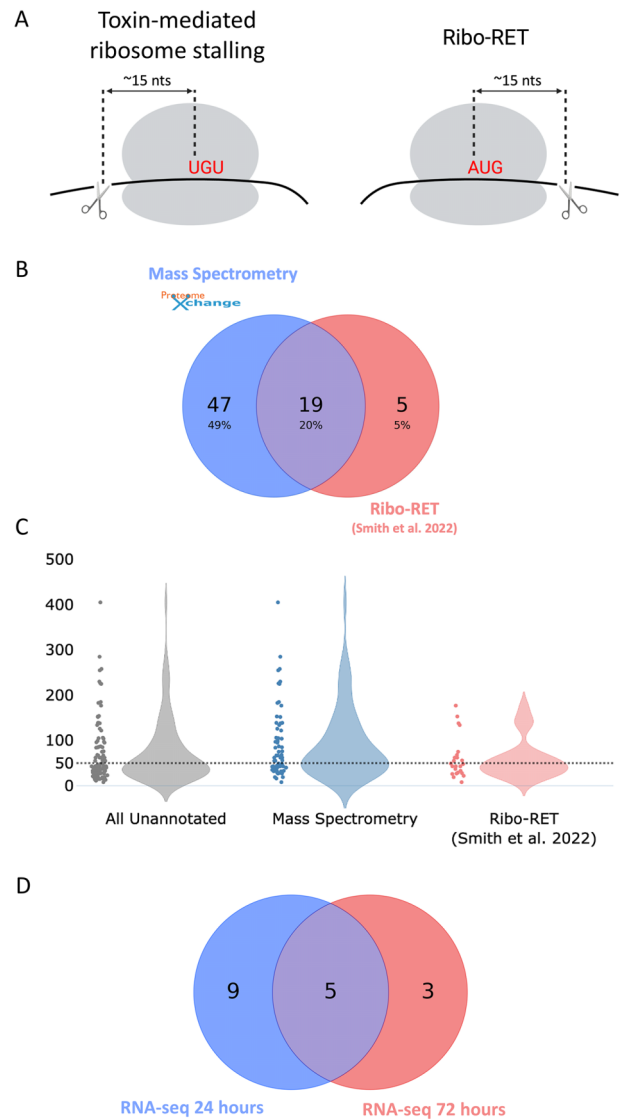


Figure 5. Validation of a subset of ORFs using Ribo-RET, mass spectrometry or RNA-seq. (A) Illustration contrasting the position of stalled ribosomes by 5'-OH RNA-seq (hungry codon) versus Ribo-RET (translation start site). (B) Venn diagram of subsets of ORFs validated by mass spectrometry, Ribo-RET or both. (C) Violin plots depicting size distribution of validated ORFs in panel (B). (D) Venn diagram of subsets of ORFs whose transcription increased upon VapC4 expression two-fold or more ($\log_2$ fold change ≥ 1 ; adjusted P -value $\leq .05$) by RNA-seq at 24 h, 72 h, or in both time points.

at Cys codon with Ribo-RET and/or mass spectrometry data.

Detection of tryptic peptides generated from 69% of the 96 ORFs is encouraging given that 29% of these are small proteins of ≤ 50 amino acids. These small proteins are notoriously difficult to detect by mass spectrometry because they contain few, if any, trypsin cleavage sites, and are often not abundant enough for reliable detection. Even when there are trypsin sites, the resulting peptides may be outside the optimal size window (too small or large) for detection by mass spectrometry [5, 19, 48]. Use of alternative proteases—chymotrypsin, LysC, LysN, AspN, GluC, and ArgC [49]—for peptide generation prior to mass spectrometry may improve identification of peptides from sORFs; however, they are not routinely

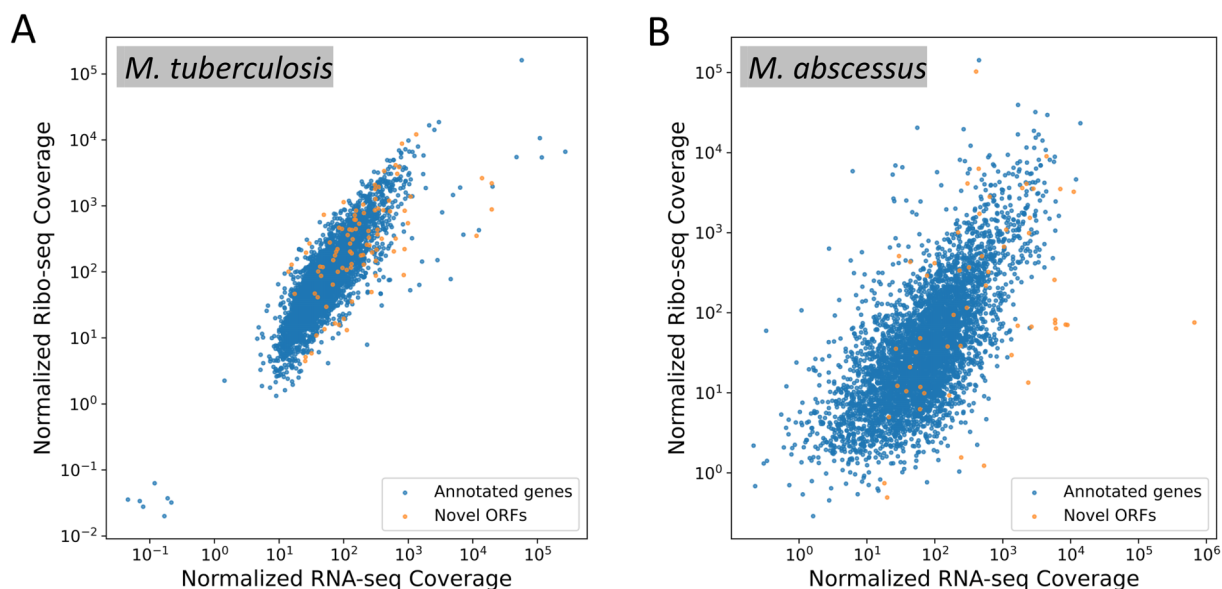


Figure 6. Published RNA-seq and Ribo-seq data support efficient translation of novel Mtb and *M. abscessus* ORFs. Pairwise comparison of normalized Mtb (A) [24] and *M. abscessus* (B) [32] RNA-seq and Ribo-seq coverage for annotated (blue dots) and novel ORFs (orange dots). Both plots used averaged replicate pairs.

used. There are other reasons for failure to detect novel ORF protein products by mass spectrometry or Ribo-RET. For example, some may only be differentially expressed under certain conditions, secreted after synthesis, or present in very low concentrations.

In fact, four of the very small Cys-rich proteins with known regulatory roles might be intrinsically unstable [22, 25]. These Cys attenuators reported by our group and others [7, 22, 25, 24] appear to function like the unstable 14 amino acid TrpL leader peptide involved in Trp operon attenuation [50]. Therefore, while the Cys attenuator protein in Fig. 4A might be unstable, we confirmed the stability (and legitimacy) of each of the four representative unannotated proteins in Fig. 4C–F by high confidence identification of multiple tryptic peptides from published Mtb mass spectrometry datasets (Supplemental Table 2).

VapC4 expression leads to differential expression of a subset of Cys-containing ORFs

Another approach to support the functional significance of the ORFs identified with Cys stalled ribosomes would be identification of their expression in the Mtb transcriptome, especially if their transcription is associated with VapC4. We mined our published RNA-seq datasets comparing Mtb with and without VapC4 expression [22]. We identified 14 and 8 ORFs from the 96 total whose expression increased by two-fold or more ($\log_2$ fold change ≥ 1) following VapC4 induction for 24 or 72 h, respectively; five of the ORFs were expressed at both 24 and 72 h (Fig. 5D and Supplemental Table 3). There was only one ORF whose transcription was downregulated two-fold or more at 72 h ($\log_2$ fold change ≤ -1 ; Supplemental Table 3). Overall, identification of these expressed transcripts supports the possibility that they are translated as well. The differential expression of these transcripts also suggests that they may have a physiological role that is linked to the function of VapC4.

Both annotated and novel Mtb ORFs exhibit coordinated transcript abundance and ribosome occupancy

We next used the published Mtb RNA-seq and Ribo-seq datasets [24] to examine the relationship between transcript abundance and ribosome occupancy for annotated Mtb genes compared to our 96 ORFs. The plotted blue dots in Fig. 6A represent individual annotated transcripts whose overall pattern reveals that RNAs present at higher levels generally harbor more ribosomes, consistent with coordinated transcription and translation. Because 93 of our 96 newly identified ORFs were also present in these datasets, we overlaid them onto the annotated reference framework (Fig. 6A, orange dots). In contrast to the novel ORFs reported by Smith *et al.*, which exhibited efficient translation from relatively low transcript levels, our novel ORFs generally follow the same RNA-ribosome scaling observed for annotated genes, indicating a closer concordance between transcript abundance and translation. The observation that 93 of our novel ORFs behave like the annotated ORFs at the population level provides strong support for their biological authenticity.

Validation of novel ORF translation by synthetic peptide matched MS/MS spectra

We used mass spectrometry to determine if peptides detected in cultured Mtb cells expressing VapC4 were chemically identical to their synthetic counterparts. Figure 7 shows four matched pairs of synthetic peptides versus those derived from Mtb cells. These four peptides were confidently identified following validation using PeptideShaker. The vertical peaks represent fragment ions generated during tandem MS/MS; the blue b-ions are fragments from the N-terminus, and the red y-ions are fragments from the C-terminus. These two ions appear at precise mass to charge ratios (m/z value). Here the MS/MS spectra are identical in each pair, i.e. the same b- and y-ions appear in both spectra, at the same m/z positions,

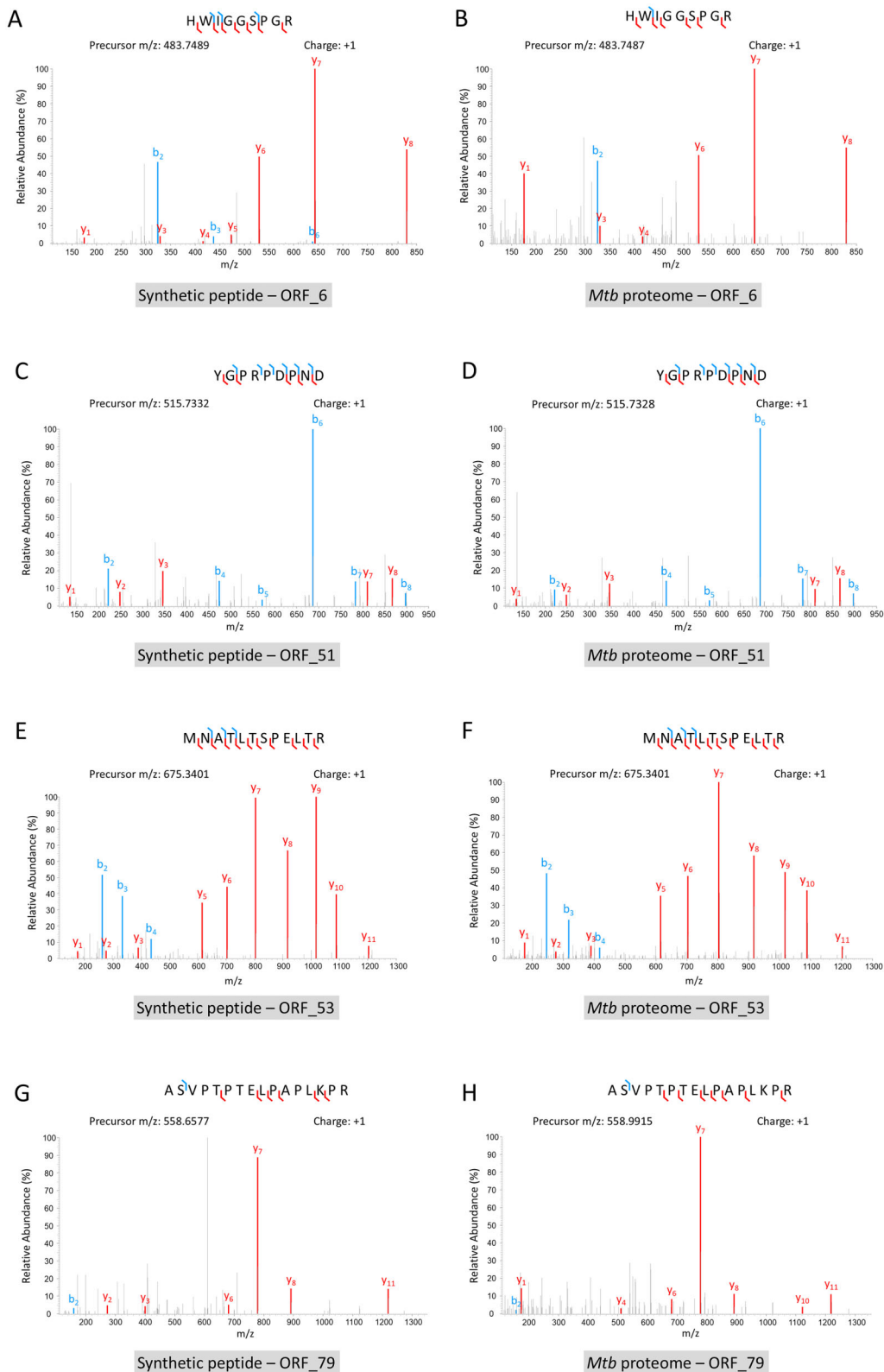


Figure 7. Validation of a subset of novel ORFs by mass spectrometry. Each pair of panels compares the MS/MS spectra from synthetic versus observed peptides. Peptide sequences from [Supplemental Table 1](#). **(A, B)** For ORF_6, amino acids 34–42. **(C, D)** For ORF_51, amino acids 126–134. **(E, F)** for ORF_53, amino acids 1–12. **(G, H)** For ORF_79, amino acids 16–31. Measured b-ions (blue) and y-ions (red). The precursor and charge are indicated below the peptide sequence.

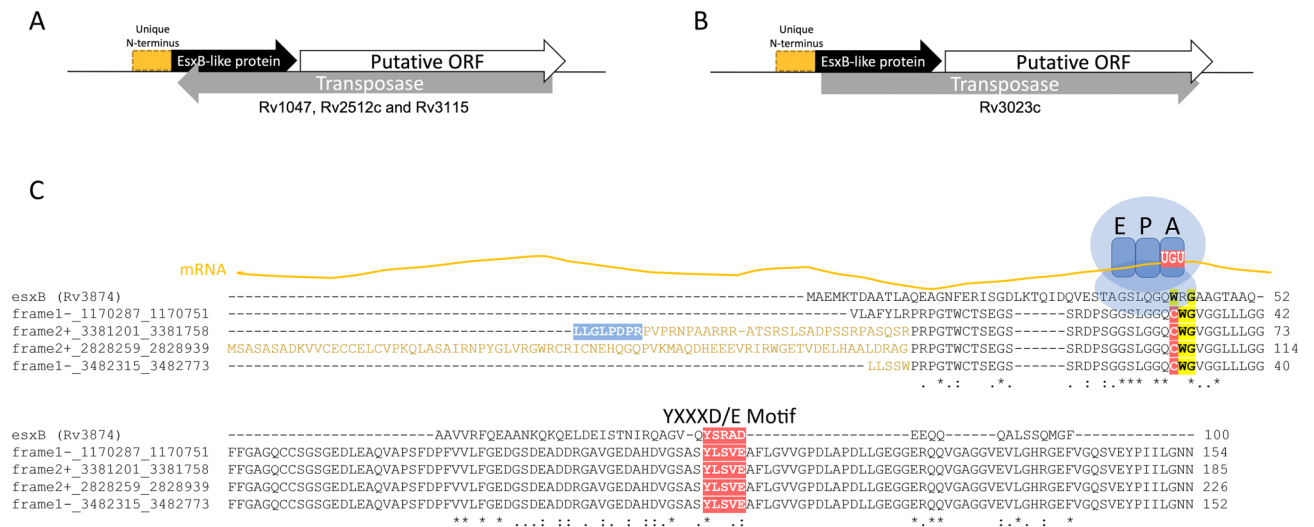


Figure 8. Four sORFs encode proteins related to EsxB/CFP-10. Illustration of the genome positions of four EsxB-like proteins highlighting variable N-termini (orange) and overlap with four transposases shown in gray (Rv1047, Rv2512c, Rv3115, and Rv3023c) that are identical except for a single nt change resulting in an A9T change very early in Rv2512c). A putative identical ORF is located downstream of the EsxB-like proteins. (A) Overlap with three transposases in the opposite orientation (B) Overlap with one transposase in the same orientation. (C) Clustal Omega sequence alignment of EsxB (Rv3874), against the four proteins derived from Cys-containing unannotated sORFs; identities (*), strong (:), and weak (.) similarities. Cys codon with stalled ribosome highlighted in red. WG motif highlighted in yellow. The ORF_42 peptide boxed in blue was identified by mass spectrometry.

with very similar intensities. Together these properties demonstrate peptide identity. This method is the gold-standard for definitive validation of peptides identified through Ribo-RET, Ribo-seq—and in our case ribosome stalling at Cys codons—because it provides direct biochemical evidence for their translation.

Four unannotated sORFs encode novel EsxB-like proteins

We performed NCBI-BLASTP searches on all unannotated proteins that harbored stalled ribosomes at Cys codons to determine if their sequences revealed functional clues. We identified four proteins with significant similarity to a 156 amino acid “ESAT-6-like” protein sequenced from an *Mtb* strain unique to a strain originating from Peru. Alignments to investigate sequence similarity of these four proteins to EsxA (formerly ESAT-6) versus EsxB (formerly CFP-10) suggest that these proteins are orthologs of EsxB (Fig. 8). However, each of the four unannotated proteins have distinctly different amino terminal sequences whose legitimacy was supported by the presence of a stalled ribosome at a Cys codon within these variable amino terminal sequences. We also identified a tryptic peptide confirmed by published mass spectrometry datasets (Supplemental Table 2) within the variable region (Fig. 8). All four proteins contain the hallmark Y-X₃-D/E motif and an appropriately aligned possible counterpart (i.e. W-G) of the W-X-G motif of EsxB (Fig. 8). The Y-X₃-D/E sequence is required for secretion of the EsxA/EsxB complex by type VII secretion systems; the W-X-G motif is thought to also be part of the secretion signal [51, 52].

Curiously, all four EsxB-like genes overlap with a portion of essentially identical transposase genes, each with its own Rv number on the opposing DNA strand (Fig. 8). These EsxB-like genes overlapping with transposases on the opposite strand (and those displayed in Fig. 4C and E) only come to light by methods that detect stalled ribosomes. The EsxA-EsxB heterodimer is a major virulence factor that resides within the

15-gene ESX-1 type VII secretion system locus. ESX-1 is essential for *Mtb* evasion of the host immune response [53]. Since EsxB is followed by EsxA in the ESX-1 locus, we examined each EsxB-like gene for a downstream ORF. Although we identified a putative ORF with some similarity to EsxA, it was more than twice the size of EsxA. Therefore, it does not appear to be a clear functional ortholog of EsxA. Because these four EsxB-like proteins do not have an apparent binding partner, they may function on their own, or with an EsxA partner encoded elsewhere in the genome. Although EsxA and the EsxA-EsxB complex are the most common focus of functional studies and EsxB is often referred to as simply a chaperone, one published report uncovered a novel biological activity for EsxB alone: recruitment and activation of human neutrophils [54]. If these EsxB-like proteins function in a similar manner, then ribosome stalling upon VapC4 expression is expected to block their translation and thus, dampen neutrophil activation through this pathway.

Overall, the majority of ORFs identified by ribosome stalling at Cys codons showed stretches of identity to portions of other bacterial or mycobacterial proteins, but no other definitive functional orthologs were identified. Some of them also contain motifs of eukaryotic zinc fingers (C-X₄-C and C-X₂-C) or the canonical heme binding motif CX₄CH in cytochrome c-type proteins. However, because there are few known zinc fingers in bacteria [55], the significance of these putative zinc-binding motifs is unclear.

Stalled ribosomes are also detectable at hungry codons in *M. abscessus* 5'-P RNA-seq datasets

Most bacterial VapC toxins are isoacceptor-specific tRNases that specifically recognize, cleave, and inactivate a single tRNA species [23,56–61]. Among these, the VapC5 toxin present in some *M. abscessus* clinical strains specifically targets tRNA^{SerCGA} for inactivation, leading to ribosome stalling at hungry Ser UCG codons (Fig. 9A) [23]. We analyzed our published VapC5 5'-P dataset in *M. abscessus* [23] and found

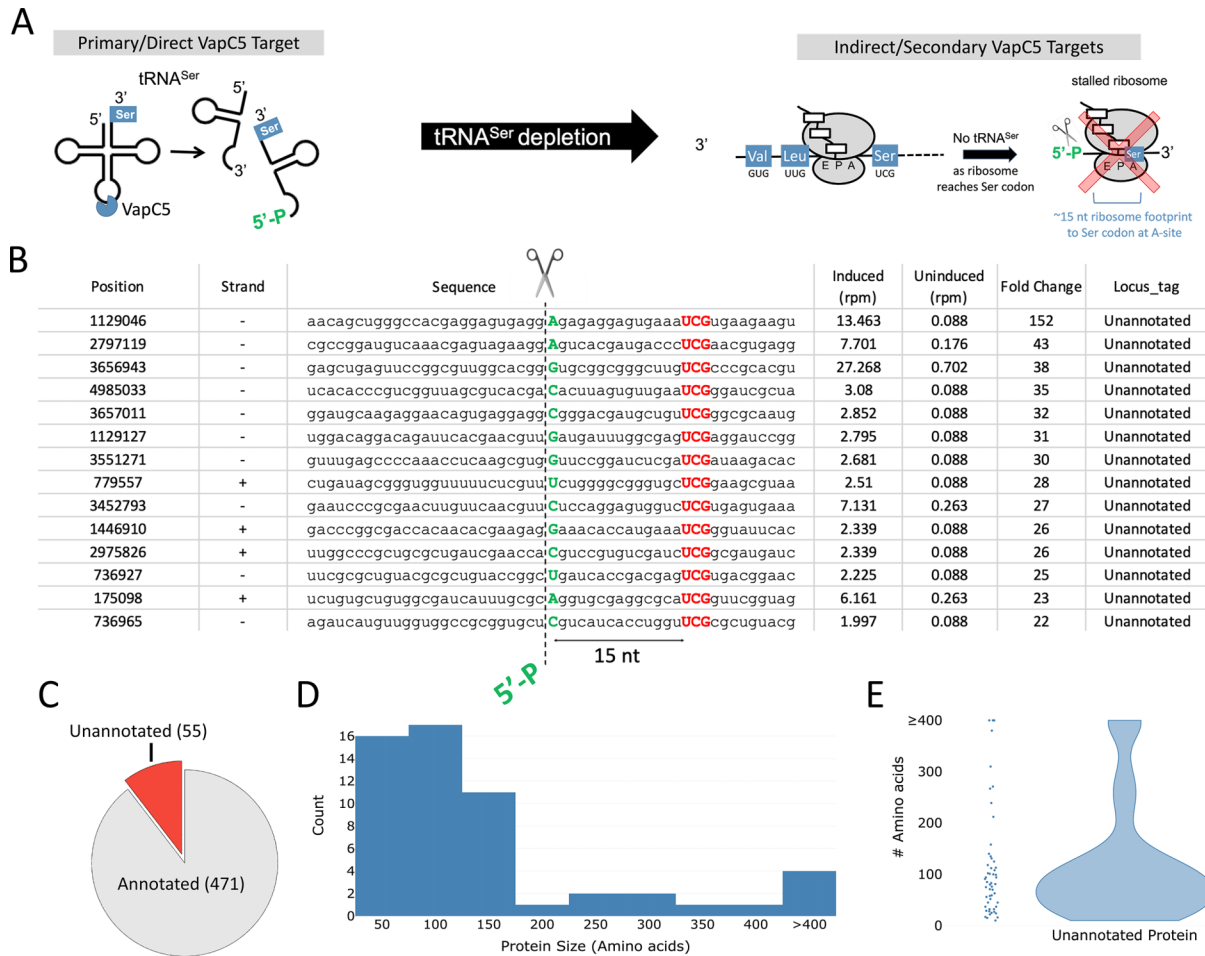


Figure 9. Depletion of tRNA^{Ser} by *M. abscessus* VapC5 leads to ribosome stalling at Ser UCG codons. **(A)** VapC5-mediated cleavage of tRNA^{Ser} leads to transcriptome-wide ribosome stalling at serine codons. **(B)** Unannotated mRNA hits in descending order of fold-change relative to the control in the 5'-P RNA-seq dataset (the RNase cleaving the stalling ribosome generates a 5'-P instead of a 5'-OH compared to Mtb) [23]. Serine codons are in bold red, ~15 nucleotides downstream of the 5'-P cleavage site (5' of the green capitalized letter). Genome position and strand shown. Unannotated Ser codon-containing ORFs were identified as follows: (i) Ser codons with a stalled ribosome by 5'-P RNA-seq were considered in-frame, (ii) minimum ORF length was set to ≥ 10 amino acids, (iii) start codons AUG, GUG, UUG, and AUU [7] and a stop UGA, UAA, or UAG codon were considered. We selected the longest possible ORF since the TSS for these ORFs are not known. **(C)** Pie chart of the 526 Ser codon stalled hits from the 5'-P RNA-seq dataset [23] displayed as annotated versus unannotated. Histogram **(D)** and violin plot **(E)** showing size distribution of the 55 unannotated ORFs.

526 Ser UCG ribosome stalling signatures at distinct transcripts (those with multiple stalled ribosomes at UCG codons were counted only once); 55/526 = 11% were unannotated (Fig. 9B and C). Of the unannotated ORFs detected, 16/55 = 29% were ≤ 50 amino acids, and 18/55 = 33% were ≤ 100 but ≥ 50 amino acids; the remaining 21 were > 101 amino acids (Fig. 9D and E). Although Ribo-RET has not yet been performed for any *M. abscessus* strains, we mined the mass spectrometry data from our VapC5 publication along with six publicly available mass spectrometry datasets for tryptic peptides derived from any of the 55 unannotated ORFs; 21/55 = 38% were validated by mass spectrometry (Supplemental Table 5).

Translation in *M. abscessus* is less tightly proportional to transcript abundance than in Mtb. Finally, using published data [32] we plotted *M. abscessus* mRNA levels and ribosome occupancy for both annotated (blue dots) and novel ORFs (orange dots) (Fig. 6B). In con-

trast to Mtb, *M. abscessus* shows a broader dispersion for annotated ORFs: ribosome occupancy is less tightly proportional to transcript abundance, there is greater vertical spread at similar RNA levels, and translation is more heterogeneous and less predictably coupled to transcription. The track of our novel *M. abscessus* ORFs shows the same dispersed distribution as the annotated ORFs. Although translation in *M. abscessus* annotated ORFs is more weakly coupled to transcript abundance than in Mtb, our newly identified ORFs follow the same RNA-ribosome relationship as annotated ORFs. This concordant behavior supports their authenticity as true coding regions rather than products of experimental or analytical artifacts.

Discussion

Here we report how the highly selective Mtb VapC4 toxin that depletes a single tRNA isoacceptor can be applied to improve annotation of the Mtb genome. Further, we show that a similar approach can be used for the genome of the poorly

annotated emerging pathogen *M. abscessus*. Our ability to capture signatures of putative stalled elongating ribosomes without antibiotic treatment represents a highly useful, alternate tactic to elucidate hidden messages within the Mtb genome. This approach is complementary to the 2022 Mtb Ribo-RET report that identified translation start sites by trapping initiating ribosomes at start codons [24]. As with Mtb and *E. coli* Ribo-RET studies [16, 17, 24, 62], we were able to identify (i) new putative small bacterial protein ORFs, (ii) other novel bacterial proteins within known ORFs (both in frame and out of frame), and (iii) proteins translated within annotated ORFs but from the antisense strand. Ribo-RET studies also used validation methods—reporter gene fusions or western detection of FLAG-tagged ORFs integrated into the genome—that supported actual *in vivo* expression or translation of a high percentage the novel ORFs uncovered [16, 17, 24, 62].

Nevertheless, while optimized bacterial Ribo-Seq [63] and Ribo-RET used together are powerful tools, they cannot comprehensively detect all functional ORFs. For example, Meydan *et al.* reported that they were able to identify ribosome occupancy peaks corresponding to 86% of annotated *E. coli* start codons [16]. They postulated that the other 14% may not be detectable by Ribo-RET due to changes in expression because of RET exposure. Also, Weaver *et al.* detected poor overlap between their predicted small ORFs and those identified by other groups using different ribosome profiling methods [64, 65]. For Mtb, Ribo-RET identified 2299 leadered and leaderless transcripts, 90 of which were not annotated [24]. The NCBI Mtb H37Rv genome lists 4018 protein coding genes, therefore the reduced Ribo-RET coverage for Mtb versus *E. coli* appears to be attributable in part to the sequence bias by micrococcal nuclease for cleavage on the 5' side of A or T DNA and A or U RNA sequences [66]. The GC rich Mtb genome is only 33%–34% AT relative to *E. coli* coding regions (46%–48% AT). Indeed, 74% of the ribosome protected Mtb Ribo-RET fragments were cleaved at an A or U [24]. Thus, there is room for further refinement of data derived from even the latest Ribo-seq methods.

Toxin-mediated ribosome stalling can help fill this gap because our 5'-OH RNA-seq method for identifying stalled ribosomes does not rely on micrococcal nuclease treatment. Here, using just the VapC4 toxin, we identified ribosomal stalling on 437 independent transcripts, 96 from putative unannotated genes and 341 from annotated genes, approaching 10% of the Mtb transcriptome. These 437 stalling events were picked up in our 5'-OH RNA-seq datasets only because there was enhancement of cleavage by an unknown RNase that cuts on the 5' side of the stalled ribosome at hungry Cys codons, leaving a 5'-OH that our assay is specifically designed to detect. Consequently, cleavage and degradation of transcripts harboring stalled ribosomes leads to a concomitant genome-wide reduction in these transcripts and their protein products [22]. Therefore, even though Cys codons are surprisingly abundant in Mtb coding regions—there are 3259 genes (81%) containing one or more Cys codons—we identify just a fraction of them because RNA harvested for each 5'-OH analysis is a snapshot of a dynamic process of ribosome stalling, transcript cleavage, and presumably ribosome release. The stalling events we detect are dependent on genome-wide expression levels of each transcript following VapC4 toxin induction, and this is expected to be influenced by length of toxin induction and the Mtb growth phase at time of induc-

tion. Thus, we would expect the depth for Cys codon stalling to continue to improve by increasing the number of 5'-OH RNA-seq runs under a variety of conditions. However, some Cys-containing transcripts may not be detectable under toxin-expressing conditions if their transcripts are low or translation is inhibited. Nevertheless, overall coverage is expected to increase after compiling data from multiple iterations of 5'-OH RNA-seq from RNAs harvested under varied growth conditions.

Although here we present depletion of tRNA^{Cys} as the example we chose to assess in depth, there are multiple isoacceptor-specific Mtb tRNase toxins that can be enlisted to enhance genome annotation and uncover novel ORFs/sORFs. We have identified more Mtb toxins that specifically target only one tRNA isoacceptor and exhibit ribosome stalling at codons for these depleted tRNAs—Gln, Lys, Phe, Pro, Ser, Trp, Asn—within the respective 5'-OH RNA-seq datasets ([57, 67]; Woychik lab, in preparation). This expands the number of available tools to uncover a substantially broader range of ORFs in the Mtb genome. As more hidden Mtb ORFs are unmasked, the next challenge will be to understand their function in Mtb virulence and stress adaptation. Our identification of new ORFs whose transcription is coincident with VapC4 toxin expression suggests that the role of some ORFs is linked to Mtb stress adaption as well.

Finally, the spectrum of Mtb isoacceptor-specific tRNases can be exploited to enhance annotation and reveal novel regulatory mechanisms in other mycobacteria (including *M. smegmatis* and the poorly annotated emerging pathogen *M. abscessus* (Fig. 9)). This approach can also be applied to other bacterial model organisms and incompletely annotated bacterial pathogens. Since we have identified a single tRNA^{Lys}-cleaving toxin in *M. smegmatis* [67] and a tRNA^{Ser}-cleaving toxin in an *M. abscessus* clinical strain to date [23], we can further expand coverage by ectopically expressing one of the Mtb isoacceptor-specific tRNase toxins to engage toxin-mediated ribosome stalling. In fact, ectopic expression of individual Mtb tRNase toxins should also induce ribosome stalling and mRNA cleavage in other bacteria that carry an RNase that rescues stalled ribosomes by cleaving mRNA on the 5' side of a stalled ribosome. *E. coli* lacks this activity but we were able to reconstitute detection of the ribosome stalling signature in *E. coli* 5'-OH RNA-seq datasets upon co-expression of a Lys tRNase toxin with Mtb RNase J [20]. The *E. coli* SmrB RNase, a ribosome rescue factor that cleaves between collided ribosomes [45], apparently does not function like Mtb RNase J. Therefore, toxin-mediated ribosome stalling can be applied to enhance genome annotation for any bacterium with established methods for co-expression of one or two proteins. For pathogen genomes with high AT content, toxin sequences can be codon optimized to improve expression. However, we routinely obtain efficient expression of Mtb TA toxins in both *E. coli* (~47% AT) and *Saccharomyces cerevisiae* (~62% AT) without codon optimization.

In summary, although our approach applies Ribo-RET data to accurately identify translation start codons, it does not depend on micrococcal nuclease and therefore avoids complications arising for its well-documented sequence bias. This distinction is particularly important for mycobacteria, which have GC-rich genomes. Therefore, mapping stalled ribosomes at hungry codons to identify translated ORFs—including unannotated ORFs—provides a powerful and

complementary strategy for illuminating the “dark matter” of bacterial proteomes.

Acknowledgements

We thank the William Jacobs laboratory (Albert Einstein College of Medicine) for providing the attenuated H37Rv strain mc² 6206 (Δ panCD Δ leuCD).

Author contributions: V.C.B. performed the VapC4 5'-OH RNA-seq experiments used in ORF analyses. E.A.T. and V.C.B. performed the analyses and figure preparation for Figs 1–9. E.A.T. performed the experiments, analyses, and figure preparation for Fig. 9. H.Z. and C.Z. provided proteomics expertise and assistance in preparation of Fig. 7. U.C. and V.C.B. compiled the data for Supplemental Table 1; E.A.T. reanalyzed and confirmed this important dataset. Supplemental Tables 1 and 2 were researched and compiled by V.C.B. and E.A.T. Supplemental Tables 3–5 were researched and compiled by E.A.T. RNA and protein from VapC4 expressed in *Mycobacterium tuberculosis* H37Rv was prepared by J.Z. in the R.N.H. laboratory. N.A.W. and E.A.T. wrote the manuscript, V.C.B., U.C., and R.N.H. reviewed the manuscript.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was funded in part by a fellowship to V.C.B. from CAPES – Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil, New Jersey Commission on Cancer Research Fellowship DFHS18PPC045 to U.C., as well as National Institutes of Health grant RO1 AI154464 to N.A.W. Funding to pay for the Open Access publication charges for this article was provided by the National Institutes of Health grant RO1 AI154464.

Data availability

The sequencing datasets used in this study were deposited in the NCBI Sequence Read Archive under BioProject accession numbers PRJNA662430 and PRJNA942981. The mass spectrometry proteomics dataset used in this study has been deposited to the ProteomeXchange Consortium via the PRIDE [68] partner repository with the dataset identifier PXD071737 and PXD071729.

References

- Haft DH, Badretdin A, Coulouris G *et al.* RefSeq and the prokaryotic genome annotation pipeline in the age of metagenomes. *Nucleic Acids Res* 2024;52:D762–9. <https://doi.org/10.1093/nar/gkad988>
- Bartholomaeus A, Kolte B, Mustafayeva A *et al.* smORFer: a modular algorithm to detect small ORFs in prokaryotes. *Nucleic Acids Res* 2021;49:e89. <https://doi.org/10.1093/nar/gkab477>
- Durrant MG, Bhatt AS. Automated prediction and annotation of small open reading frames in microbial genomes. *Cell Host Microbe* 2021;29:121–31. <https://doi.org/10.1016/j.chom.2020.11.002>
- F RC, Vasconcelos ATR. OCCAM: prediction of small ORFs in bacterial genomes by means of a target-decoy database approach and machine learning techniques. *Database (Oxford)* 2020;2020:baaa067.
- Fuchs S, Engelmann S. Small proteins in bacteria – big challenges in prediction and identification. *Proteomics* 2023;23:e2200421. <https://doi.org/10.1002/pmic.202200421>
- Cortes T, Schubert OT, Rose G *et al.* Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*. *Cell Rep* 2013;5:1121–31. <https://doi.org/10.1016/j.celrep.2013.10.031>
- Shell SS, Wang J, Lapierre P *et al.* Leaderless transcripts and small proteins are common features of the Mycobacterial translational landscape. *PLoS Genet* 2015;11:e1005641. <https://doi.org/10.1371/journal.pgen.1005641>
- Duval M, Cossart P. Small bacterial and phagic proteins: an updated view on a rapidly moving field. *Curr Opin Microbiol* 2017;39:81–8. <https://doi.org/10.1016/j.mib.2017.09.010>
- Storz G, Wolf YI, Ramamurthi KS. Small proteins can no longer be ignored. *Annu Rev Biochem* 2014;83:753–77. <https://doi.org/10.1146/annurev-biochem-070611-102400>
- Hucker SM, Ardern Z, Goldberg T *et al.* Discovery of numerous novel small genes in the intergenic regions of the *Escherichia coli* O157:H7 Sakai genome. *PLoS One* 2017;12:e0184119. <https://doi.org/10.1371/journal.pone.0184119>
- Sberro H, Fremin BJ, Zlitni S *et al.* Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell* 2019;178:1245–59. <https://doi.org/10.1016/j.cell.2019.07.016>
- Gray T, Storz G, Papenfort K. Small proteins; big questions. *J Bacteriol* 2022;204:e0034121. <https://doi.org/10.1128/JB.00341-21>
- Fijalkowska D, Fijalkowski I, Willems P *et al.* Bacterial riboproteogenomics: the era of N-terminal proteoform existence revealed. *FEMS Microbiol Rev* 2020;44:418–31. <https://doi.org/10.1093/femsre/fuaa013>
- Sawyer EB, Cortes T. Ribosome profiling enhances understanding of mycobacterial translation. *Front Microbiol* 2022;13:976550. <https://doi.org/10.3389/fmicb.2022.976550>
- Vazquez-Laslop N, Sharma CM, Mankin A *et al.* Identifying small open reading frames in prokaryotes with ribosome profiling. *J Bacteriol* 2022;204:e0029421. <https://doi.org/10.1128/JB.00294-21>
- Meydan S, Marks J, Klepacki D *et al.* Retapamulin-assisted ribosome profiling reveals the alternative bacterial proteome. *Mol Cell* 2019;74:481–93. <https://doi.org/10.1016/j.molcel.2019.02.017>
- Weaver J, Mohammad F, Buskirk AR *et al.* Identifying small proteins by ribosome profiling with stalled initiation complexes. *mBio* 2019;10:e02819–18. <https://doi.org/10.1128/mBio.02819-18>
- Seefeldt AC, Nguyen F, Antunes S *et al.* The proline-rich antimicrobial peptide Onc112 inhibits translation by blocking and destabilizing the initiation complex. *Nat Struct Mol Biol* 2015;22:470–5. <https://doi.org/10.1038/nsmb.3034>
- Ahrens CH, Wade JT, Champion MM *et al.* A practical guide to small protein discovery and characterization using mass spectrometry. *J Bacteriol* 2022;204:e0035321. <https://doi.org/10.1128/jb.00353-21>
- Barth VC, Zeng JM, Vvedenskaya IO *et al.* Toxin-mediated ribosome stalling reprograms the *Mycobacterium tuberculosis* proteome. *Nat Commun* 2019;10:3035. <https://doi.org/10.1038/s41467-019-10869-8>
- Crooks GE, Hon G, Chandonia JM *et al.* WebLogo: a sequence logo generator. *Genome Res* 2004;14:1188–90. <https://doi.org/10.1101/gr.849004>

22. Barth VC, Chauhan U, Zeng J *et al.* *Mycobacterium tuberculosis* VapC4 toxin engages small ORFs to initiate an integrated oxidative and copper stress response. *Proc Nat Acad Sci USA* 2021;118:e2022136118.
23. Troian EA, Maldonado HM, Chauhan U *et al.* *Mycobacterium abscessus* VapC5 toxin potentiates evasion of antibiotic killing by ribosome overproduction and activation of multiple resistance pathways. *Nat Commun* 2023;14:3705. <https://doi.org/10.1038/s41467-023-38844-4>
24. Smith C, Canestrari JG, Wang AJ *et al.* Pervasive translation in *Mycobacterium tuberculosis*. *eLife* 2022;11:e73980. <https://doi.org/10.7554/eLife.73980>
25. Canestrari JG, Lasek-Nesselquist E, Upadhyay A *et al.* Polycysteine-encoding leaderless short ORFs function as cysteine-responsive attenuators of operonic gene expression in mycobacteria. *Mol Microbiol*;2020;114:93–108. <https://doi.org/10.1111/mmi.14498>
26. Hummon AB, Lim SR, Difilippantonio MJ *et al.* Isolation and solubilization of proteins after TRIzol extraction of RNA and DNA from patient material following prolonged storage. *BioTechniques* 2007;42:467–470, 472. <https://doi.org/10.2144/000112401>
27. Vaudel M, Barsnes H, Berven FS *et al.* SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* 2011;11:996–9. <https://doi.org/10.1002/pmic.201000595>
28. Vaudel M, Burkhart JM, Zahedi RP *et al.* PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotechnol* 2015;33:22–4. <https://doi.org/10.1038/nbt.3109>
29. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004;20:1466–7. <https://doi.org/10.1093/bioinformatics/bth092>
30. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* 2014;5:5277. <https://doi.org/10.1038/ncomms6277>
31. Demichev V, Messner CB, Vernardis SI *et al.* DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* 2020;17:41–4. <https://doi.org/10.1038/s41592-019-0638-x>
32. Miranda-CasoLuengo AA, Staunton PM, Dinan AM *et al.* Functional characterization of the *Mycobacterium abscessus* genome coupled with condition specific transcriptomics reveals conserved molecular strategies for host adaptation and persistence. *BMC Genomics* 2016;17:553. <https://doi.org/10.1186/s12864-016-2868-y>
33. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 2011;17:200.
34. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–9. <https://doi.org/10.1038/nmeth.1923>
35. Ramirez F, Ryan DP, Gruning B *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 2016;44:W160–5. <https://doi.org/10.1093/nar/gkw257>
36. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 2012;28:2184–5. <https://doi.org/10.1093/bioinformatics/bts356>
37. Ramage HR, Connolly LE, Cox JS. Comprehensive functional analysis of *Mycobacterium tuberculosis* toxin–antitoxin systems: implications for pathogenesis, stress responses, and evolution. *PLoS Genet* 2009;5:e1000767. <https://doi.org/10.1371/journal.pgen.1000767>
38. Sala A, Bordes P, Genevaux P. Multiple toxin–antitoxin systems in *Mycobacterium tuberculosis*. *Toxins* 2014;6:1002–20. <https://doi.org/10.3390/toxins6031002>
39. Fraikin N, Goormaghtigh F, Van Melderden L. Type II toxin–antitoxin systems: evolution and revolutions. *J Bacteriol* 2020;202:e00763–19. <https://doi.org/10.1128/JB.00763-19>
40. Cruz JW, Sharp JD, Hoffer ED *et al.* Growth-regulating *Mycobacterium tuberculosis* VapC-mt4 toxin is an isoacceptor-specific tRNase. *Nat Commun* 2015;6:7480. <https://doi.org/10.1038/ncomms8480>
41. Sharp JD, Cruz JW, Raman S *et al.* Growth and translation inhibition through sequence-specific RNA binding by *Mycobacterium tuberculosis* VapC toxin. *J Biol Chem* 2012;287:12835–47. <https://doi.org/10.1074/jbc.M112.340109>
42. Schifano JM, Vvedenskaya IO, Knoblauch JG *et al.* An RNA-seq method for defining endoribonuclease cleavage specificity identifies dual rRNA substrates for toxin MazF-mt3. *Nat Commun* 2014;5:3538. <https://doi.org/10.1038/ncomms4538>
43. McKenzie JL, Duyvestyn JM, Smith T *et al.* Determination of ribonuclease sequence-specificity using Pentaprobates and mass spectrometry. *RNA* 2012;18:1267–78. <https://doi.org/10.1261/rna.031229.111>
44. Woolstenhulme CJ, Guydosh NR, Green R *et al.* High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. *Cell Rep* 2015;11:13–21. <https://doi.org/10.1016/j.celrep.2015.03.014>
45. Saito K, Kratzat H, Campbell A *et al.* Ribosome collisions induce mRNA cleavage and ribosome rescue in bacteria. *Nature* 2022;603:503–8. <https://doi.org/10.1038/s41586-022-04416-7>
46. Fesenko I, Sahakyan H, Dhyani R *et al.* The hidden bacterial microproteome. *Mol Cell* 2025;85:1024–41. <https://doi.org/10.1016/j.molcel.2025.01.025>
47. Kelkar DS, Kumar D, Kumar P *et al.* Proteogenomic analysis of *Mycobacterium tuberculosis* by high resolution mass spectrometry. *Mol Cell Proteomics* 2011;10:M111.011445. <https://doi.org/10.1074/mcp.M111.011627>
48. Hemm MR, Weaver J, Storz G. *Escherichia coli* small proteome. *EcoSal Plus* 2020;9:ecosalplus.ESP–0031–2019. <https://doi.org/10.1128/ecosalplus.esp-0031-2019>
49. Giansanti P, Tsiatsiani L, Low TY *et al.* Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat Protoc* 2016;11:993–1006. <https://doi.org/10.1038/nprot.2016.057>
50. Merino E, Jensen RA, Yanofsky C. Evolution of bacterial trp operons and their regulation. *Curr Opin Microbiol* 2008;11:78–86. <https://doi.org/10.1016/j.mib.2008.02.005>
51. Ates LS, Houben ENG, Bitter W. Type VII secretion: a highly versatile secretion system. *Microbiol Spectr* 2016;4:VMBF–0011–2015. <https://doi.org/10.1128/microbiolspec.VMBF-0011-2015>
52. Daleke MH, Ummels R, Bawono P *et al.* General secretion signal for the mycobacterial type VII secretion pathway. *Proc Natl Acad Sci USA* 2012;109:11342–7. <https://doi.org/10.1073/pnas.1119453109>
53. Groschel MI, Sayes F, Simeone R *et al.* ESX secretion systems: mycobacterial evolution to counter host immunity. *Nat Rev Micro* 2016;14:677–91. <https://doi.org/10.1038/nrmicro.2016.131>
54. Welin A, Bjornsdottir H, Winther M *et al.* CFP-10 from *Mycobacterium tuberculosis* selectively activates human neutrophils through a pertussis toxin-sensitive chemotactic receptor. *Infect Immun* 2015;83:205–13. <https://doi.org/10.1128/IAI.02493-14>
55. Malgieri G, Palmieri M, Russo L *et al.* The prokaryotic zinc-finger: structure, function and comparison with the eukaryotic counterpart. *FEBS J* 2015;282:4480–96. <https://doi.org/10.1111/febs.13503>
56. Chauhan U, Barth VC, Woychik NA. tRNA(fMet) inactivating *Mycobacterium tuberculosis* VapBC toxin–antitoxin systems as therapeutic targets. *Antimicrob Agents Chemother* 2022;66:e0189621. <https://doi.org/10.1128/aac.01896-21>
57. Cintron M, Zeng JM, Barth VC *et al.* Accurate target identification for *Mycobacterium tuberculosis* endoribonuclease toxins requires expression in their native host. *Sci Rep* 2019;9:5949. <https://doi.org/10.1038/s41598-019-41548-9>
58. Lopes AP, Lopes LM, Fraga TR *et al.* VapC from the leptospiral VapBC toxin–antitoxin module displays ribonuclease activity on

- the initiator tRNA. *PLoS One* 2014;9:e101678. <https://doi.org/10.1371/journal.pone.0101678>
59. Walling LR, Butler JS. Homologous VapC toxins inhibit translation and cell growth by sequence-specific cleavage of tRNA(fMet). *J Bacteriol* 2018;200:e00582–17. <https://doi.org/10.1128/JB.00582-17>
 60. Winther K, Tree JJ, Tollervey D *et al.* VapCs of *Mycobacterium tuberculosis* cleave RNAs essential for translation. *Nucleic Acids Res* 2016;44:9860–71. <https://doi.org/10.1093/nar/gkw781>
 61. Winther KS, Gerdes K. Enteric virulence associated protein VapC inhibits translation by cleavage of initiator tRNA. *Proc Natl Acad Sci USA* 2011;108:7403–7. <https://doi.org/10.1073/pnas.1019587108>
 62. Stringer A, Smith C, Mangano K *et al.* Identification of novel translated small ORFs in *Escherichia coli* using complementary ribosome profiling approaches. *J Bacteriol* 2021;204:JB0035221. <https://doi.org/10.1128/JB.00352-21>
 63. Mohammad F, Green R, Buskirk AR. A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *eLife* 2019;8:e42591. <https://doi.org/10.7554/eLife.42591>
 64. Baek J, Lee J, Yoon K *et al.* Identification of unannotated small genes in *Salmonella*. *G3* 2017;7:983–9. <https://doi.org/10.1534/g3.116.036939>
 65. Nakahigashi K, Takai Y, Kimura M *et al.* Comprehensive identification of translation start sites by tetracycline-inhibited ribosome profiling. *DNA Res* 2016;23:193–201. <https://doi.org/10.1093/dnares/dsw008>
 66. Dingwall C, Lomonosoff GP, Laskey RA. High sequence specificity of micrococcal nuclease. *Nucl Acids Res* 1981;9:2659–74. <https://doi.org/10.1093/nar/9.12.2659>
 67. Barth VC, Woychik NA. The sole *Mycobacterium smegmatis* MazF toxin targets tRNA(Lys) to impart highly selective, codon-dependent proteome reprogramming. *Front Genet* 2019;10:1356. <https://doi.org/10.3389/fgene.2019.01356>
 68. Perez-Riverol Y, Bai J, Bandla C *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* 2022;50:D543–52. <https://doi.org/10.1093/nar/gkab1038>