Pervasive Translation in *Mycobacterium tuberculosis*

Carol Smith[1†], Jill G. Canestrari[1†], Jing Wang[1†], Keith M. Derbyshire[1,2*], Todd A. Gray[1,2*], and Joseph T. Wade[1,2*]

[1]Wadsworth Center, New York State Department of Health, Albany, New York, USA.

[2]Department of Biomedical Sciences, School of Public Health, University at Albany, Albany, New York, USA.

[†]Equal author contribution.

[*]Corresponding authors:

joseph.wade@health.ny.gov

todd.gray@health.ny.gov

keith.derbyshire@health.ny.gov

ABSTRACT

ORF boundaries in bacterial genomes have largely been drawn by gene prediction algorithms. These algorithms often fail to predict ORFs with non-canonical features. Recent developments in genome-scale mapping of translation have facilitated the empirical identification of ORFs. Here, we use ribosome profiling approaches to map initiating and elongating ribosomes in *Mycobacterium tuberculosis*. Thus, we identify over 1,000 novel ORFs, revealing that much of the genome encodes proteins in overlapping reading frames, and/or on both strands. Most of the novel ORFs are short (sORFs), impeding their identification by traditional methods. The strong codon bias that characterizes annotated mycobacterial ORFs is not evident in the aggregate novel sORFs; hence most are unlikely to encode functional proteins. Our data suggest that bacterial transcriptomes are subject to pervasive translation. We speculate that the inefficiency of expressing spurious sORFs may be offset by positive contributions to *M. tuberculosis* biology through activities of a small subset.

INTRODUCTION

Advances in DNA sequencing technology have led to an explosion in the number of sequenced bacterial genomes (*1*), with the vast majority of genes identified using automated prediction algorithms (*2–4*). While these algorithms typically have a low false-positive rate, they often fail to identify non-canonical open reading frames (ORFs), including short ORFs (sORFs; ≤50 codons; most algorithms have a lower size limit of 30-50 codons), overlapping ORFs (*5*), and leaderless ORFs that are translated from mRNAs that lack a 5' UTR (*6*, *7*). Ribosome profiling is a powerful experimental approach to identify the translated regions of mRNAs by mapping ribosome-protected RNA fragments (*8*). We recently used ribosome profiling to investigate translation in *Mycobacterium smegmatis* (*9*). Our data demonstrated that there is widespread leaderless translation in *M. smegmatis*, and suggested that many leaderless *M. smegmatis* mRNAs encode sORFs. Leaderless translation is also widespread in other mycobacteria (*9–11*), including *Mycobacterium tuberculosis* (*9*, *10*), the causative agent of tuberculosis, suggesting active 5' sORF translation in this species. Moreover, the enormous potential for leadered translation initiation to contribute to the unannotated *M. tuberculosis* proteome has not been explored.

RESULTS AND DISCUSSION

Two previous studies of *M. tuberculosis* identified 369 transcription start sites corresponding to transcripts beginning with the sequence "RUG", where R = A or G (*9*, *10*). There is strong evidence that mycobacterial RNAs with a 5' RUG sequence are efficiently translated as leaderless mRNAs (*9*). To determine whether the 369 RUG identified transcription start sites correspond to the starts of leaderless mRNA ORFs, we performed ribosome profiling in *M. tuberculosis*. All genome-scale data described in this manuscript can be viewed in our interactive genome browser (http://mtb.wadsworth.org/). Consistent with previous studies (*12*, *13*), we observed increased ribosome occupancy within the set of all annotated, leadered ORFs that was accentuated at the start and stop codons(Fig. S1). We reasoned that if the 369 transcripts with a 5' RUG are genuine leaderless mRNAs, they would also exhibit increased ribosome occupancy at start and stop codons relative to other locations within the putative ORF. However, we limited our analysis to putative stop codons, because we observed increased signal in ribosome profiling libraries at the 5' ends of leadered mRNAs, likely reflecting an artifact of library construction (Fig. S2). We observed enhanced ribosome occupancy at the predicted stop codons of most of the 369 RNAs (Fig. 1), indicating that most, if not all, are fully translated as leaderless mRNAs. Strikingly, 268 of these mRNAs encode sORFs (≤50 codons), with 45 sORFs having ≤5 codons. We conclude that *M. tuberculosis* encodes hundreds of actively translated sORFs on leaderless mRNAs.

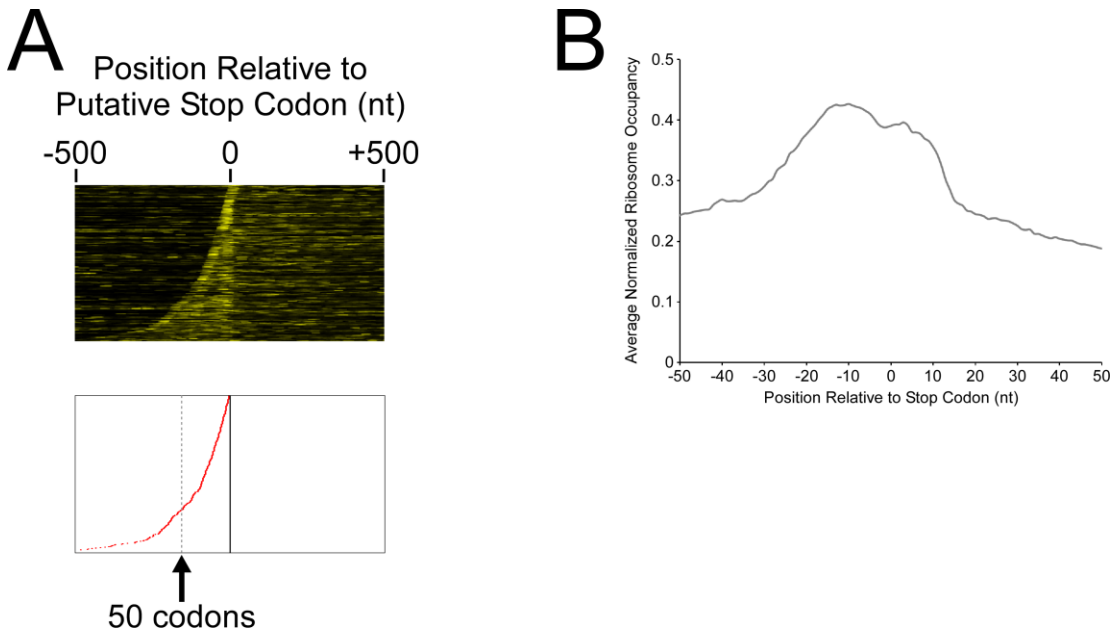
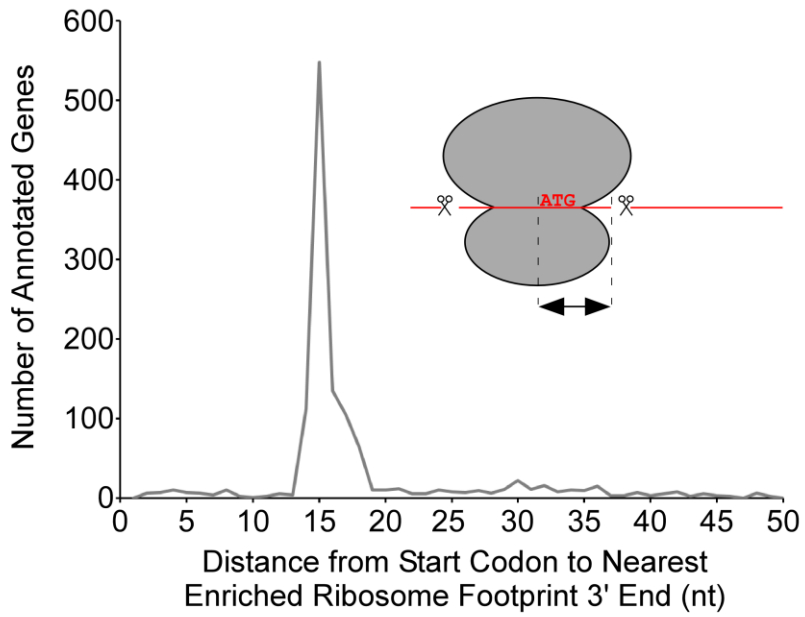


Figure 1. Translation of hundreds of unannotated ORFs from leaderless mRNAs. (A) Heatmap showing normalized sequence coverage from ribosome profiling data for *M. tuberculosis*. The heatmap shows data for 369 putative ORFs found on leaderless mRNAs. Sequence coverage is shown for the region from -500 bp to +500 bp relative to the putative stop codon. The plot below the heatmap shows the expected position of start codons (red datapoints) and stop codons (black datapoints). The position of ORFs with 50 codons is indicated by a dashed vertical line. **(B)** Average normalized coverage for all 369 putative ORFs in the region from -50 to +50 bp relative to the putative stop codon.

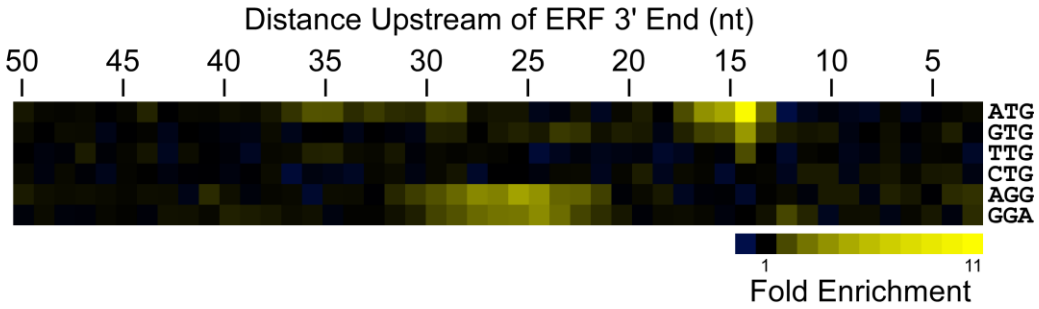
Based on our earlier study (*9*), we estimate that a quarter to a third of all *M. tuberculosis* mRNAs are leaderless. Given that our data support the existence of >300 unannotated ORFs on leaderless mRNAs, we speculated that there are many more unannotated ORFs on leadered mRNAs. Whereas leaderless translation initiation sites stand out, generating characteristically clear RNA-seq and Ribo-seq profiles, leadered initiation profiles are less clear. Identification of novel leadered ORFs from conventional ribosome profiling data is challenging due to (i) the potential for artifactual signal in 5' UTRs due to the binding of RNA-binding proteins (*14*), and (ii) masking of signal by overlapping ORFs on the same strand. Identification of novel ORFs can be facilitated by performing ribosome profiling on cells that have been treated with an antibiotic that traps initiating ribosomes, allowing elongating ribosomes to release their mRNAs (*15–20*). Two recent studies highlight the use of the antibiotics retapamulin and Onc112 for application of this approach to *Escherichia coli* (*18, 20*). We performed a similar ribosome profiling of *M. tuberculosis* treated with retapamulin ("Ribo-RET" (*20*)). We aligned the ribosome-protected RNA fragment sequences to the *M. tuberculosis* genome and identified "Enriched Ribosome Footprints" (ERFs), sites of ribosome occupancy that exceed the local background (Table S1). We hypothesized that ERFs correspond to sites of translation initiation. In support of this idea, there is a strong enrichment of ERF 3' ends 15 nt downstream of the start codons of annotated genes (Fig. 2A; note that genes encoded at the start of leaderless mRNAs were excluded from this analysis).

We determined the abundance of all trinucleotide sequences in the 50 nt upstream of ERF 3' ends. Strikingly, there is a strong enrichment of ATG, GTG and TTG (start codons), but not CTG, 15 nt upstream of ERF 3' ends, and a strong enrichment of AGG and GGA (components of a consensus Shine-Dalgarno sequence) 24-29 nt upstream of ERF 3' ends (Fig. 2B). We also observed >2-fold enrichment of GTG 16 nt upstream of ERF 3' ends, and of ATG 14, 16, 17 and 18 nt upstream of ERF 3' ends (Fig. 2B). In addition to enrichment of start codon and Shine-Dalgarno-like sequences upstream of ERF 3' ends, we observed a strong enrichment of A/T immediately downstream of the 3' end, i.e. on the other side of the site cleaved by micrococcal nuclease (MNase) during the Ribo-RET procedure (Fig. S3). This sequence bias is likely not due to a biological phenomenon, but rather to the sequence preference of MNase, which is known to display sequence bias when cutting DNA (*21*), and RNA (*13*). This bias is evidenced in our libraries, with 74% of sequenced ribosome-protected fragments having an "A" or "U" at the 5' end, i.e. immediately downstream of the upstream cleavage site. Given that the genomic A/T content in *M. tuberculosis* is only 34%, it is likely that this sequence bias underrepresented or prevented detection of many initiation sites.

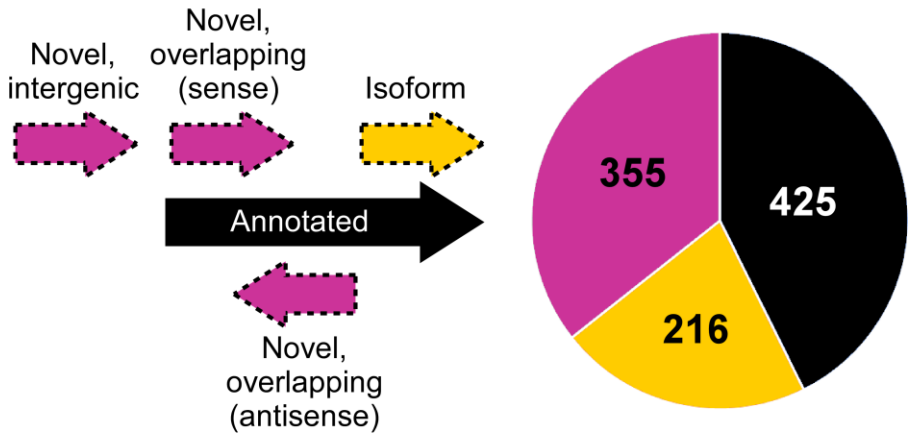
A


Number of Annotated Genes
600
500
400
300
200
100
0
0 5 10 15 20 25 30 35 40 45 50
Distance from Start Codon to Nearest
Enriched Ribosome Footprint 3' End (nt)
ATG

B


Distance Upstream of ERF 3' End (nt)
50 45 40 35 30 25 20 15 10 5
ATG
GTG
TTG
CTG
AGG
GGA
1
11
Fold Enrichment

C


Novel, intergenic
Novel, overlapping (sense)
Isoform
Annotated
Novel, overlapping (antisense)
355
425
216

D

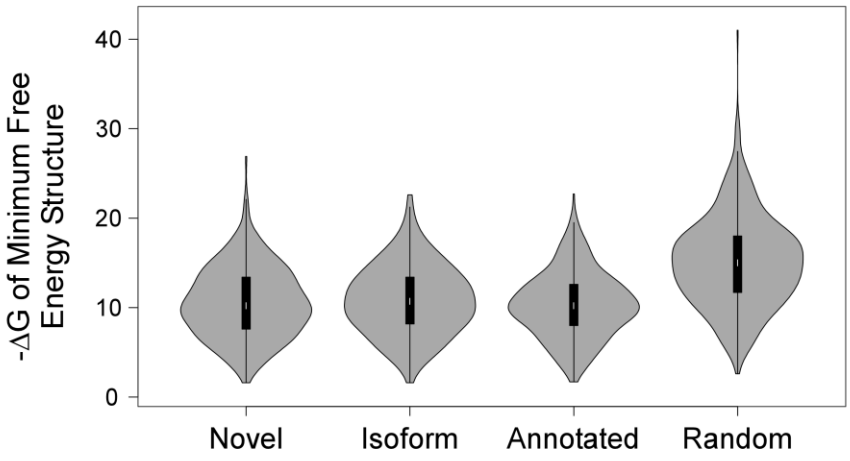

-ΔG of Minimum Free Energy Structure
40
30
20
10
0
Novel
Isoform
Annotated
Random

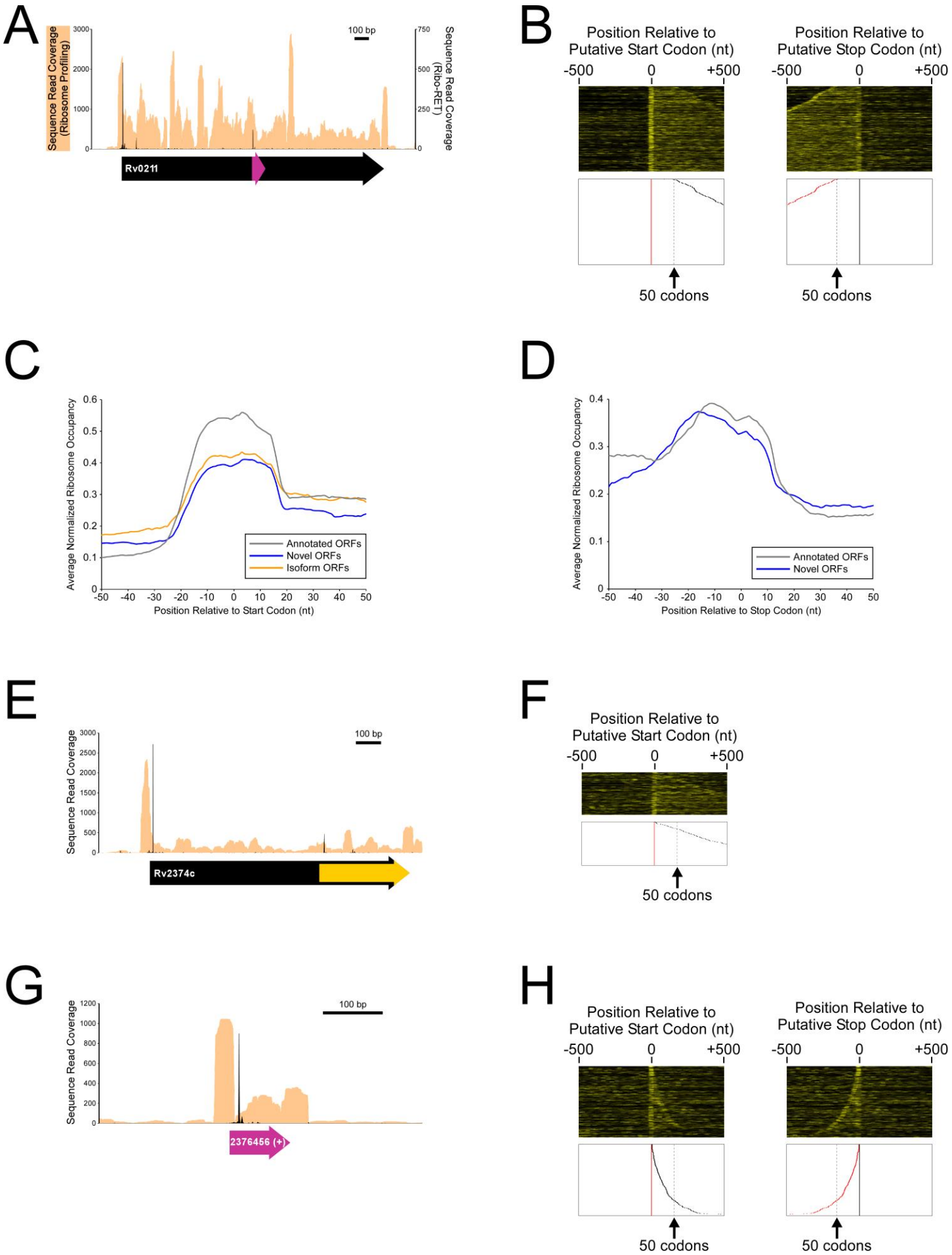
Figure 2. Ribo-RET of *M. tuberculosis* identifies sites of translation initiation. (A) Frequency distribution of the distances between the start codons of annotated genes and EFRs. **(B)** Heatmap showing the enrichment of trinucleotide sequences upstream of EFRs relative to randomly selected control regions. **(C)** Distribution of different classes of ORFs identified by Ribo-RET. **(D)** Violin plots showing the -ΔG for the predicted minimum free energy structure for the regions from -40 to +10 nt relative to putative start codons for the different classes of ORF.

We identified ERFs with putative start codons 14-18 nt upstream of their 3' ends in each of two replicate experiments (Table S2; see Methods for details). To be conservative, we generated a list of 996 high-confidence start codons by requiring that they be identified in both replicates. We estimate a false discovery rate (FDR) for this list of 15% (see Methods for details). Annotated genes accounted for 43% of the putative start codons; 22% of putative start codons are located within annotated genes, in-frame relative to the overlapping gene ("isoform ORFs"), and 36% represent completely unannotated ORFs ("novel ORFs") that are either found entirely in intergenic regions or overlap partly or completely with annotated genes in the sense and/or antisense orientation (Fig. 2C). Strikingly, 78% of the novel ORFs we identified are <50 codons in length, with 88 ORFs consisting only of a start and stop codon (Table S2). We reasoned that if the isoform ORFs and novel ORFs are genuine, they should have Shine-Dalgarno sequences upstream, and their start codons should each be associated with a region of reduced RNA secondary structure, as has been described for ORFs in *Escherichia coli* (*22*). As we had observed for the set of all ERFs, regions upstream of isoform ORFs and novel ORFs are associated with an enrichment of AGG and GGA sequences in the expected location of a Shine-Dalgarno sequence (Fig. S4). This enrichment is lower than for annotated genes (Fig. S4), but it is important to note that a Shine-Dalgarno sequence was also likely a criterion in computationally predicting the initiation codons of annotated genes. We also assessed the level of RNA secondary structure upstream of all the predicted ORFs. The predicted secondary structure for a set of random genomic sequences was significantly higher than the predicted secondary structure around the start of the identified annotated ORFs (Mann-Whitney U Test $p = 5.3e^{-49}$), novel ORFs (Mann-Whitney U Test $p = 1.8e^{-39}$), and isoform ORFs (Mann-Whitney U Test $p = 3.3e^{-26}$; Fig. 2D). Moreover, the predicted secondary structure around the start of the annotated ORFs was not significantly higher than that of novel ORFs (Mann-Whitney U Test $p = 0.24$) or isoform ORFs (Mann-Whitney U Test $p = 0.052$). Collectively, retapamulin-arrested ribosomes generated ERFs that exhibit the content and context characteristics of genuine translation initiation sites.

Although our data suggested that isoform ORFs and novel ORFs are genuine, we could not exclude the possibility that treatment of cells with retapamulin diverted ribosomal subunits to bind cryptic translation start sites, or to bind sites without transitioning to a productive translation elongation phase. To determine if isoform ORFs and novel ORFs are fully translated in cells not treated with retapamulin, we examined our standard ribosome profiling data generated from cells without drug treatment. We reasoned that genuine ORFs would be associated with increased ribosome occupancy at start and stop codons, as we and others have observed for annotated ORFs (Fig. S1) (*12*, *13*). We observed enriched ribosome occupancy signal

at the start and stop codons of the annotated genes we identified (Fig. 3A-D). We also observed enriched ribosome occupancy at the start codons of isoform ORFs (Fig. 3D-F), and at the start and stop codons of novel ORFs (Fig. 3A, C-D, G-H). Thus, our data indicate that most novel ORFs are actively translated from start to stop codon, independent of retapamulin treatment. The stop codons of isoform ORFs are expected to have strong ribosome occupancy signal since they are also the stop codons of annotated genes; hence, we cannot use this approach to determine whether isoform ORFs are fully translated, although our data indicate that their start codons are occupied in the absence of retapamulin (Fig. 3C, E-F). The ribosome signal at the start codons of isoform ORFs, and at the start/stop codons of many novel ORFs, is likely to be lower relative to local ribosome signal when compared to that of annotated genes, because all isoform ORFs and 80% of novel ORFs partially or completely overlap annotated genes on the same strand, and annotated genes are more highly expressed as a group than isoform ORFs or novel ORFs (Fig. S5).

In addition to the 996 ORFs predicted from duplicate Ribo-RET datasets, we identified 2,053 putative ORFs from only a single Ribo-RET experiment. We refer to these as "lower-confidence" ORFs. The estimated FDRs for lower-confidence ORFs (8% and 31% for each respective replicate) are sufficiently low that we hypothesized this list contains many genuine ORFs. Consistent with this expectation, 478 correspond to annotated genes. Of the remainder, 580 are isoform ORFs (including 12 isoforms of novel ORFs identified from the high-confidence set of putative ORFs), and 995 are novel ORFs (Fig. S6). The predicted secondary structure for a set of random genomic sequences was significantly higher than the predicted secondary structure around the start of the lower-confidence annotated ORFs (Mann-Whitney U Test $p = 7.3e^{-38}$), lower-confidence novel ORFs (Mann-Whitney U Test $p = 7.0e^{-43}$), and lower-confidence isoform ORFs (Mann-Whitney U Test $p = 3.8e^{-18}$; Fig. S7). Moreover, the predicted secondary structure around the start of the lower-confidence annotated ORFs was not significantly lower than that of lower-confidence novel ORFs (Mann-Whitney U Test $p = 0.09$). However, the predicted secondary structure around the start of lower-confidence annotated ORFs was significantly lower than that of lower-confidence isoform ORFs (Mann-Whitney U Test $p = 1.4e^{-09}$), suggesting a larger number of false positives in this set relative to the high-confidence isoform ORFs. We also found that regions upstream of lower-confidence isoform ORFs and novel ORFs are associated with an enrichment of AGG and GGA sequences in the expected location of a Shine-Dalgarno sequence (Fig. S4). Lastly, we examined ribosome occupancy at the start and stop codons of the lower-confidence ORFs from our ribosome profiling data generated from cells treated without drug treatment. As expected, lower-confidence ORFs matching annotated genes showed enriched ribosome occupancy at their start and stop codons (Fig. S8A-C). Many


A
Sequence Read Coverage (Ribosome Profiling)
Sequence Read Coverage (Ribo-RET)
100 bp
Rv0211
B
Position Relative to Putative Start Codon (nt)
Position Relative to Putative Stop Codon (nt)
-500
0
+500
50 codons
C
Average Normalized Ribosome Occupancy
Position Relative to Start Codon (nt)
Annotated ORFs
Novel ORFs
Isoform ORFs
D
Position Relative to Stop Codon (nt)
E
Sequence Read Coverage
Rv2374c
F
G
2376456 (+)
H

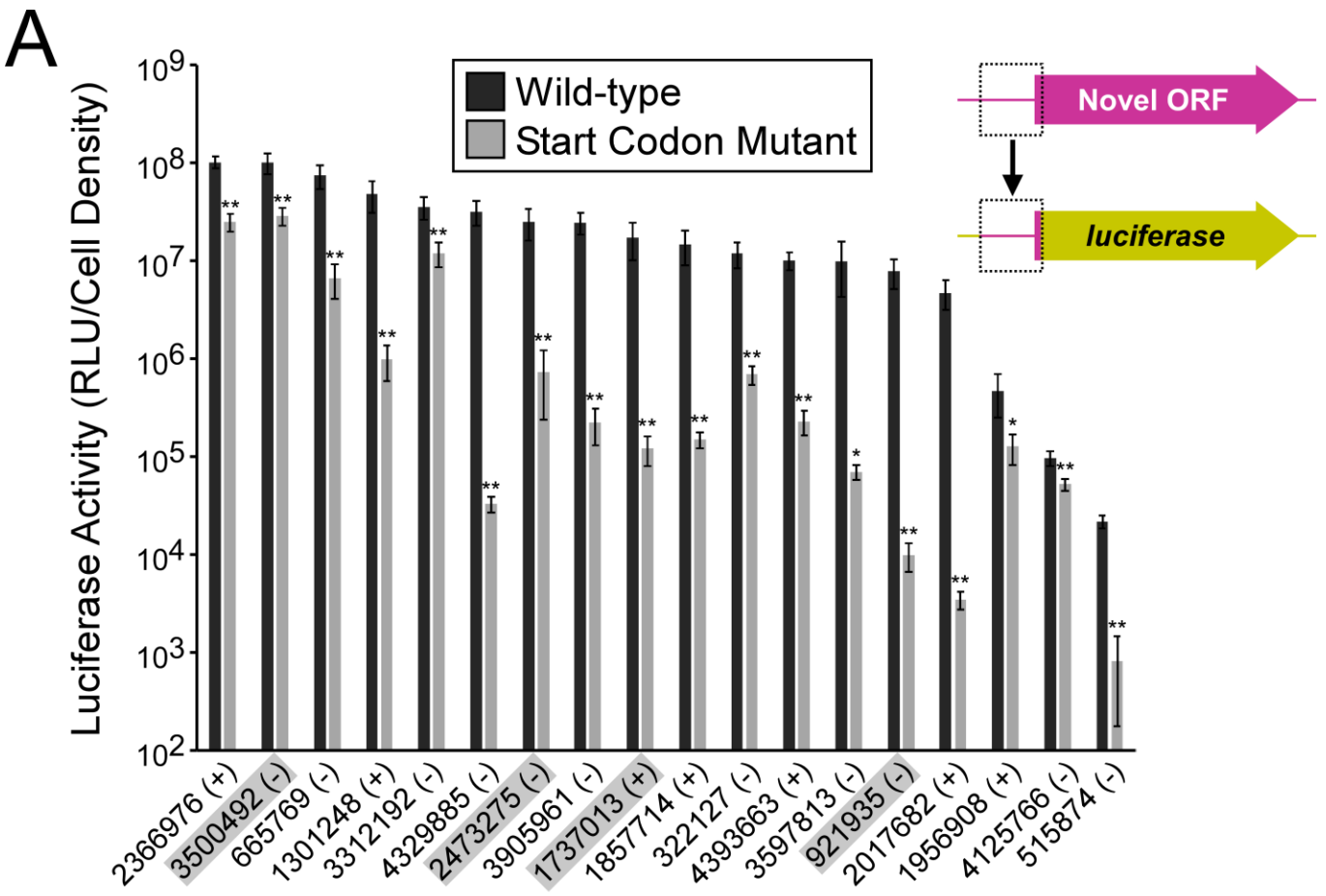
155

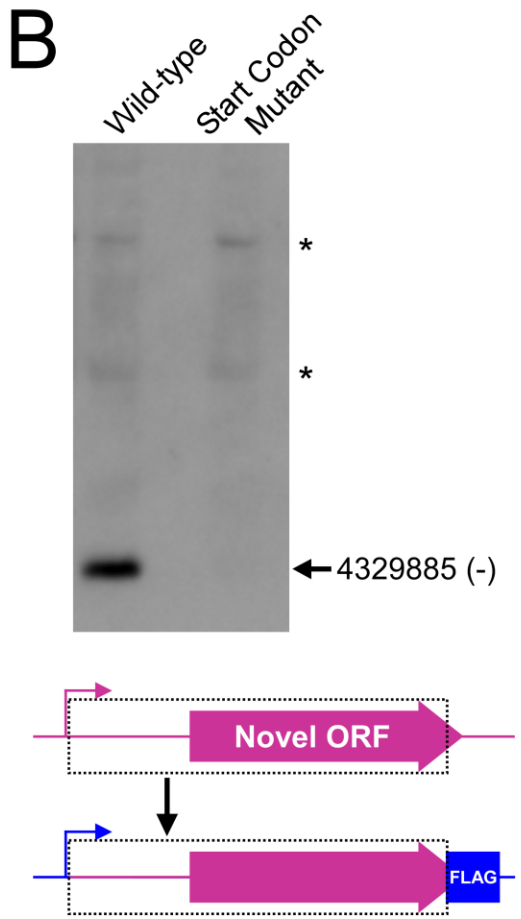
Figure 3. Widespread translation of isoform and novel ORFs. (A) Ribosome profiling (without drug treatment) and Ribo-RET sequence read coverage across the Rv0211 gene, an example of a gene in the "annotated" category of ORFs identified by Ribo-RET. Sequence read coverage for ribosome profiling data is shown in peach. Sequence read coverage for Ribo-RET data is shown in black. Note that Ribo-RET coverage is plotted only for the 3' end of sequence reads, whereas Ribosome profiling coverage is plotted for entire reads. The Rv0211 gene and a gene-internal novel ORF are indicated below the graph by a black and pink arrow, respectively. **(B)** Heatmaps showing normalized sequence coverage from ribosome profiling data for *M. tuberculosis* cells without drug treatment. Sequence coverage is shown for the regions from -500 bp to +500 bp relative to the putative start (left panel) and stop (right panel) codons of annotated ORFs identified by Ribo-RET. The plots below the heatmaps show the expected position of start codons (red datapoints) and stop codons (black datapoints). The position of ORFs with 50 codons is indicated by a dashed vertical line. **(C)** Average normalized coverage for all annotated, isoform, and novel ORFs identified by Ribo-RET in the region from -50 to +50 bp relative to the putative start codon. **(D)** Average normalized coverage for all annotated and novel ORFs identified by Ribo-RET in the region from -50 to +50 bp relative to the putative stop codon. **(E)** Same as panel (A) but for an isoform ORF (indicated by a tallow arrow) and the overlapping, full-length ORF Rv2374c, both of which were identified by Ribo-RET. **(F)** Same as panel (B) but for isoform ORFs (start codons only). **(G)** Same as panel (A) but for a novel ORF. **(H)** Same as panel (B) but for novel ORFs.

isoform lower-confidence ORFs showed enriched ribosome occupancy at their start codons (Fig. S8B-D), and many novel lower-confidence ORFs showed enriched ribosome occupancy at their start and stop codons (Fig. S8B-C, E). We conclude that our high confidence lists of isoform ORFs and novel ORFs represent only a subset of all such ORFs that can be detected using Ribo-RET. As we observed for high-confidence novel ORFs, many of the lower-confidence novel ORFs are short, with 76% being <50 codons in length.

Given that there are (i) >300 novel leaderless ORFs, (ii) almost as many novel leadered ORFs as annotated ORFs in our high-confidence set, and (iii) hundreds of novel leadered ORFs with the expected start/stop codon ribosome occupancy in our lower-confidence set, we conclude that there are well over 1,000 novel ORFs in *M. tuberculosis*. Indeed, the relative numbers of ORFs we detect from the different categories suggests that there are fewer annotated ORFs than there are novel and isoform ORFs combined (Fig. 2C and S6). Since we detect strong ribosome occupancy at both the start and stop codons of most novel ORFs, we can conclude that these ORFs are translated to completion. The distribution of ribosome occupancy levels for novel ORFs is significantly lower than that of the annotated ORFs we detected (Mann Whitney U test $p = 2.3e^{-38}$). Nonetheless, there are many highly expressed novel ORFs; of the 249 high confidence ORFs we identified with the top 25% values for ribosome occupancy, 35 are novel ORFs.

Our data also suggest that hundreds of genes have additional isoforms due to in-frame internal start codons, with some genes having as many as four different possible isoforms. The distribution of ribosome occupancy levels (a combination of RNA level and translation efficiency) for isoform ORFs is significantly lower than that of the annotated ORFs we detected (Fig. S5; Mann-Whitney U Test $p = 6.5e^{-25}$). Indeed, for the 23 instances where we detected the start of an annotated ORF and an associated isoform ORF, the full-length ORF was associated with an average of 28 times higher Ribo-RET signal than the isoform ORF. Nonetheless, three of the isoform ORFs were associated with <2-fold lower Ribo-RET signal than the corresponding full-length ORF. 100 of the 216 isoform ORFs initiate within ±200 nt of the start of an annotated gene. These ORFs are significantly more highly translated than other isoform ORFs (Mann Whitney U Test $p = 3.9e^{-07}$), suggesting that many represent primary product reannotations of the full-length gene rather than secondary isoforms. Consistent with this interpretation, isoform ORFs that initiate within ±200 nt of an annotated gene are less likely to overlap an annotated gene whose start codon was identified in our dataset (Fisher's Exact Test $p = 0.010$), although there are five such isoform ORFs.


A
Luciferase Activity (RLU/Cell Density)
Wild-type
Start Codon Mutant
Novel ORF
luciferase
2366976 (+)
3500492 (-)
665769 (-)
1301248 (+)
3312192 (-)
4329885 (-)
2473275 (-)
3905961 (-)
1737013 (+)
1857714 (+)
322127 (-)
4393663 (+)
3597813 (-)
921935 (-)
2017682 (+)
1956908 (+)
4125766 (-)
515874 (-)


B
Wild-type
Start Codon Mutant
4329885 (-)
Novel ORF
FLAG

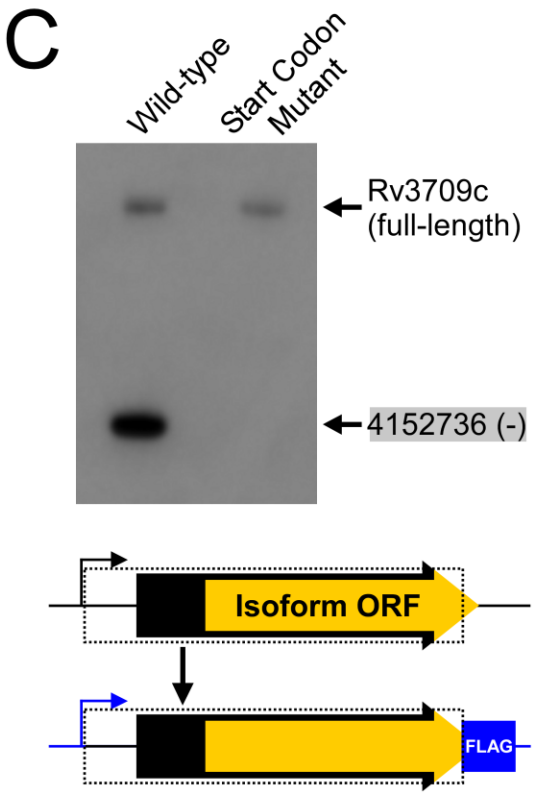

C
Wild-type
Start Codon Mutant
Rv3709c
(full-length)
4152736 (-)
Isoform ORF
FLAG

Figure 4. Validation of selected novel and isoform ORFs. (A) Luciferase reporter assays for constructs consisting of the regions surrounding novel ORF start codons fused translationally to a luciferase reporter gene. The region cloned from the chromosome is indicated by a dashed box. Wild-type and mutant start codon reporter constructs were integrated into the *M. smegmatis* chromosome. The genome coordinate and strand of each selected novel ORF start codon is indicated. Shaded coordinates indicate novel ORFs identified in a single Ribo-RET replicate. **(B)** Western blot with anti-FLAG antibody to detect a FLAG-tagged novel ORF integrated into the *M. smegmatis* chromosome with either an intact (wild-type) or mutated start codon. The integrated constructs included the entire 5' UTR and open-reading frame (indicated by a dashed box), but not the native promoter. The band corresponding to the tagged novel ORF is indicated with an arrow. Asterisks indicate the positions of common cross-reacting proteins. **(C)** Western blot with anti-FLAG antibody to detect a FLAG-tagged isoform ORF integrated into the *M. smegmatis* genome with either an intact (wild-type) or mutated start codon. The integrated constructs included the overlapping full-length, annotated ORF and its entire 5' UTR. Bands corresponding to the tagged full-length and isoform ORFs are indicated.

We sought to validate selected novel and isoform ORFs. We initially tested 18 novel predicted start codons by fusing them to a luciferase reporter ORF, and including 25 bp of upstream context for each site. We constructed equivalent reporter fusions with a single base substitution in the predicted start codon. The reporter plasmids were integrated into the chromosome of *M. smegmatis*. Luciferase expression from all of the 18 luciferase fusions, including four from our lower-confidence list, was significantly reduced by mutation of the start codon (Fig. 4A; $p < 0.05$ or 0.01, as indicated, one-way Student's T-test). Thus, these data demonstrate that Ribo-RET identified functional start codons. To assess translation of putative ORFs with more sequence context, we generated constructs for complete novel ORFs (see Methods for details), with 3 x FLAG tags fused at the C-terminus. We generated equivalent constructs with a single base substitution in the putative start codon. The tagged constructs were integrated into the chromosome of *M. smegmatis*. Three ORFs were detected by Western blot, and their expression was abolished in the mutant constructs (Fig. 4B; Fig. S9). We generated equivalent 3 x FLAG-tagged strains for putative isoform ORFs. We detected the overlapping, full-length protein by Western blot, and expression of these full-length proteins was unaffected by mutation of the isoform ORF start codon (Fig. 4C; Fig. S9). In two cases, we also detected a protein of smaller size, corresponding to the expected size of the isoform protein; expression of these small isoform proteins was not detected in the mutant constructs (Fig. 4C; Fig. S9). Thus, these data strongly support the ORF predictions from the Ribo-RET data. Moreover, the luciferase reporter data and the Western blot data support the existence of novel and isoform ORFs identified from only a single replicate of Ribo-RET data (Fig. 4A, C; Fig. S9), consistent with the idea that there are many novel and isoform ORFs beyond the high-confidence set.

The *M. tuberculosis* genome has an unusually high G/C content (65.6%). Consequently, there is a G/C bias within codons of annotated genes. Specifically, the second position of codons is biased towards A/T relative to the overall genome G/C content, and the wobble position is biased towards G/C; this bias reflects purifying selection on the specific codons within an ORF (*23*). We reasoned that if all of the novel ORFs were functional, this group would show a codon bias similar to that of annotated genes. The G/C content at the second and third positions of codons for novel ORFs (excluding those that overlap annotated ORFs) differs modestly, but significantly, from the overall G/C content (Binomial Test $p = 2.6e^{-03}$ and $p = 0.03$ for second and third codon positions, respectively; n = 2085 codons; Fig. 5), but the differences are greatly and significantly smaller than for the annotated ORFs we identified in our dataset (Chi Squared Test $p < 1e^{-05}$ and $p < 1e^{-05}$ for second and third codon positions, respectively; Fig. 5). We observed similar position-specific G/C content biases within codons for lower-confidence ORFs (Fig. 5). These data suggest that the large majority of the novel ORFs (in this non-

	123	123	123	123	123	123	123	123	123	123
ORF #1	**GCA**	**AAG**	**ACC**	**CAG**	**GTC**	**AAT**	**GTC**	**CAG**	**GTG**	**TTC**
ORF #2	**GCG**	**GTC**	**GGC**	**GAT**	**GAT**	**AAG**	**GCA**	**AGG**	**TGC**	**GCG**

G/C Frequency at Codon Position

1	2	3
6/10=0.6	2/10=0.2	8/10=0.8
7/10=0.7	6/10=0.6	7/10=0.7

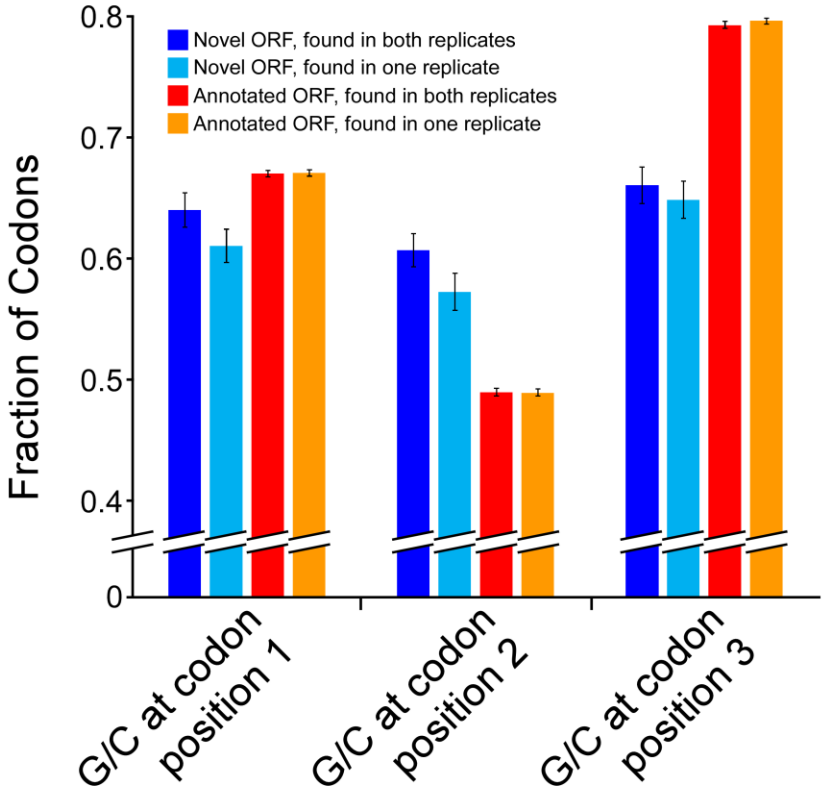


Figure 5. G/C skew within codons of novel and annotated ORFs. The schematic shows how G/C counts would be calculated for two example 12-codon ORFs (note that only 10 codons are shown because start and stop codons were excluded from the analysis). The histogram shows the frequency of G/C nucleotides at each of the three codon positions for (i) novel ORFs identified from two Ribo-RET replicates, (ii) novel ORFs identified from a single Ribo-RET replicate, (iii) annotated ORFs identified from two Ribo-RET replicates, and (iv) annotated ORFs identified from a single Ribo-RET replicate. Note that only novel ORFs that do not overlap an annotated ORF were analyzed.

overlapping subset) have not experienced purifying selection. The small size of novel ORFs provides too little information content to assess G/C content biases of most individual sORFs, and with over a thousand actively translated sORFs, a small functional fraction should not be ignored. Indeed, a few novel sORFs show particularly striking G/C content biases within their codons, highlighting these as likely candidates for functional ORFs (Table S3), and indicating the likely importance of small proteins in bacterial physiology.

A recent study used Ribo-RET to identify novel and isoform ORFs in *E. coli* (*20*). Although Ribo-RET has been applied to only two bacterial species to date, the abundance of novel and isoform ORFs in two distantly related species suggests that many bacterial transcriptomes are pervasively translated. This is analogous to the phenomenon of pervasive transcription that has been described in numerous bacterial species, whereby spurious promoters scattered throughout the coding sequences of bacterial genomes drive transcription of short, non-coding RNAs (*24*, *25*). Due to the short length and low translation initiation levels of proteins generated by pervasive translation (Fig. 3, S5, S8) (*18*, *20*), they likely represent a small fraction of total cellular translation. Moreover, our analysis of G/C content within codons strongly suggests that most of the proteins themselves are unlikely to be functional in *M. tuberculosis* (Fig. 5). We suggest that this lack of function also applies to most of the equivalent proteins identified in *E. coli* (*18*, *20*), although an equivalent analysis of G/C content within codons cannot be applied to *E. coli* data since the G/C content of the genome is 50.7%. This raises the question of whether pervasive translation has a function in bacteria. We propose that it is an inevitable consequence of the relatively low specificity associated with the selection of start codons by initiating ribosomes: at least five possible start codon choices (AUG, GUG, UUG, AUU, AUC) (*9*, *26*), limited sequence specificity in the Shine-Dalgarno sequence (*27*), and considerable flexibility in the Shine-Dalgarno to start codon distance (*28*). The cost of active spurious start codons must be balanced with the beneficial functions that these features are providing in their larger context. We propose that the proteins generated by pervasive translation provide a rich source for the creation of novel, functional proteins, especially for small proteins, whose importance to bacterial physiology and gene regulation is becoming increasingly apparent (*29*, *30*). Strikingly, far fewer new ORFs were identified in *E. coli* than we found in *M. tuberculosis*, with only ~3% and ~7% of all *E. coli* ORFs being isoform and novel ORFs, respectively (*20*). The much greater extent of pervasive translation in *M. tuberculosis* relative to *E. coli* suggests that the selective pressure against pervasive translation is lower in some bacterial species than in others.

MATERIALS AND METHODS

Strains and plasmids

Ribosome profiling and Ribo-RET experiments were performed using the *M. tuberculosis* strain mc^{2}7000 (*31*). All oligonucleotides used in this study are listed in Table S4. *M. tuberculosis* mc^{2}7000 cells were grown in 7H9 medium supplemented with 10% OADC (Oleic Albumin Dextrose Catalase), 0.2 % glycerol, 100 ug/ml pantothenic acid and 0.05 % Tween80 at 37 °C, without shaking.

We constructed a shuttle vector, pRV1133C, to allow integration of luciferase or FLAG-tag fusion constructs into the *M. smegmatis* mc^{2}155 (*32*) chromosome, with a constitutive promoter driving transcription. pRV1133C was derived from pMP399, retaining its *oriE* for episomal maintenance in *E. coli*, its integrase and *attP* site for integration at the L5 *attB* site in mycobacteria, and apramycin resistance (*33*). The *hsp60* promoter of pMP399 was replaced by the promoter of the *M. tuberculosis Rv1133c* (*metE*) gene. The criterion for selecting *Rv1133c* was its strong constitutive expression assessed by transcription start site metrics (*9*).

A luciferase (NanoLuc) gene amplified from pNL1.1 (Promega, cat no 1001) was cloned downstream of the *Rv1133c* promoter to generate pGE190. To construct individual reporter fusion plasmids, the entire pGE190 plasmid was amplified by inverse PCR using Q5 High Fidelity DNA polymerase (NEB) with oligonucleotides TGD4006 and TGD5162. Sequences corresponding to the 25 bp upstream of, and including the start codons for selected ORFs were PCR-amplified using oligonucleotide pair TGD5163 and TGD5164, to amplify template oligonucleotides TGD5165-5173, TGD5175, and TGD5178-5186. PCR products were cloned into the linearized pGE190 using the In-Fusion cloning system (Takara). The oligonucleotide templates had a “Y” (mixed base “C” or “T”) at the position corresponding to the central position of the start codon. Clones were sequenced to identify wild-type and mutant constructs, where the central position of the start codon was a “T” or a “C”, respectively. Plasmid DNA was electroporated into *M. smegmatis* mc^{2}155 for chromosomal integration before assaying luciferase activity.

A 3 x FLAG-epitope-tag sequence was integrated into pRV1133C to generate pGE450. To construct individual FLAG-tagged constructs, the entire pGE450 plasmid was amplified by inverse PCR using Q5 High Fidelity DNA polymerase

(NEB) with oligonucleotide TGD4981 and TGD4982. Sequences from predicted transcription start sites to immediately before the stop codon for selected ORFs were PCR-amplified using oligonucleotide pairs TGD5208 and TGD5209, TGD5216 and TGD5217, TGD5241 and TGD5242, or TGD5247 and TGD5248. PCR products were cloned into pGE450 using the In-Fusion cloning system (Takara). Start codon mutant constructs were made by inverse PCR-amplification of the wild-type constructs using primers that introduce a start codon mutation ("T" to "C" change at the central position of the start codon; oligonucleotides TGD5210, TGD5211, TGD5218, TGD5219, TGD5256, TGD5257, TGD5258, and TGD5259). PCR products were treated with *Dpn*I and cloned using the In-Fusion cloning system (Takara). Plasmid DNA was electroporated into *M. smegmatis* mc^2155 for chromosomal integration before performing expression analysis by Western blot.

Ribosome profiling without drug treatment

10 ml of *M. tuberculosis* (OD_{600} of 0.4) was used to inoculate 400 ml of medium and grown to an OD_{600} of 1 (2-3 weeks). Cells were collection by filtration through a 0.22 µm filter. Libraries were prepared for sequencing as described previously for *M. smegmatis* (*9*).

Ribosome profiling with retapamulin treatment (Ribo-RET)

10 ml of *M. tuberculosis* (OD_{600} of 0.4) was used to inoculate 400 ml of medium and grown to an OD_{600} of 1 (2-3 weeks). Cells were treated with retapamulin (Sigma CDS023386) at a final concentration of 0.125 mg/ml for 15 minutes at room temperature, with occasional manual shaking, and collected by filtration through a 0.22 µm filter. Cells were flash frozen in liquid nitrogen with 0.7 ml lysis buffer (20 mM Tris pH 8.0, 10 mM MgCl2, 100 mM NH4Cl, 5 mM CaCl2, 0.4 % Triton X100, 0.1 % NP-40, 1 mM chloramphenicol, 100 U/mL DNase I). Frozen cells were milled using a Retsch MM400 mixer mill for 8 cycles of 3 minutes each at 15 Hz. Milling cups were re-frozen in liquid nitrogen in between each milling cycle. Cell extracts were thawed and incubated on ice for 30 min. Samples were clarified by centrifugation, supernatants removed and quantified by nanodrop, and 1 mg aliquots of cell extracts were flash-frozen in liquid nitrogen. Monosomes were isolated by digesting 1 mg of cell extract with 1,500 units of micrococcal nuclease for 1 hour at room temperature on a rotisserie rotator. The reaction was quenched by adding 2 µl 0.5 M EGTA, after which the digest was fractionated through a 10-50 % sucrose gradient. Fractions from the sucrose gradients were electrophoresed on a 1 % agarose gel with 1 % bleach to identify

ribosomal RNA peaks. Those fractions were selected, pooled, and monosomes isolated by acid phenol:chloroform extraction and isopropanol precipitation.

Libraries for sequencing were prepared using a previously described method (*34*). RNA from monosomes was run on a 15% denaturing gel alongside a 31 nt RNA oligonucleotide to size-select 31±5 nt fragments. Samples were gel-extracted in 500 ml RNA gel extraction buffer (300 mM NaOAc (pH 5.5), 1 mM EDTA, 0.1 U/mL SUPERase-In RNase inhibitor) followed by isopropanol precipitation. The samples were dephosphorylated by incubating with 10 U of T4 Polynucleotide Kinase (NEB) for 1 hour at 37 °C, before extraction with phenol:chloroform:isoamyl alcohol, and ethanol precipitation. The dephosphorylated RNAs were ligated to the 3' linker oligonucleotide JW9371 using T4 RNA Ligase 2 (truncated, K227Q) at a 1:4 RNA:linker ratio. The ligation reactions were incubated for 3 hours at 37 °C, followed by 20 minutes at 65 °C. The reactions were separated on a 15% polyacrylamide denaturing gel with a control RNA oligonucleotide (JW9370) of the expected size of the ligated product. The RNA-ligation products were excised and extracted in 500 µL RNA extraction buffer and concentrated by ethanol precipitation. Reverse transcription was performed on the RNA samples using Superscript III (Life Technologies) and oligo JW8875, as described previously (*34*). The reactions were separated through a 10% polyacrylamide denaturing gel and cDNAs excised and extracted in 500 µL DNA extraction buffer (300 mM NaCl, 10 mM Tris-Cl (pH 8), 1 mM EDTA). Reverse-transcribed cDNA was circularized using CircLigase, and PCR-amplified as described previously (*34*). Between 4 and 9 cycles of PCR were performed using Phusion High Fidelity DNA Polymerase, JW8835 as the standard forward primer and JW3249, 3250, 8876 or 8877, corresponding to Illumina index numbers 1, 2, 34 or 39 respectively, as the reverse primer. Samples were separated through an 8% polyacrylamide acrylamide gel. DNAs of the appropriate length (longer than the control adapter band) were excised from the gel and extracted in 500 µL of DNA extraction buffer. DNAs were concentrated by isopropanol precipitation. Samples were quantified and subject to DNA sequence analysis on a NextSeq instrument.

Inferring ORF positions from ribosome profiling data

Using sequencing data from Ribo-RET, reads were trimmed to remove adapter sequences using a custom python script. Reads were then aligned to the *M. tuberculosis* genome (NCBI Reference Sequence: NC_000962.3) to identify genomic strand and position information corresponding to the 3' ends of ribosome-protected RNA fragments. We refer to the number of 3' end positions at any given genomic position as the "footprint 3' coverage". Footprint 3' coverage values were set to

zero for specific positions: (i) all tRNA genes, rRNA genes, and the first 50 nt of putative leaderless ORFs (i.e. transcription start sites where the first three nucleotides are RUG) identified by (*9*); and (ii) all genomic positions where footprint 3' coverage was not higher than 90% of the genomic positions from 50 nt upstream to 51 nt downstream, on the same strand. We then selected genome positions with footprint 3' coverage above a threshold value (22 reads for replicate #1 and 10 reads for replicate #2; these values are proportional to the total number of aligned sequence reads for each replicate). We further selected genome positions that were the highest value in the 41 nt region centered on the position being analyzed, on the same strand. All selected genome positions were designated as ERFs. ERFs were inferred to represent an ORF if the ERF position was 15 nt downstream of an ATG, GTG or TTG, 16 nt downstream of an ATG or GTG, and 14, 17, or 18 nt downstream of an ATG. These trinucleotide sequences and distances were selected based on a >2-fold enrichment upstream of ERFs (Fig. 2B).

Calculating False Discovery Rates for ORF prediction from Ribo-RET data

The likelihood of randomly selecting a genome coordinate with an associated start codon sequence (as defined above for EFRs) was estimated by repeating the ERF and ORF identification analysis using genome sequences rotated by 1,000 bp at a time, 100 times. We refer to this frequency as "R". We refer to the total number of EFRs as "T". We refer to the number of EFRs associated with a putative start codon as "S". We refer to the number of falsely discovered start codons as "F", and the number of correctly discovered start codons as "C". Based on these definitions, we can assume the following:

$$S = F + C$$

$$R(T - C) = F$$

Combining these two equations gives:

$$F = R(T - S)/(1 - R)$$

Since we have experimentally determined values for R, T and S, we can solve this equation to determine the value of F, and then determine the value of C. For analysis of the two replicates combined, "T" was estimated by combining the number of EFRs identified in each replicate and removing any from the second replicate that were within 3 nt of an EFR in the first

replicate. The FDRs calculated using this approach are based on the assumption that there are no systematic biases in the data that lead to EFRs being located ~15 nt upstream of potential start codons.

RNA folding prediction

The sequence from -40 to +10 relative to each start codon, or for 500 x 50 nt sequences randomly selected from the *M. tuberculosis* genome, were selected for prediction of the free energy of the predicted minimum free energy structure using ViennaRNA RNAfold, version 2.1 (*35*) through the Galaxy web server (usegalaxy.org) (*36*) using default settings, except 310.15 K for temperature.

Analysis of ribosome profiling sequence coverage around start and stop codons

Sequence reads from ribosome profiling datasets (without drug treatment) were trimmed at their 3' ends to the position immediately upstream of the first instance of "AAA" using a custom Python script. "AAA" was selected because the RNA fragments were polyadenylated at their 3' ends, and "AAA" sequences are very rare in *M. tuberculosis*. Reads lacking an "AAA" sequence were discarded. Trimmed sequence reads were then aligned to the *M. tuberculosis* genome (NCBI Reference Sequence: NC_000962.3), and genome coverage values were determined, using Rockhopper (*37*). Coverage scores were selected for regions ±500 nt relative to selected start or stop codons. All values were normalized to the highest value in the 1,001 nt range. Heatmaps were generated using TreeView (*38*).

Analysis of G/C usage within codons

ORFs were scored at the first, second or third position of all codons for the presence of a G or C, excluding start and stop codons. For statistical comparison of ORFs between classes, the G/C frequency at the second or third position of all ORFs within a category was determined, with the exception of the longest high-confidence novel ORF. We excluded this ORF because it contributes 18% of the codons for the high-confidence novel ORF category, and it represents an isoform of *Rv1807*, a gene not included in the NC_00962.3 annotation but included in other annotations. A cumulative binomial test was then used to compare G/C content at the second or third codon position relative to the genome average. Values plotted in Figure 5 represent the average of values for each individual ORF (including the isoform of *Rv1807*). ORFs with <22 codons (including start and stop codons) were discarded from this analysis, but not from the statistical comparison.

Analysis of EFR position relative to annotated start codons

EFR positions from one Ribo-RET replicate dataset were compared to the start codon positions of all annotated ORFs (NC_000962.3), except for ORFs predicted to be leaderless by (*9*).

Analysis of trinucleotide sequence content upstream of EFRs

The frequency of each trinucleotide was determined for the 50 nt upstream of all EFRs. In parallel, the frequency of each trinucleotide sequence was determined for a set of 1,000 random, 50 nt genomic DNA sequences. The frequency of each trinucleotide sequence was averaged for all positions in the set of random sequences. The frequency of trinucleotide sequences in all EFRs was normalized using the averaged values from the set of random sequences. The heatmap was plotted using TreeView (*38*). The frequency of AGG and GGA trinucleotide sequences upstream of putative start codons was determined similarly.

Luciferase reporter assays

M. smegmatis mc^2155 strains with integrated luciferase reporter constructs were grown in TSB with Tween80 overnight at 37 °C to an OD_{600} of ~1.0. Relative Light Units for each strain were determined using the Nano-Glo Luciferase Assay (Promega) and normalized to the corresponding OD_{600} value. Assays were performed in triplicate (biological replicates).

Western blots

M. smegmatis MC^2155 strains with integrated FLAG-tagged constructs were grown in TSB with Tween overnight at 37 °C to an OD_{600} of ~1.0. Cells were harvested by centrifugation and resuspended in 1 x NuPage LDS sample buffer (Invitrogen) + 5 mM sodium metabisulfite. Samples were heated at 95 °C for 10 min before loading onto a 4-12% gradient Bis-Tris mini-gel (Invitrogen). After separation, proteins were transferred to a nitrocellulose membrane (Life Technologies) or a PVDF membrane (Thermo Scientific). The membrane was probed with a monoclonal mouse anti-FLAG antibody (Sigma). Secondary antibody and detection reagents were obtained from Lumigen (ECL plus kit) and used according to the manufacturer's instructions.

Integrated Genome Browser

All ribosome profiling and Ribo-RET data are available for visualization on our interactive genome browser (*9*): http://mtb.wadsworth.org/

ACKNOWLEDGEMENTS

We thank Mike Palumbo and Dan Muller for computational assistance. We thank Yunlong Li and Yong Yang for technical support. We thank David Grainger for comments on the manuscript. This work was supported by R21AI117158 and R21AI119427 grants from the National Institutes of Health to JTW, KMD and TAG. Raw Illumina sequencing data are available from the ArrayExpress and European Nucleotide Archive repositories with accession number E-MTAB-8039.

REFERENCES

1. M. J. Dark, Whole-genome sequencing in bacteriology: state of the art. *Infect Drug Resist*. **6**, 115–123 (2013).

2. J. Besemer, M. Borodovsky, GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* **33**, W451-454 (2005).

3. A. L. Delcher, K. A. Bratke, E. C. Powers, S. L. Salzberg, Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*. **23**, 673–679 (2007).

4. D. Hyatt *et al.*, Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. **11**, 119 (2010).

5. C. B. Burge, S. Karlin, Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8**, 346–354 (1998).

6. A. Lomsadze, K. Gemayel, S. Tang, M. Borodovsky, Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Res.* **28**, 1079–1089 (2018).

7. H. J. Beck, I. Moll, Leaderless mRNAs in the Spotlight: Ancient but Not Outdated! *Microbiol Spectr*. **6** (2018), doi:10.1128/microbiolspec.RWR-0016-2017.

8. N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, J. S. Weissman, Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*. **324**, 218–223 (2009).

9. S. S. Shell *et al.*, Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLoS Genet.* **11**, e1005641 (2015).

10. T. Cortes *et al.*, Genome-wide Mapping of Transcriptional Start Sites Defines an Extensive Leaderless Transcriptome in Mycobacterium tuberculosis. *Cell Reports*. **5**, 1121–1131 (2013).

11. A. A. Miranda-CasoLuengo, P. M. Staunton, A. M. Dinan, A. J. Lohan, B. J. Loftus, Functional characterization of the Mycobacterium abscessus genome coupled with condition specific transcriptomics reveals conserved molecular strategies for host adaptation and persistence. *BMC Genomics*. **17**, 553 (2016).

12. E. Oh *et al.*, Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell*. **147**, 1295–1308 (2011).

13. C. J. Woolstenhulme, N. R. Guydosh, R. Green, A. R. Buskirk, High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. *Cell Rep*. **11**, 13–21 (2015).

14. Z. Ji, R. Song, H. Huang, A. Regev, K. Struhl, Transcriptome-scale RNase-footprinting of RNA-protein complexes. *Nat. Biotechnol.* **34**, 410–413 (2016).

15. S. Lee *et al.*, Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E2424-2432 (2012).

16. C. Fritsch *et al.*, Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res.* **22**, 2208–2218 (2012).

17. N. T. Ingolia, L. F. Lareau, J. S. Weissman, Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*. **147**, 789–802 (2011).

18. J. Weaver, F. Mohammad, A. R. Buskirk, G. Storz, Identifying Small Proteins by Ribosome Profiling with Stalled Initiation Complexes. *MBio*. **10** (2019), doi:10.1128/mBio.02819-18.

19. S. Meydan, N. Vázquez-Laslop, A. S. Mankin, Genes within Genes in Bacterial Genomes. *Microbiol Spectr*. **6** (2018), doi:10.1128/microbiolspec.RWR-0020-2018.

20. S. Meydan *et al.*, Retapamulin-Assisted Ribosome Profiling Reveals the Alternative Bacterial Proteome. *Mol. Cell* (2019), doi:10.1016/j.molcel.2019.02.017.

21. C. Dingwall, G. P. Lomonossoff, R. A. Laskey, High sequence specificity of micrococcal nuclease. *Nucleic Acids Res.* **9**, 2659–2673 (1981).

22. C. Del Campo, A. Bartholomäus, I. Fedyunin, Z. Ignatova, Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. *PLoS Genet.* **11**, e1005613 (2015).

23. M. J. Bibb, P. R. Findlay, M. W. Johnson, The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene*. **30**, 157–166 (1984).

24. J. T. Wade, D. C. Grainger, Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat Rev Micro*. **12**, 647–653 (2014).

25. M. Lybecker, I. Bilusic, R. Raghavan, Pervasive transcription: detecting functional RNAs in bacteria. *Transcription*. **5**, e944039 (2014).

26. J. K. Sussman, E. L. Simons, R. W. Simons, Escherichia coli translation initiation factor 3 discriminates the initiation codon in vivo. *Mol. Microbiol.* **21**, 347–360 (1996).

27. S. A. Evfratov *et al.*, Application of sorting and next generation sequencing to study 5´-UTR influence on translation efficiency in Escherichia coli. *Nucleic Acids Res.* **45**, 3487–3502 (2017).

28. H. Chen, M. Bjerknes, R. Kumar, E. Jay, Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. *Nucleic Acids Res.* **22**, 4953–4957 (1994).

29. E. C. Hobbs, F. Fontaine, X. Yin, G. Storz, An expanding universe of small proteins. *Curr. Opin. Microbiol.* **14**, 167–173 (2011).

30. M. A. Kriner, A. Sevostyanova, E. A. Groisman, Learning from the Leaders: Gene Regulation by the Transcription Termination Factor Rho. *Trends Biochem. Sci.* **41**, 690–699 (2016).

31. A. K. Ojha *et al.*, Growth of Mycobacterium tuberculosis biofilms containing free mycolic acids and harbouring drug-tolerant bacteria. *Mol. Microbiol.* **69**, 164–174 (2008).

32. S. B. Snapper, R. E. Melton, S. Mustafa, T. Kieser, W. R. Jacobs, Isolation and characterization of efficient plasmid transformation mutants of Mycobacterium smegmatis. *Mol. Microbiol.* **4**, 1911–1919 (1990).

33. S. A. Consaul, M. S. Pavelka, Use of a novel allele of the Escherichia coli aacC4 aminoglycoside resistance gene as a genetic marker in mycobacteria. *FEMS Microbiol. Lett.* **234**, 297–301 (2004).

34. N. T. Ingolia, Genome-wide translational profiling by ribosome footprinting. *Methods Enzymol.* **470**, 119–142 (2010).

35. R. Lorenz *et al.*, ViennaRNA Package 2.0. *Algorithms Mol Biol*. **6**, 26 (2011).

36. E. Afgan *et al.*, The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537–W544 (2018).

37. R. McClure *et al.*, Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res.* **41**, e140 (2013).

38. A. J. Saldanha, Java Treeview--extensible visualization of microarray data. *Bioinformatics*. **20**, 3246–3248 (2004).