# Computational modeling of mRNA degradation in *Mycolicibacterium smegmatis*

by

Huaming Sun

A Dissertation
Submitted to the Faculty of the
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Bioinformatics and Computational Biology

April 29th, 2024

APPROVED BY:

Prof. Patrick Flaherty, Committee Member. Mathematics and Statistics, Umass Amherst

Prof. Lane Harrison, Committee Member. Computer Science, WPI

Prof. Dmitry Korkin, Advisor. Computer Science, WPI

Prof. Scarlet Shell, Advisor. Biology and Biotechnology, WPI

# Abstract

The increasingly severe and frequent drug resistance highlights the need to better understand the stress response strategies that the causative agent of tuberculosis, *Mycobacterium tuberculosis*, employs to successfully adapt and persist within the host. One of the stress response strategies is regulation of mRNA degradation, which can contribute to mycobacterial survival in energy-limited environments by reprogramming gene expression, altering mRNA abundance, and modulating energy usage. However, the regulatory mechanisms that control mRNA degradation are not well understood.

In this work, I investigated mRNA degradation mechanisms in the nonpathogenic model *Mycolicibacterium smegmatis* from two perspectives: a targeted study of the impact of an important RNase, and an agnostic study of the impact of a diverse compendium of mRNA properties on degradation rates. In Chapter 2, we characterized the role and cleavage site preferences of an essential endoribonuclease, RNase E, in mycobacteria. By repressing transcription of *rne*, the gene encoding RNase E, we showed that RNase E has a major impact on mRNA degradation rates transcriptome-wide in *M. smegmatis*. Through the comparison of RNAseq coverage between *rne* knockdown and control strains, we showed that RNase E cleavage regions are enriched for cytidines in both *M. smegmatis* and *M. tuberculosis*, allowing us to attribute to RNase E a number of cleavage sites previously mapped with high resolution *in vivo*. These preferences for cytidines at RNase E cleavage sites were further confirmed *in vitro* for *M. smegmatis*. Together, these findings defined the dominant role of RNase E in transcriptome-wide mRNA degradation along with its cleavage targets at high resolutions in mycobacteria.

In Chapter 3, we developed an experimental and computational framework to identify the intrinsic transcript properties that are associated with transcript stability in *M. smegmatis*. We quantified transcriptome-wide mRNA half-life in log phase growth and hypoxia-induced growth arrest using RNAseq. Through machine learning, we showed that transcript stability is influenced by the collective effect of diverse transcript features. Our results highlighted the impact of 5′ UTRs on the stability of leadered transcripts. We also identified transcript properties whose associations with transcript stability differ between leadered and leaderless transcripts as well as between different growth conditions. In sum, these results provided a comprehensive and enhanced understanding of the impacts of intrinsic transcript features on mRNA degradation rates in *M. smegmatis*.

# Table of Contents

# List of Figures

# List of Supplemental Figures

# Acknowledgements

To my two advisors, Dr. Scarlet Shell and Dr. Dmitry Korkin, whose guidance helped me navigate through all the turmoil to finish my PhD. Scarlet, thank you for being the person I leaned on countless times over the years. Dmitry, thank you for all the inspiring conversations and all the valuable experience you shared with me. I am extremely fortunate to have both of you standing by my side, patiently allowing me to explore and learn, to make mistakes and improve, and teaching me everything I need to know to grow as a scientist, maybe one day, a scientist like you.

To my dissertation committee members, Dr. Patrick Flaherty and Dr. Lane Harrison, thanks for your incredible patience, support and understanding. Thanks for all the valuable questions, feedbacks to help me ground my work.

To the Shell lab, thank you all for listening to me talking about machine learning over the years. To the Korkin lab, thank you all for listening to me talking about mycobacteria over the years. Thanks all of you for giving me the opportunity to learn from you, for helping me shape my work, for supporting me, and for bringing laugher and joy to my PhD life. You are all amazing to work with.

To my parents, thanks for the unconditional support and for pushing me to where I am today. To my wonderful wife and two precious girls, thanks for being the excuses for my procrastination. You are the people who motivate me, and put a smile on my face every day.

To all the lovely people in Biology and Biotechnology Department at WPI, thank you all for making the department feel like a home for everyone. Thanks for all of your insights to help me improve my work.

To *Mycolicibacterium smegmatis*, what a journey we had together. Thanks for the company, all the beautiful and painful memories, which I will cherish for the rest of my life.

To everyone annoyed by my impatience, grumpiness, mumbling, reticence, irregular emotional highs among constant lows over the years, I apologize. However, it is just part of my PhD, which I am proud to say that I have finished.

# Chapter 1 : Review of computational modeling to depict mRNA degradation in bacteria

# Introduction

As the intermediate product between DNA and protein, mRNA plays a critical role in expressing genes and maintaining basic cell functions. The abundance of each mRNA is the result of the balance between its transcription and its degradation. Compared to transcription, regulation of mRNA degradation is still poorly understood yet essential for various bacteria. As a stress response mechanism, regulation of mRNA degradation, specifically mRNA stabilization, is widely used by bacteria to survive energy-limited microenvironments. Given the increasingly severe antibiotic resistance problem, a better understanding of mRNA degradation mechanisms is important for more broadly understanding stress response strategies in bacteria, which is needed for development of novel therapeutics. Here we will briefly review the current knowledge of mRNA degradation processes and the factors that might affect it in bacteria, and then discuss the computational models that have been developed to facilitate understanding of mRNA degradation.

# Overview of mRNA degradation in bacteria and its influencing factors

In many bacteria including *E. coli* and mycobacteria, mRNA degradation is thought to be initiated by endonucleolytic cleavage by the key enzyme RNase E, followed by exonucleolytic cleavage by other enzymes that cleave at either the 5′ or 3′ ends of the RNA fragments. The critical role of RNase E in this process is evidenced by studies in which its deletion or depletion in *E. coli* and mycobacteria causes global increases in mRNA half-life (1-3). Therefore, factors related to the interactions between RNase E and mRNAs are often incorporated into models of the mRNA

degradation. Here, we are listing some of these factors to be considered and will discuss them further in the section on modeling.

It is thought that one of the factors affecting mRNA degradation rates in bacteria is the steady-state abundance of mRNAs. This idea arose from multiple observations of inverse relationships between mRNA abundance and mRNA half-life in bacteria (3-13). It has been proposed that higher mRNA abundance could increase the probability of RNase E binding to RNAs, leading to faster degradation (5,8). As for the binding itself, it is known that RNase E has preferences for cleaving in certain sequence contexts (1,14-16). Therefore, the actual cleavage activity also depends on the accessibility of the RNA to RNases at these preferred sequence locations. Specific RNA regions may be less accessible if they are folded into secondary structure or bound by small RNAs, since RNase E only cleaves single-stranded RNA (reviewed in (17)). RNAs may also be protected by RNA binding proteins or ribosomes, which can physically block access of RNases. In other cases, factors such as small RNAs may specifically trigger degradation of mRNAs by recruiting RNases (reviewed in (17)), while stalled ribosomes may trigger mRNA degradation by recruiting ribosome rescue factors (reviewed in (17)). The 5′ UTRs of mRNAs have been shown to impact degradation in many cases (reviewed in (18,19)), and this raises interesting questions for mycobacteria in particular, since mycobacteria encode many leaderless genes lacking 5′ UTRs.

These are just some examples illustrating the wide-ranging biological factors that could be involved in the complicated process of mRNA degradation. Besides the diversity of the potential factors that impact mRNA degradation, the unknown underlying mechanisms by which they interact with each other make the modeling mRNA degradation inherently complicated. Furthermore, growth and stress conditions are known to impact mRNA degradation rates in many

bacteria (9,13,20,21), and such adaptation to microenvironments may involve different regulatory mechanisms and thereby add addition layers of complexity. Ultimately, understanding of the mRNA degradation mechanisms and their regulation in bacteria may require a combination of multiple models to better describe the various aspects involved.

## Methods for measuring mRNA degradation in bacteria

Regardless of the specific computational method being used, modeling of mRNA degradation relies on experimental measurements of mRNA degradation rates. The most commonly used approach in bacteria is to inhibit transcription initiation with the antibiotic rifampicin, then quantify residual mRNA abundance over time to estimate degradation rates. mRNA abundance can be measured for individual genes by northern blotting or quantitative PCR, and on transcriptome-wide scales by RNAseq. Given the fact that abundance data is collected separately at each time point, normalization must be employed to allow quantitative comparisons between time points. Abundance data are then fit to single-exponential fits or more complicated models to calculate half-lives for each gene and, in transcriptome-wide studies, generate a portrait of mRNA degradation in a given experimental setup.

There are two major caveats to the method described above. One is that rifampicin blocks transcription initiation but not transcription elongation, leading to delays in the observed effects on mRNA abundance, as continued elongation produces new transcript after addition of rifampicin (reviewed in (6,22-25)). The resulting biphasic degradation trend complicates the modeling of degradation, especially for the models assuming a constant degradation rate from the beginning of the abundance measurements. The other major limitation of the rifampicin

method is that transcriptional block is a dramatic perturbation to cellular physiology, and the mRNA degradation observed during this treatment may therefore not faithfully reflect that which would occur in an unperturbed cell. We and other investigators have reported changes in apparent mRNA degradation rates within minutes after addition of rifampicin, which may reflect physiological responses to transcriptional block (1,20,23,26,27).

In summary, besides the complex underlying mechanisms with diverse factors involved, the measurement of mRNA degradation itself could also contribute to the challenges of computational modeling of mRNA degradation in bacteria.

# Computational modeling of mRNA degradation

## Kinetic modeling of mRNA degradation

The simplest type of model for mRNA degradation is the single exponential decay function commonly used to fit mRNA abundance data over time as described above to calculate degradation rates, mRNA half-lives, and/or mRNA lifetimes (reviewed in (28)). These models are derived directly from experimental data and their functions are typically descriptive. In their simplest forms they assume that degradation rate is constant over time after the addition of rifampicin. They can also be modified to incorporate continuing transcription elongation after addition of rifampicin by conversion to piecewise functions (6,27) that capture both the initial delay and exponential decay period. They can also be extended to include slower mRNA degradation that may follow the rapid exponential decay period due to physiological responses to rifampicin (1,27). Recently, more advanced models, such as Bayesian hierarchical modeling,

have been developed to improve the mRNA half-life estimations by modeling the effect of transcription elongation and RNA baseline concentration (23,25).

Assessment of correlations between mRNA half-lives and properties such as G+C content and steady-state abundance can be used to generate hypotheses about factors that affect mRNA degradation rates, and comparisons of half-lives between conditions and strains reveal the impacts of growth conditions and roles of specific RNases (1,3-13,27,29,30). Rates of transcription can be inferred from measured rates of mRNA degradation and steady-state abundance (1,4).

Given the essential role of RNase E in the degradation process as we mentioned early, another study extended the basic exponential model is by adding the concentration of RNase E (28), which allowed the model to incorporate the interactions between RNAs and free RNase E in the cell. Through simulations, they found that the competitions between mRNAs with the limited amount of RNase E could provide an alternative explanation for the widely observed delay in degradation following addition of rifampicin. Additionally, their model indicated that competition could explain the negative correlation between mRNA half-life and abundance, thus providing a potential explanation for this widely observed but never explained phenomenon.

Overall, this type of model provides a broad description of mRNA degradation, highlights the possible contributions of multiple cellular processes to regulation of mRNA degradation, and in some cases suggest mechanisms to explain the observations.

## Modeling the impacts of transcript features on mRNA degradation

For any given organism and growth condition, mRNA degradation rates vary among genes. Many studies therefore seek to elucidate the features of mRNAs that dictate their degradation rates.

Some recent studies have sought to move beyond correlating individual features with half-life as described above, and instead to investigate multiple intrinsic transcript features simultaneously, to identify their impacts on mRNA degradation.

The first challenge of these models is to identify the candidate features that may be important. Transcript features that have shown correlations with mRNA half-life in various organisms including growth rate in *L. lactis* and *E. coli* (4,12), transcript abundance in *E. coli* and *L. lactis* (5,8), GC content in *B. cereus*, *E. coli* and *S. cerevisiae* (5,29,31), 3′ UTR and 5′ UTR sequence motifs in *S. cerevisiae* (32), gene function and essentiality in *B. cereus* and *E. coli* (29,30), transcript length in *L. lactis*, *E. coli*, and *S. cerevisiae* (12,30,31), ribosome density in *S. cerevisiae* (31), and adjacent codon pair usage in *S. cerevisiae* (33). Some studies have reported experiments to show that features correlated with half-life indeed impact it in a causal fashion; for example, two studies showed that manipulating transcription rates affected half-lives of the resulting mRNAs (8,13), providing support for the hypothesis that steady-state abundance affects degradation, although another study did not replicate this result (20).

Another challenge is to employ the appropriate models to investigate the underlying collective effect of these transcript features on stability. Some studies have used linear regression models to quantify feature contributions to variance in mRNA half-lives (5,31,32). These models can quantitatively compare the impacts of features on mRNA half-life, as well as evaluate the contribution of features that have not been studied by adding them to the existing model. However, these models simplify the relationship between the features by assuming that they can be combined linearly to determine transcript half-life. Our recent study using a machine learning approach confirmed the non-linear relationships between features associated with mRNA half-life

in *Mycolicibacterium smegmatis*. We also found that the variance in mRNA stability can be best explained by the collective effect of diverse transcript features. Additionally, being built upon mRNA half-life measurements in log phase growth and hypoxia conditions, these machine learning models were able to identify the transcript features differentially associated with mRNA degradation in each condition. Thereby, they can provide insights about the effect of transcript features on mRNA degradation as well as on stress response mechanisms.

More advanced sequence-based deep learning models were also applied to predict mRNA stability in mammalian systems with the goal of achieving accurate predictions (34,35). These methods have greater predictive power, but at the expense of reduced ability to shed light on the underlying mechanisms. Their performances also rely on large amounts of data for training.

In summary, compared to kinetic modeling, these feature-based prediction models utilize, and in many cases shed light on, the associations between transcript features and mRNA degradation. However, they also depend on the accurate measurement of mRNA degradation rate from the kinetic models. Future work is needed for these feature-based models to further elucidate the relationship between transcript features and their working mechanisms to regulate mRNA degradation.

## Our work on modeling mRNA degradation in *Mycolicibacterium smegmatis*

Considering the potential diverse factors involved in the mRNA degradation process as discussed above, the following Chapter 2 and Chapter 3 describe in detail our recent efforts to utilize these

factors to obtain a more comprehensive model of mRNA degradation in *Mycolicibacterium smegmatis*.

In Chapter 2, we investigated the role of essential enzyme RNase E on mRNA degradation in *Mycolicibacterium smegmatis*. The mRNA half-lives of *M. smegmatis* strains with and without *rne* knockdown were measured using the method that combined rifampicin and RNAseq as mentioned above. These results confirmed the broad impact of RNase E on mRNA degradation in *M. smegmatis*. Furthermore, our results also revealed the influences of sequence context and transcript type (leadered vs leaderless) on cleavage by RNase E. All of these provided important information of RNase E cleavage while highlighting the necessity to further incorporate RNase E in modeling mRNA degradation using more comprehensive computational approaches, which is discussed in Chapter 4.

In Chapter 3, we focused on identifying the features that are important for mRNA degradation in *M. smegmatis* among the diverse pool of potential candidates as we discussed above. We further characterized the influence of these important features in the context of transcript type (leadered vs leaderless) and microenvironments (log phase vs hypoxia). These results confirmed the contributions of transcript features to mRNA degradation rate in *M. smegmatis*. Although these important features cannot fully explain the mRNA degradation variance in our machine learning models, our results highlighted that transcript stability in *M. smegmatis* is shaped by the complex interplay between transcript features and microenvironments.

Consistent with the current understanding of mRNA degradation processes in mycobacteria, our results reaffirmed the impacts of two critical aspects, the key cleavage enzyme RNase E and

intrinsic transcript features, on mRNA degradation. The fact that neither of the two aspects itself can provide a complete model of mRNA degradation highlights the need for a more comprehensive model that could harness the information from both of them together. We discussed one of the potential future approaches in more detail in Chapter 4. In summary, our work established the landscape of the mRNA degradation regulatory mechanisms in mycobacteria, and provided a foundation to facilitate further development of more comprehensive and advanced models of these mechanisms.

## Conclusions

Our current knowledge suggests that the underlying mechanisms that specify and regulate transcript degradation in bacteria include the complex interplay among transcript features, microenvironments, and other cellular processes. Here, we have summarized and discussed the main computational models aiming to describe the mRNA degradation process and identify the factors that impact the process. Given the limitations of these models, further improvements of model performance and accuracy will greatly enhance our understanding of the mRNA degradation mechanisms, and provide valuable information to inspire other studies of RNA metabolism in bacteria.

## Reference

1. Zhou, Y., Sun, H., Rapiejko, A.R., Vargas-Blanco, D.A., Martini, M.C., Chase, M.R., Joubran, S.R., Davis, A.B., Dainis, J.P., Kelly, J.M. *et al.* (2023) Mycobacterial RNase E cleaves with a distinct sequence preference and controls the degradation rates of most Mycolicibacterium smegmatis mRNAs. *J Biol Chem*, **299**, 105312.

2. Sousa, S., Marchand, I. and Dreyfus, M. (2001) Autoregulation allows Escherichia coli RNase E to adjust continuously its synthesis to that of its substrates. *Mol Microbiol*, **42**, 867-878.

3. Bernstein, J.A., Khodursky, A.B., Lin, P.H., Lin-Chao, S. and Cohen, S.N. (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci U S A*, **99**, 9697-9702.

4. Esquerre, T., Laguerre, S., Turlan, C., Carpousis, A.J., Girbal, L. and Cocaign-Bousquet, M. (2014) Dual role of transcription and transcript stability in the regulation of gene expression in Escherichia coli cells cultured on glucose at different growth rates. *Nucleic Acids Res*, **42**, 2460-2472.

5. Esquerre, T., Moisan, A., Chiapello, H., Arike, L., Vilu, R., Gaspin, C., Cocaign-Bousquet, M. and Girbal, L. (2015) Genome-wide investigation of mRNA lifetime determinants in Escherichia coli cells cultured at different growth rates. *BMC Genomics*, **16**, 275.

6. Chen, H., Shiroguchi, K., Ge, H. and Xie, X.S. (2015) Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli. *Mol Syst Biol*, **11**, 808.

7. Esquerre, T., Bouvier, M., Turlan, C., Carpousis, A.J., Girbal, L. and Cocaign-Bousquet, M. (2016) The Csr system regulates genome-wide mRNA stability and transcription and thus gene expression in Escherichia coli. *Sci Rep*, **6**, 25057.

8. Nouaille, S., Mondeil, S., Finoux, A.L., Moulis, C., Girbal, L. and Cocaign-Bousquet, M. (2017) The stability of an mRNA is influenced by its concentration: a potential physical mechanism to regulate gene expression. *Nucleic Acids Res*, **45**, 11711-11724.

9. Morin, M., Enjalbert, B., Ropers, D., Girbal, L. and Cocaign-Bousquet, M. (2020) Genomewide Stabilization of mRNA during a "Feast-to-Famine" Growth Transition in Escherichia coli. *mSphere*, **5**.

10. Redon, E., Loubiere, P. and Cocaign-Bousquet, M. (2005) Transcriptome analysis of the progressive adaptation of Lactococcus lactis to carbon starvation. *J Bacteriol*, **187**, 3589-3592.

11. Redon, E., Loubiere, P. and Cocaign-Bousquet, M. (2005) Role of mRNA stability during genome-wide adaptation of Lactococcus lactis to carbon starvation. *J Biol Chem*, **280**, 36380-36385.

12. Dressaire, C., Picard, F., Redon, E., Loubiere, P., Queinnec, I., Girbal, L. and Cocaign-Bousquet, M. (2013) Role of mRNA stability during bacterial adaptation. *PLoS One*, **8**, e59059.

13. Rustad, T.R., Minch, K.J., Brabant, W., Winkler, J.K., Reiss, D.J., Baliga, N.S. and Sherman, D.R. (2013) Global analysis of mRNA stability in Mycobacterium tuberculosis. *Nucleic Acids Res*, **41**, 509-517.

14. Mackie, G.A. (1992) Secondary structure of the mRNA for ribosomal protein S20. Implications for cleavage by ribonuclease E. *J Biol Chem*, **267**, 1054-1061.

15. McDowall, K.J., Lin-Chao, S. and Cohen, S.N. (1994) A+U content rather than a particular nucleotide order determines the specificity of RNase E cleavage. *Journal of Biological Chemistry*, **269**, 10790-10796.

16. Martini, M.C., Zhou, Y., Sun, H. and Shell, S.S. (2019) Defining the Transcriptional and Post-transcriptional Landscapes of Mycobacterium smegmatis in Aerobic Growth and Hypoxia. *Front Microbiol*, **10**, 591.

17. Mackie, G.A. (2013) RNase E: at the interface of bacterial RNA processing and decay. *Nat Rev Microbiol*, **11**, 45-57.

18. Rauhut, R. and Klug, G. (1999) mRNA degradation in bacteria. *FEMS Microbiol Rev*, **23**, 353-370.

19. Laalami, S., Zig, L. and Putzer, H. (2014) Initiation of mRNA decay in bacteria. *Cell Mol Life Sci*, **71**, 1799-1828.

20. Vargas-Blanco, D.A., Zhou, Y., Zamalloa, L.G., Antonelli, T. and Shell, S.S. (2019) mRNA Degradation Rates Are Coupled to Metabolic Status in Mycobacterium smegmatis. *mBio*, **10**.

21. Anderson, K.L., Roberts, C., Disz, T., Vonstein, V., Hwang, K., Overbeek, R., Olson, P.D., Projan, S.J. and Dunman, P.M. (2006) Characterization of the Staphylococcus aureus heat shock, cold shock, stringent, and SOS responses and their effects on log-phase mRNA turnover. *J Bacteriol*, **188**, 6739-6756.

22. Campbell, E.A., Korzheva, N., Mustaev, A., Murakami, K., Nair, S., Goldfarb, A. and Darst, S.A. (2001) Structural mechanism for rifampicin inhibition of bacterial rna polymerase. *Cell*, **104**, 901-912.

23. Wanney, W.C., Youssar, L., Kostova, G. and Georg, J. (2023) Improved RNA stability estimation indicates that transcriptional interference is frequent in diverse bacteria. *Commun Biol*, **6**, 732.

24. Mosteller, R.D. and Yanofsky, C. (1970) Transcription of the tryptophan operon in Escherichia coli: rifampicin as an inhibitor of initiation. *J Mol Biol*, **48**, 525-531.

25. Jenniches, L., Michaux, C., Popella, L., Reichardt, S., Vogel, J., Westermann, A.J. and Barquist, L. (2024) Improved RNA stability estimation through Bayesian modeling reveals most Salmonella transcripts have subminute half-lives. *Proc Natl Acad Sci U S A*, **121**, e2308814121.

26. Nguyen, T.G., Vargas-Blanco, D.A., Roberts, L.A. and Shell, S.S. (2020) The Impact of Leadered and Leaderless Gene Structures on Translation Efficiency, Transcript Stability, and Predicted Transcription Rates in Mycobacterium smegmatis. *J Bacteriol*, **202**.

27. Moffitt, J.R., Pandey, S., Boettiger, A.N., Wang, S. and Zhuang, X. (2016) Spatial organization shapes the turnover of a bacterial transcriptome. *Elife*, **5**.

28. Etienne, T.A., Cocaign-Bousquet, M. and Ropers, D. (2020) Competitive effects in bacterial mRNA decay. *J Theor Biol*, **504**, 110333.

29. Kristoffersen, S.M., Haase, C., Weil, M.R., Passalacqua, K.D., Niazi, F., Hutchison, S.K., Desany, B., Kolsto, A.B., Tourasse, N.J., Read, T.D. *et al.* (2012) Global mRNA decay analysis at single nucleotide resolution reveals segmental and positional degradation patterns in a Gram-positive bacterium. *Genome Biol*, **13**, R30.

30. Esquerre, T., Moisan, A., Chiapello, H., Arike, L., Vilu, R., Gaspin, C., Cocaign-Bousquet, M. and Girbal, L. (2015) Genome-wide investigation of mRNA lifetime determinants in Escherichia coli cells cultured at different growth rates. *BMC Genomics*, **16**, 275.

31. Neymotin, B., Ettorre, V. and Gresham, D. (2016) Multiple Transcript Properties Related to Translation Affect mRNA Degradation Rates in Saccharomyces cerevisiae. *G3 (Bethesda)*, **6**, 3475-3483.

32. Cheng, J., Maier, K.C., Avsec, Z., Rus, P. and Gagneur, J. (2017) Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast. *RNA*, **23**, 1648-1659.

33. Harigaya, Y. and Parker, R. (2017) The link between adjacent codon pairs and mRNA stability. *BMC Genomics*, **18**, 364.

34. Agarwal, V. and Kelley, D.R. (2022) The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biol*, **23**, 245.

35. Yaish, O. and Orenstein, Y. (2022) Computational modeling of mRNA degradation dynamics using deep neural networks. *Bioinformatics*, **38**, 1087-1101.

# Chapter 2 : Mycobacterial RNase E cleaves with a distinct sequence preference and controls the degradation rates of most *Mycolicibacterium smegmatis* mRNAs

This chapter is a reproduction of following published paper without further changes. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/). To cite this paper:

Zhou Y, Sun H, Rapiejko AR, et al. Mycobacterial RNase E cleaves with a distinct sequence preference and controls the degradation rates of most Mycolicibacterium smegmatis mRNAs. *J Biol Chem*. 2023;299(11):105312. doi:10.1016/j.jbc.2023.105312

In this work, we investigated the impact of RNase E on mRNA degradation in mycobacteria. The transcriptome-wide degradation profiles were collected separately at different time points in *Mycolicibacterium smegmatis* strains. Other lab members did the wet lab work including constructing strains, harvesting RNA, performing qPCR, and performing biochemical experiments. I developed a normalization method that utilized targeted qPCR to obtain a more accurate and comparable quantification of transcript abundance between time points. The degradation profile was further used for transcriptome-wide half-life calculation. To characterize the cleavage events of RNase E, I modified a published method for identifying potential cleavage site based on RNAseq libraries, and applied it to strains in *M. smegmatis* and *M. tuberculosis*. This allowed us to identify the consistent enrichment of cytidines around RNase E cleavage sites in mycobacteria. Besides methodology development and data analysis, I also contributed to the manuscript visualization, writing and editing.

# Mycobacterial RNase E cleaves with a distinct sequence preference and controls the degradation rates of most *Mycolicibacterium smegmatis* mRNAs

**Ying Zhou[1]*, Huaming Sun[2]*, Abigail R. Rapiejko[1], Diego A. Vargas-Blanco[1], Maria Carla Martini[1], Michael R. Chase[3], Samantha R. Joubran[1], Alexa B. Davis[1], Joseph P. Dainis[1], Jessica M. Kelly[1], Thomas R. Ioerger[4], Louis A. Roberts[1], Sarah M. Fortune[3], Scarlet S. Shell[1,2]**

[1] Department of Biology and Biotechnology, Worcester Polytechnic Institute, Worcester, Massachusetts, USA.
[2] Program in Bioinformatics and Computational Biology, Worcester Polytechnic Institute, Worcester, Massachusetts, USA.
[3] Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA.
[4] Department of Computer Science & Engineering, Texas A&M University, College Station, Texas, USA.
* These authors contributed equally to this work

This chapter corresponds to a manuscript that was published as:
Zhou Y, Sun H, Rapiejko AR, et al. Mycobacterial RNase E cleaves with a distinct sequence preference and controls the degradation rates of most Mycolicibacterium smegmatis mRNAs. *J Biol Chem*. 2023;299(11):105312. doi:10.1016/j.jbc.2023.105312

**Author Contributions**

Conceptualization: T.G.N., L.A.R., and S.S.S. Methodology: D.A.V.-B. and T.G.N. Data Analysis: D.A.V.-B., T.G.N., and S.S.S. Writing – Original Draft: D.A.V.-B., S.S.S, and T.G.N. Writing – Review and Editing: D.A.V.-B., L.A.R., and S.S.S. Y.Z., H.S., D.A.V.-B., M.R.C., T.R.I., S.M.F., S.S.S. conceptualization; Y.Z., H.S., A.R.R., D.A.V.-B., T.R.I., L.A.R., and S.S.S. methodology; H.S. and M.R.C. software; A.R.R. validation; Y.Z., H.S., M.R.C., J.M.K., T.R.I., and S.S.S. formal analysis; Y.Z., A.R.R., D.A.V.-B., M.C.M., S.R.J., A.B.D., J.P.D., J.M.K., and L.A.R. investigation; H.S. data curation; S.S.S. writing–original draft; Y.Z., H.S., D.A.V.-B., M.C.M., T.R.I., L.A.R., and S.M.F. writing–review & editing; H.S., M.R.C., and S.S.S. visualization; S.M.F. and S.S.S. supervision; S.M.F. and S.S.S. project administration; T.R.I., S.M.F., and S.S.S. funding acquisition.

# Abstract

The mechanisms and regulation of RNA degradation in mycobacteria have been subject to increased interest following the identification of interplay between RNA metabolism and drug resistance. Mycobacteria encode multiple ribonucleases predicted to participate in mRNA degradation and/or processing of stable RNAs. RNase E is an endoribonuclease hypothesized to play a major role in mRNA degradation due to its essentiality in mycobacteria and its role in mRNA degradation in gram-negative bacteria. Here, we defined the impact of RNase E on mRNA degradation rates transcriptome-wide in the non-pathogenic model *Mycolicibacterium smegmatis*. RNase E played a rate-limiting role in degradation of the transcripts encoded by at least 89% of protein-coding genes, with leadered transcripts often being more affected by RNase E repression than leaderless transcripts. There was an apparent global slowing of transcription in response to knockdown of RNase E, suggesting that *M. smegmatis* regulates transcription in responses to changes in mRNA degradation. This compensation was incomplete, as the abundance of most transcripts increased upon RNase E knockdown. We assessed the sequence preferences for cleavage by RNase E transcriptome-wide in *M. smegmatis* and *Mycobacterium tuberculosis*, and found a consistent bias for cleavage in C-rich regions. Purified RNase E had a clear preference for cleavage immediately upstream of cytidines, distinct from the sequence preferences of RNase E in gram-negatives. We furthermore report a high-resolution map of mRNA cleavage sites in *M. tuberculosis*, which occur primarily within the RNase E-preferred sequence context, confirming that RNase E has a broad impact on the *M. tuberculosis* transcriptome.

# Introduction

Mycobacteria are a globally important group of bacteria including the pathogen *Mycobacterium tuberculosis,* which kills over a million people each year (1), as well as numerous environmental bacteria and opportunistic pathogens. Mycobacteria are phylogenetically distant from better-studied models such as *Escherichia coli*, and consequently, numerous aspects of their fundamental biology remain poorly understood. mRNA metabolism is a critical aspect of mycobacterial biology, as regulation of gene expression facilitates adaptation to stressors both during infection and in the environment, and regulation of mRNA degradation permits energy conservation during severe stress. The roles and regulation of mycobacterial mRNA degradation enzymes remain largely undefined; however, recent reports of interplay between RNA metabolism and drug resistance have highlighted the relevance of these pathways (2-6).

The endoribonuclease RNase E is a critical component of the bulk mRNA degradation machinery in gram-negative bacteria. In *E. coli*, RNase E cleaves single-stranded mRNAs in A/U-rich regions and interacts with other RNA degradation proteins to increase the efficiency of mRNA degradation ((7-11) and reviewed in (12)). In contrast, many gram-positive bacteria such as *Bacillus subtilis* and *Staphylococcus aureus* lack RNase E completely and rely on other RNases such as RNase J and RNase Y. Mycobacteria are phylogenetically more closely related to gram-positive bacteria than gram-negatives, despite having cell envelopes that prevent gram staining. However, they encode orthologs of RNase E, and these genes are essential in both *M. tuberculosis* and the non-pathogenic model *Mycolicibacterium smegmatis* (13-15). The essentiality of RNase E suggests it may be a critical component of the bulk mRNA degradation machinery in mycobacteria.

Consistent with this, mycobacterial RNase E was shown to interact with other RNases such as RNase J and PNPase (16). It was also shown to contribute to rRNA maturation (15).

We previously showed that the *M. smegmatis* transcriptome is shaped by endonucleolytic cleavage events that produce mRNA fragments with monophosphorylated 5′ ends (17). RNase E is known to produce cleavage products with monophosphorylated 5′ ends in other organisms. Taken together with the observation that the mycobacterial cleavage sites occurred preferentially in single-stranded regions, and the paucity of other candidate RNases predicted to cleave with those properties, we hypothesized that RNase E was responsible for the majority of the cleavage sites we mapped in *M. smegmatis*. However, the mycobacterial cleavage sites occurred primarily in a sequence context distinct from that reported to be cleaved by *E. coli* RNase E. Most mycobacterial mRNA cleavages occurred immediately upstream of a cytidine, with a preference for 1-2 purines immediately upstream and uridine three nt downstream of the cleavage site (RR↓**C**NU). A previous report tested the cleavage specificity of *M. tuberculosis* RNase E on several short substrates *in vitro*; however, none of the substrates used in that study contained the motif "RRCNU" (18).

Given the clear importance of RNase E in mycobacteria and lack of information on its role, we sought to define its function in mycobacterial mRNA metabolism. First, we used an inducible system to interrogate the effects of knockdown of *rne*, the gene encoding RNase E, in *M. smegmatis*. We found that RNase E has a rate-limiting role in degradation of most mRNAs, with a larger influence on leadered transcripts compared to leaderless transcripts. Its cleavage signature is ubiquitous across the transcriptomes of both *M. smegmatis* and *M. tuberculosis* and is distinct from that reported in gram-negative bacteria. We then used purified RNase E to confirm its

cleavage specificity *in vitro*. Finally, we report a transcriptome-wide high-resolution map of major RNA cleavage sites in *M. tuberculosis,* which occur in sequence contexts corresponding to the RNase E signature. Together, our results implicate RNase E as the predominant source of 5′ monophosphorylated, cleaved mRNAs in the transcriptomes of both *M. smegmatis* and *M. tuberculosis* as well as a critical mediator of bulk mRNA degradation in these organisms.

# Results

## RNase E has a global role in *M. smegmatis* mRNA degradation

Given its essentiality in mycobacteria and its broad role in mRNA degradation, we sought to determine the role of RNase E in mRNA degradation transcriptome-wide in a mycobacterial model. We therefore constructed an *M. smegmatis* strain in which we could repress transcription of *rne* (msmeg_4626), the gene encoding RNase E. Replacement of the native *rne* promoter and 5′ UTR (17) with the P766(8G) promoter and associated 5′ UTR (19) produced a strain in which anhydrotetracycline (ATc) caused a constitutively expressed reverse TetR to bind the promoter and repress *rne* transcription (Figure 2-1A-B, Table 2-1). We hereafter refer to this as the repressible *rne* strain. Consistent with the known essentiality of *rne*, growth slowed approximately 15 hours after addition of ATc and later ceased (Figure 2-1C). As RNase E is untagged in our strains, we were unable to quantify depletion at the protein level. Notably, the amount of essential protein depletion required to affect growth in *M. tuberculosis* was shown to vary dramatically among essential proteins (20). Construction of the repressible strain involved insertion of a hygromycin resistance gene upstream of *rne*. We therefore constructed an isogenic strain in which the

hygromycin resistance gene was inserted upstream of the native copy of *rne*, hereafter referred to as the control strain (Figure 2-1A).

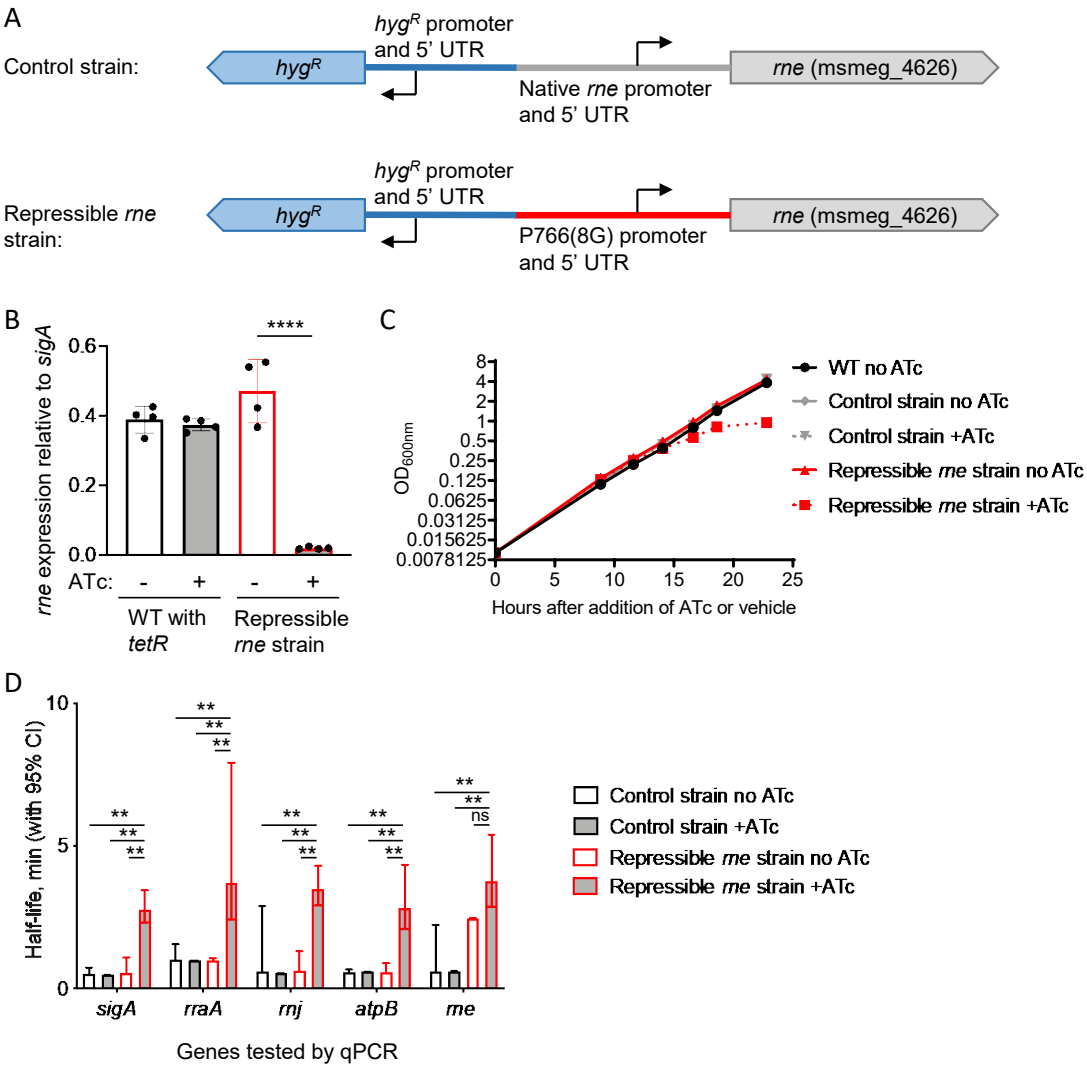



**Figure 2-1. Knockdown of rne expression causes growth cessation and altered transcript abundance in *M. smegmatis*.**

**A.** Promoter replacement strategy to construct a strain in which *rne* expression is repressed by addition of ATc. **B.** *rne* transcript levels were reduced in the repressible *rne* strain following 3 hrs of exposure to ATc. **** $P < 0.001$, two-tailed t test. **C.** Growth of the repressible *rne* strain slowed approximately 15 hours after addition of ATc. **D.** Eight hours after addition of ATc or vehicle, rifampicin was added to block new transcription and mRNA levels of the indicated genes were measured at several time-points by qPCR to determine their half-lives. ** $P < 0.01$, pair-wise comparisons by linear regression.

While the essentiality of *rne* could be due to its role in mRNA degradation, rRNA maturation, or both, we were specifically interested in determining the role of RNase E in mRNA metabolism. We therefore evaluated the impact of *rne* knockdown on mRNA degradation rates prior to the slowing of bacterial growth. We measured the half-lives of several mRNAs by adding rifampicin (RIF) to block transcription initiation and quantifying transcript abundance at timepoints thereafter by quantitative PCR (qPCR)The half-life of the repressible *rne* transcript itself was longer than that of the native *rne* transcript even in the absence of ATc, but this appeared to be a feature of the transcript rather than a generalized phenomenon, as the half-lives of the transcripts of other tested genes were unaffected (Figure 2-1D). In contrast, the half-lives of all tested transcripts were lengthened upon *rne* knockdown (Figure 2-1D). To determine the generalizability of this observation, we used RNAseq to measure mRNA half-lives transcriptome-wide. RNAseq libraries were constructed from RNA extracted from triplicate cultures of each strain and condition at various timepoints after the addition of RIF. qPCR was used to establish relative abundance values for a set of calibrator genes, and these were used to normalize the coverage values obtained from the RNAseq libraries as described in detail in the methods section. Libraries were made from the repressible *rne* strain following 8 hours of treatment with ATc (*rne* knockdown condition), the repressible *rne* strain in the absence of ATc, and the control strain harboring the native *rne* promoter in the presence and absence of ATc. The timepoint for analysis of the *rne* knockdown condition was carefully chosen to maximize our power to detect relevant phenotypes, but prior to the slowing of growth. We expected growth changes would themselves affect mRNA stability as has been reported by us and many others (21-28).

To identify transcripts that were direct targets of RNase E, we calculated half-lives for transcripts of each gene in each condition as described in the methods section and Figure S2-1-S2-3 (Table S2-1). It is important to note that the RNAseq libraries presumably contained mixtures of full-length mRNA and degradation products, as the RNA extraction and library constructions protocols were expected to quantitatively capture most RNAs ~>=150 nt in length. Half-lives were calculated from the summed coverage of reads across each coding sequence at various timepoints following addition of RIF, and due to the relatively short reads produced by Illumina sequencing, it was not possible to distinguish reads arising from full-length transcripts vs degradation products. This caveat is inherent to most published transcriptome-wide studies of mRNA half-life in bacteria. We determined high-confidence half-lives for transcripts of 1643 genes and medium-confidence half-lives for transcripts of an additional 3565 genes in the *rne* knockdown condition. We were able to calculate high-confidence half-lives for 4,068 of these transcripts in the repressible *rne* strain in the absence of ATc as well. Half-lives were similar in comparisons between control conditions, indicating that mRNA degradation rates were not substantially affected by the presence of ATc or by replacement of the native *rne* promoter and 5′ UTR with the tet-repressible promoter (Figure S2-4). In contrast, the half-lives of most transcripts were longer in the *rne* knockdown (Figure 2-2A-B and Table S2-2). The half-lives of the transcripts of 3,622 genes increased by 2-fold or more, and the transcripts of an additional 78 genes had no measurable degradation in the *rne* knockdown. Together, these data are consistent with RNase E playing a rate-limiting step in the degradation of at least 89% of the transcriptome.

While the transcripts of most genes had longer half-lives in the *rne* knockdown condition, the magnitude of the increase in half-life varied substantially among genes (Figure 2-2B). To

investigate the factors that influence transcript sensitivity to RNase E, we examined fold-change half-life in the *rne* knockdown as a function of other potentially relevant characteristics. There was a very weak correlation between mRNA abundance in the control condition and fold-change in half-life upon *rne* knockdown (Figure S2-5). Previous work has reported conflicting observations about the relationship between mRNA abundance and degradation rates in bacteria. Some studies, including one on *M. tuberculosis* and several on *E. coli*, reported inverse relationships between steady-state mRNA abundance and half-lives, such that more abundant transcripts tended to be degraded more quickly (22,25,27-31). Other studies of *E. coli* and *B. subtilis* reported that mRNA abundance and half-life were uncorrelated or weakly positively correlated (24,32,33). We found a weak but statistically significant negative correlation between mRNA abundance and half-life when *rne* was expressed at normal levels, and this correlation disappeared upon *rne* knockdown (Figure S2-6).

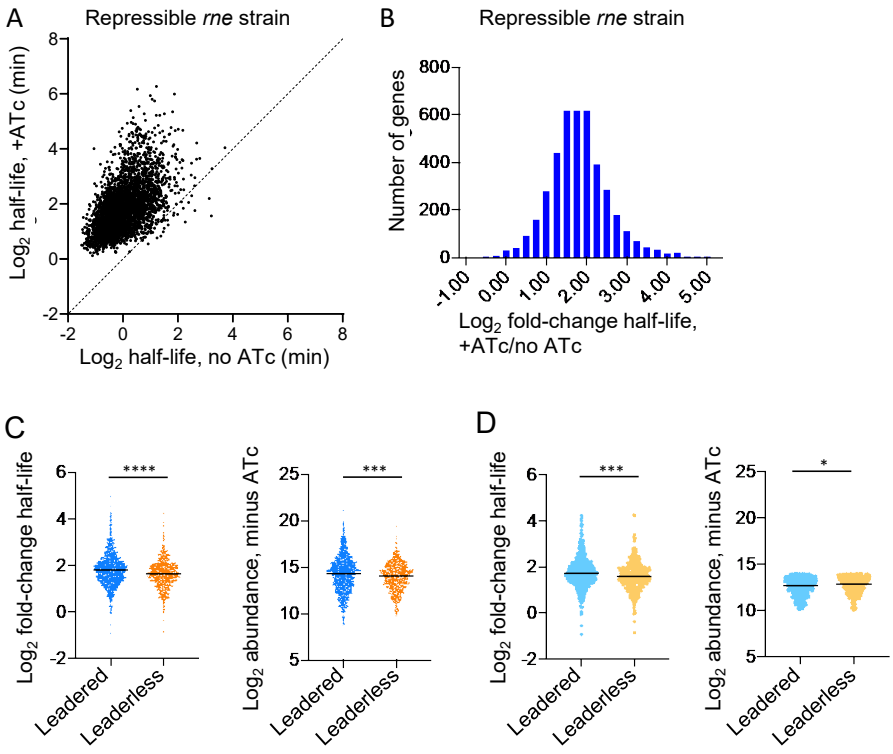


**Figure 2-2. Knockdown of *rne* expression causes stabilization of most of the *M. smegmatis* transcriptome, with leadered transcripts tending to be stabilized more than leaderless transcripts.**

Eight hours after addition of ATc (or vehicle) to knock down (or not) *rne*, rifampicin was added to block new transcription and mRNA levels were measured transcriptome-wide at several time-points by RNAseq to determine half-lives. **A.** Dots represent transcripts with measurable half-lives in both conditions. **B.** The distribution of fold-change in half-life for the transcripts shown in A. **C.** The median fold-change in half-life upon *rne* knockdown was higher for leadered transcripts than for leaderless transcripts (left). The median abundance of leadered transcripts was higher prior to *rne* knockdown (right). **D.** Only transcripts with 10< $\log_2$ abundance <14 were considered, which reversed the difference in abundance trend between leadered and leaderless transcripts. The median fold-change in half-life upon *rne* knockdown was still higher for leadered transcripts than for leaderless transcripts.

In both *M. tuberculosis* and *M. smegmatis*, approximately 15% of genes are transcribed in a leaderless fashion, meaning that transcription and translation start at the same position and there is no 5′ UTR (17,34,35). Other genes are transcribed as leadered genes with 5′ UTRs, or in polycistronic transcripts. Leader status affects translation efficiency in different conditions and in some cases alters mRNA stability (36-38). On average, leaderless genes were less affected by *rne* knockdown than leadered genes (Figure 2-2C, left). Leaderless genes also had lower median abundance than leadered genes in the control condition (Figure 2-2C, right). We then considered only genes where 10<log2 abundance<14 (Figure 2-2D, right). Within this group, the median abundance of leaderless transcripts was slightly higher than that of leadered transcripts. Nonetheless, the leadered transcripts within this group still had a greater median increase in half-life upon *rne* knockdown than leaderless transcripts (Figure 2-2D, left). This suggests that the difference in response of leaderless vs leadered transcripts to *rne* knockdown cannot be explained by differences in steady-state abundance of those transcripts. Leadered transcripts may therefore be generally more sensitive to RNase E than leaderless transcripts. However, both groups included genes that were unaffected by *rne* knockdown as well as genes that were strongly affected, indicating that additional factors are likely larger drivers of RNase E sensitivity. Given that RNase

E is strongly stimulated by engagement of transcript 5′ ends in *E. coli* ((39,40) and others), we considered that accessible 5′ ends might make transcripts more sensitive to RNase E. However, we did not find correlations between fold-change in half-life upon *rne* knockdown and predicted secondary structure near the 5′ ends of transcripts (Figure S2-7).

## Knockdown of *rne* affects mRNA abundance through both direct and indirect mechanisms in *M. smegmatis*

To assess the impact of *rne* knockdown on mRNA abundance, we examined transcript abundance in the *rne* knockdown strain with and without ATc prior to transcriptional blockage with RIF. These were the same samples used for the 0 minutes RIF treatment condition for mRNA half-life calculations, harvested 8 hours after addition of ATc or vehicle control. Our normalization method allowed us to measure mRNA abundance relative to total RNA abundance, in arbitrary units. As total RNA yields were similar for all strains and conditions, this roughly approximates mRNA abundance per cell, measured in arbitrary units. A large majority of genes had increased abundance upon *rne* knockdown (Table S2-2). We therefore could not statistically assess differential expression using a standard pipeline such as DESeq2, for which the identification of differential expressed gene relies on the assumption that mean gene expression is similar in the conditions being compared. Instead, we compared transcript abundance using Clipper, which does not rely on the specific data distributions of the two conditions (41). Of 6,922 total genes with mean read counts >0 in both conditions, 2,561 genes had increased abundance upon *rne* knockdown using cutoffs of $q < 0.05$ and fold change >= 2 (Table S2-3). In contrast, only 9 genes that met these criteria had decreased abundance.

There was a significant positive correlation between increase in half-life upon *rne* knockdown and increase in abundance (Spearman r = 0.3565, $p < 0.0001$; Figure 2-3A). These observations are consistent with the idea that slower mRNA degradation leads to accumulation of mRNA in the cell. However, the changes in mRNA abundance were of a smaller magnitude than would be expected if transcription rates remained unchanged (compare the dashed and solid lines in Figure 2-3A). We therefore used the measured mRNA abundance and half-life values to estimate transcription rates. A majority of genes had lower estimated transcription rates in the *rne* knockdown condition, suggesting the existence of a feedback process in which transcription is slowed to partially compensate for the longer mRNA half-lives (Figure 2-3B, Table S2-4).

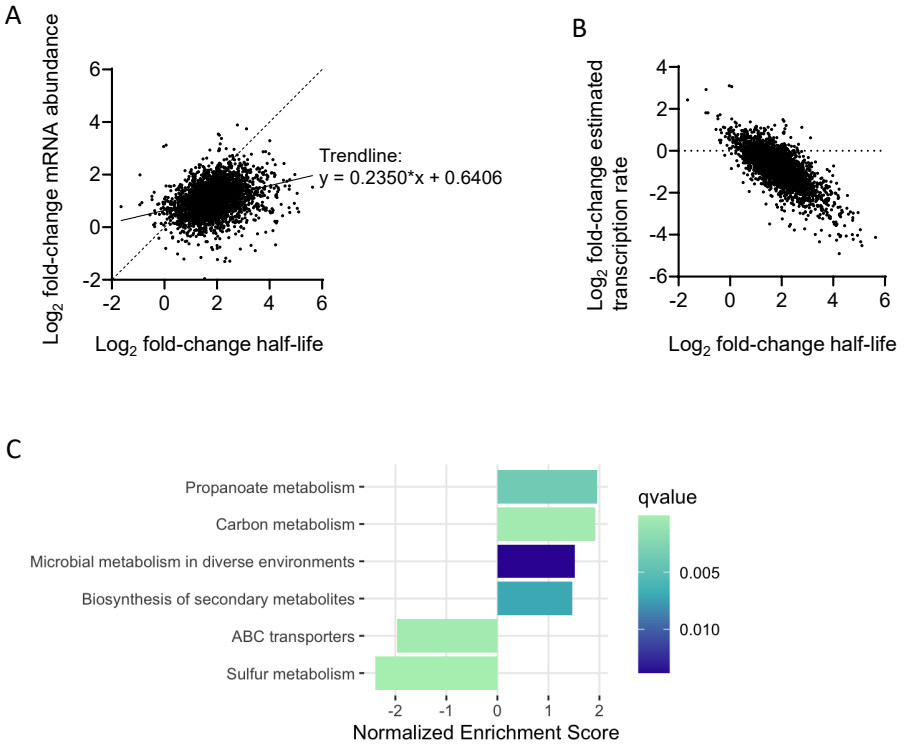



**Figure 2-3. Knockdown of *rne* impacts mRNA abundance both directly and indirectly.**
**A.** Each dot represents a gene for which $\log_2$ fold change in transcript abundance upon *rne* repression is shown as a function of $\log_2$ fold change in half-life. The solid line shows the linear regression fit where y = 0.2350*x + 0.6406. The dashed line shows the expected relationship between $\log_2$ fold change half-life and $\log_2$ fold change abundance if transcription rate were unchanged. **B.** Estimated transcription rates were

calculated from the measured mRNA half-lives and steady-state abundance. The same genes shown in panel A are shown here. **C.** For each gene, the expected change in abundance was calculated as a function of change in half-life according to the equation in panel A. The differences between expected and observed changes in abundance were then calculated, and genes with large differences were considered more likely to be subject to active regulation. Gene Set Enrichment Analysis was performed on the observed/expected $\log_2$ fold change abundance, and the gene categories with statistically significant enrichment or depletion are shown. Genes in the categories with positive enrichment scores had larger than expected increases in transcript abundance, and genes in the categories with negative enrichment scores had lower than expected increases (or had decreases) in transcript abundance. The q value is a p value corrected for multiple comparisons.

The results described above suggested that many of the transcript abundance changes caused by *rne* knockdown were direct consequences of slower degradation that was only partially compensated for by globally reduced transcription. However, some genes did not follow the bulk trend. We hypothesized that the stress imposed by *rne* knockdown led to active transcriptional changes of some specific genes and were therefore indirect effects of *rne* knockdown. To distinguish direct and indirect transcript abundance changes, we fit the bulk relationship between $\log_2$ abundance change and $\log_2$ half-life change by linear regression to determine predicted abundance changes as a function of change in half-life (Table S2-2). The difference between expected and actual abundance change reflects the extent to which a gene deviated from the bulk trend. This approach makes the assumption that most abundance changes are direct. Genes with positive differences between observed and expected abundance change had higher abundance than expected upon *rne* knockdown, while genes with negative differences had lower abundance than expected upon *rne* knockdown. To investigate the nature of the genes that did not follow the bulk trend and therefore appeared to be actively regulated at the point of transcription in response to *rne* knockdown, we used Gene Set Enrichment Analysis (42) to identify gene categories that were overrepresented among genes with large differences between observed and expected abundance. Genes with higher-than-expected abundance were most enriched for

carbon metabolism and propanoate metabolism, while genes with lower-than-expected abundance were enriched for sulfur metabolism and ABC transporters (Figure 2-3C). Transcripts for the genes encoding the RNA helicase RhlE1 (msmeg_1540) and predicted RNA binding protein KhpB (msmeg_6941) had higher-than-expected abundance, suggesting that they are transcriptionally upregulated in response to *rne* knockdown. These two proteins have reported roles as components of mycobacterial RNA degradosomes (16). It is possible that they are upregulated to partially compensate for the decrease in RNase E abundance. However, the genes encoding two other major degradosome constituents, PNPase and RNase J, did not have substantially different abundance than expected, suggesting that their abundance is not regulated in response to RNase E deficiency.

## RNase E cleavage site regions in *M. smegmatis* and *M. tuberculosis* are enriched for cytidines

Given the global role for RNase E implied by our data, we hypothesized that RNase E was the enzyme responsible for many of the mRNA cleavage events that we previously mapped (17). Those cleavage events occurred across the transcriptome at a sequence motif not previously associated with any RNase in any organism. The dominant feature of the cleavage site sequence context was a cytidine immediately downstream of the cleavage site. To assess the impact of *rne* knockdown on mRNA cleavage in *M. smegmatis*, we modified a recently published method for assessment of mRNA cleavage from standard paired-end RNAseq libraries, without construction of separate 5′-targeted libraries (43) (Figure 2-4A and S2-8). This method harnesses the fact that in a standard mRNA expression library, fewer reads are obtained in regions containing cleavage sites compared

to longer stretches of uncleaved RNA. When comparing the reads obtained from strains with and without knockdown of an endoribonuclease, one therefore expects to find regions of genes that have fewer reads when the RNase is expressed at higher levels. To apply this method to our *M. smegmatis rne* knockdown data, we first quantified the number of reads aligning to each

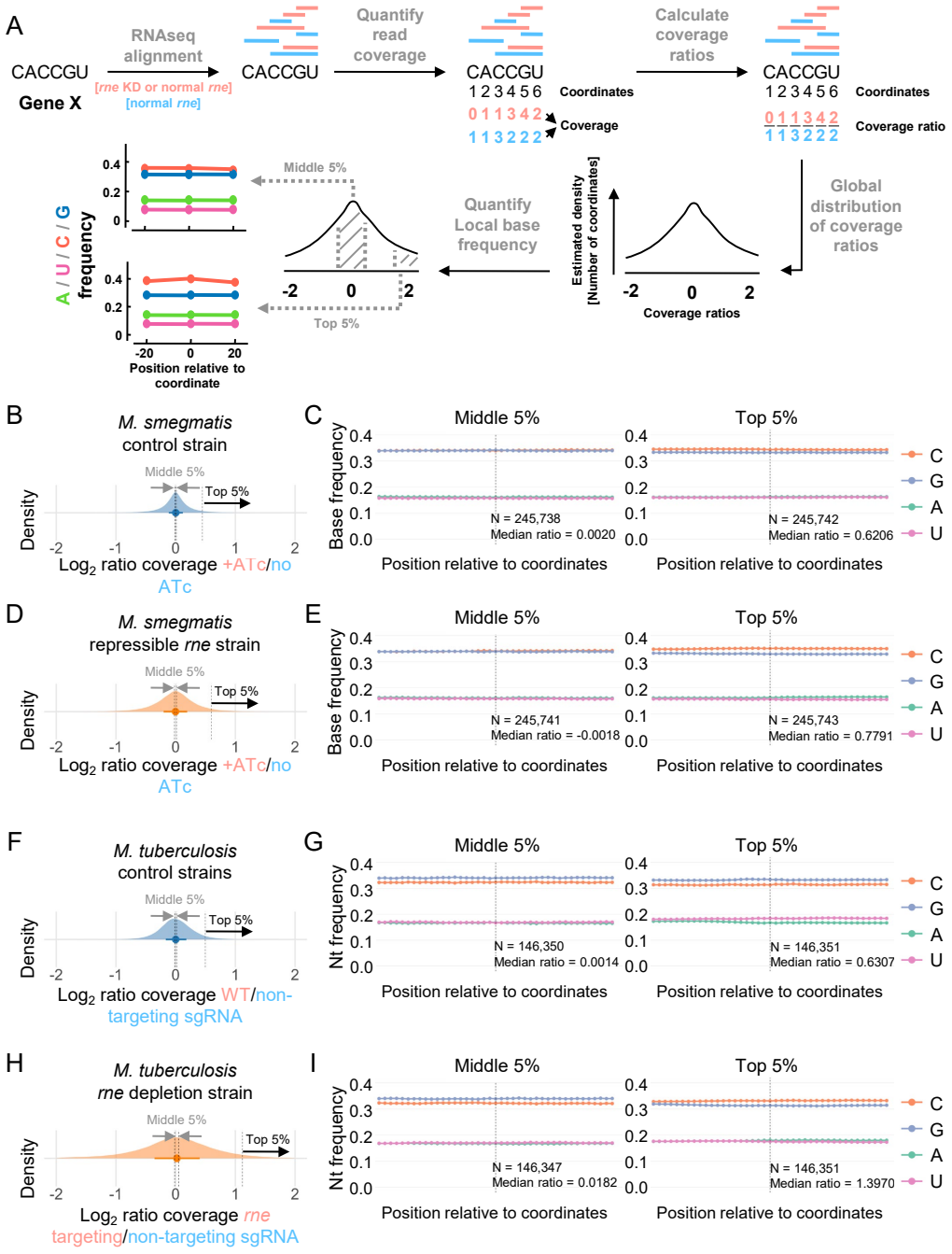


**Figure 2-4. Cytidines are enriched in regions of RNase E-dependent mRNA cleavage in both *M. smegmatis* and *M. tuberculosis*.**

**A.** Overview of the method for relative quantification of mRNA cleavage events using standard RNAseq data. Illumina RNAseq data for both *M. smegmatis* (this work) and *M. tuberculosis* (16) were used. Both datasets included an *rne* knockdown condition and multiple control conditions. Cleavage events result in a lower proportion of reads in the immediate vicinity of the cleavage site compared to uncleaved regions of the transcript. Read depth (coverage) for each coordinate within each coding sequence across the genome was determined for each sample, then normalized by the average read depth within that gene in that sample. For each coordinate, the $\log_2$ ratio of coverage in the *rne* knockdown compared to a control (or two distinct controls compared to each other) was determined. The median $\log_2$ ratio should be approximately zero for all comparisons, due to the method of normalization. Coordinates at or near RNase E cleavage sites are expected to have high ratios in the *rne* knockdown/control comparison. The regions surrounding coordinates with $\log_2$ ratios in the top 5% and middle 5% were then assessed for base composition bias (A, U, C, G frequency). The bases at each position within 20 coordinates up and downstream of coordinates of interest (those having $\log_2$ ratios in the middle 5% or top 5%) were determined. **B.** $\text{Log}_2$ ratios from the *M. smegmatis* control strain in the presence and absence of ATc, which is not expected to affect RNase E activity. **C.** The base frequencies in 41-nt regions centered on coordinates with $\log_2$ ratios in the middle 5% or top 5% of the distribution shown in panel B. **D.** $\text{Log}_2$ ratios from the *M. smegmatis* repressible *rne* strain in the +ATc condition (*rne* repressed) vs the no-ATc condition (*rne* expressed). **E.** The base frequencies in 41-nt regions centered on coordinates with $\log_2$ ratios in the middle 5% or top 5% of the distribution shown in panel D. Coordinates with $\log_2$ ratios in the top 5% are expected to be enriched for RNase E cleavage site-containing regions. **F.** $\text{Log}_2$ ratios from two *M. tuberculosis* strains that are expected to have similar RNase E activity (a WT strain and a strain expressing a CRISPRi system with a non-targeting sgRNA). **G.** The base frequencies in 41-nt regions centered on coordinates with $\log_2$ ratios in the middle 5% or top 5% of the distribution shown in panel F.. **H.** $\text{Log}_2$ ratios from an *M. tuberculosis* strain expressing an sgRNA to knock down expression of *rne* vs a strain with a non-targeting sgRNA. **I.** The base frequencies in 41-nt regions centered on coordinates with $\log_2$ ratios in the middle 5% or top 5% of the distribution shown in panel H. . Coordinates with $\log_2$ ratios in the top 5% are expected to be enriched for RNase E cleavage site-containing regions.

coordinate within each gene in the same RNAseq libraries that were used for expression analyses in the previous section (0 minutes RIF treatment, harvested 8 hours after addition of ATc or vehicle control) (Figure 2-4A). The number of reads aligned to each coordinate is henceforth referred to as that coordinate's coverage (Figure 2-4A). The coverage at each coordinate in each coding sequence in each sample was then normalized to the summed coverage of all coordinates in that coding sequence, to avoid confounding by genes whose mRNA abundance varied among conditions (Figure S2-8). Coding sequences and coordinates with low coverage were filtered out. We then calculated the $\log_2$ ratios of coverage for each coordinate in the repressible *rne* strain in

the presence vs absence of ATc, as well as for the control strain in the presence vs absence of ATc (Figure 2-4A, B, and D). If RNase E was responsible for cleavage at a particular site, we predicted that a smaller proportion of transcripts would exist in the cleaved form in the *rne* knockdown compared to the control conditions. We therefore expected coordinates that were very close to cleavage sites to have higher coverage in the repressible *rne* strain in the presence of ATc compared to the absence of ATc. In contrast, we did not expect coverage near RNase E cleavage sites to be affected by ATc in the control strain.

For each of the two comparisons (presence vs absence of ATc in the repressible *rne* strain and presence vs absence of ATc in the control strain), we obtained distributions of $\log_2$ coverage ratios for each coordinate in each gene that passed our coverage filters (Figure 2-4B and D). The distributions of $\log_2$ coverage ratios in the presence and absence of ATc were centered around 0 for both comparisons, due to our normalization method (Figure 2-4B and D). The distribution was broader for the *rne* knockdown strain, consistent with the expectation that RNase E levels affect the relative abundance of cleaved vs intact transcripts. RNase E both makes cleavage products and degrades cleavage products into pieces too small to be captured by our RNAseq library construction strategy; we therefore expect the effects of *rne* knockdown on the steady-state abundance of detectable cleavage products to be complex, with some cleaved RNAs decreasing in abundance and others increasing in abundance.

Nonetheless, many of the coordinates at or near RNase E cleavage sites should have high $\log_2$ coverage ratios for +ATc/no ATc comparison in the repressible *rne* strain, but this should not be true for the control strain where ATc does not affect RNase E levels. We used this assumption to assess the sequence context of RNase E cleavage sites by examining the sequence contexts of

coordinates with high $\log_2$ ratios in the +ATc/no ATc comparison in the repressible *rne* strain. Specifically, we determined the sequence context of the 5% of coordinates with the highest $\log_2$ ratios (Figure 2-4E). We compared this to the sequence context of coordinates with $\log_2$ ratios in the highest 5% in the +ATc/no ATc comparison in the control strain, as well as to the sequence context of coordinates with $\log_2$ ratios in the middle 5% for both strains (Figure 2-4C and E). For the control strain, the relative frequencies of each base were equivalent for the coordinates with $\log_2$ ratios in the middle 5% and highest 5% (Figure 2-4C), with G and C having similar frequencies that were much higher than A and U, as expected for an organism with a genomic GC content of ~65%. In the *rne* knockdown strain, the same was true for the coordinates with $\log_2$ ratios in the middle 5% (Figure 2-4E). In contrast, coordinates with $\log_2$ ratios in the highest 5% in the repressible *rne* strain showed a clear enrichment for cytidines (Figure 2-4E). This is consistent with the hypothesis suggested by our previous work that RNase E has a preference for cleaving near cytidines (17). The enrichment for cytidines may appear modest compared to our previous finding that >90% of mapped cleavage events were immediately upstream of cytidines (cytidine at the +1 position). However, this modest enrichment is consistent with the nature of the method. At any given endonucleolytic cleavage site, when comparing $\log_2$ coverage in a strain with lower cleavage to a strain with higher cleavage, coordinates at the -1 position are expected to have equally high $\log_2$ ratios as coordinates at the +1 position, but only the +1 position shows a preference for cytidines. Furthermore, nearby coordinates (eg, -2, -3, -4, +2, +3, +4) are also likely to have relatively high $\log_2$ ratios. Cytidines were not enriched at any position besides +1 in our previously mapped cleavage sites, and in several of those positions there was reduced presence of cytidines (17). Adding to the complexity of the interpretation of these data, RNAseq expression

library coverage is typically bumpy, with stochastic factors leading to variability in coverage among adjacent coordinates. The coordinates with $\log_2$ ratios in the highest 5% are therefore likely to include many -1 and +1 positions of cleavage sites, but also many other coordinates that are in the general vicinities of cleavage sites. The observed modest enrichment of cytidines in Figure 2-4E is broadly consistent with the averaging of the previously observed sequence preferences in the vicinity of cleavage sites.

RNAseq data have been previously published for *M. tuberculosis* with *rne* knockdown (16). We therefore applied the method described above to investigate the extent to which *M. tuberculosis* RNase E preferentially cleaves cytidine-rich regions. This was done by comparing RNAseq read coverage from a strain in which *rne* was knocked down by CRISPRi to coverage from a strain expressing a non-targeting CRISPRi sgRNA (Figure 2-4H). As a control, we compared RNAseq read coverage from WT H37Rv to coverage from the strain expressing a non-targeting CRISPRi sgRNA (Figure 2-4F). In the control strain comparison, the coordinates with $\log_2$ ratios in the middle 5% and top 5% $\log_2$ ratios had similar sequence contexts, which differed from the *M. smegmatis* data in having a greater proportion of guanosines than cytosines (Figure 2-4G). This is consistent with differences in the overall nucleotide usage in the two organisms; *M. tuberculosis* coding sequences contain more guanosines than cytidines, while *M. smegmatis* coding sequences have roughly equal usage of guanosines and cytidines (Figure S2-9). When comparing the strain containing an *rne*-targeting CRISPRi sgRNA to the non-targeting sgRNA strain, we found that coordinates with $\log_2$ ratios in the middle 5% had base frequencies similar to the control comparison, but the coordinates with $\log_2$ ratios the highest 5% showed a higher frequency of cytidines compared to

guanosines (Figure 2-4I). The preference for RNase E to cleave cytidine-rich regions is therefore conserved in *M. tuberculosis*.

### *M. smegmatis* RNase E cleaves immediately 5′ of cytidines

To assess RNase E's cleavage site sequence preference with higher resolution, we performed two additional analyses. First, we used 5′ RACE to qualitatively compare the abundance of 5′ ends arising from a putative RNase E cleavage event in the rRNA precursor (Figure S2-10A). We mapped a 5′ end in the spacer region between the 16S and 23S rRNAs resulting from cleavage at the sequence UG↓CU (Figure S2-10A). Consistent with the idea that RNase E is responsible for cleaving this site, the band corresponding to the 5′ end produced by the cleavage event was fainter in the *rne* knockdown (Figure S2-10B and C). This is consistent with a previously reported role for RNase E in cleaving near this location (15), although the method used in that report did not permit precise identification of the 5′ end as we did here.

Next, we overexpressed and purified *M. smegmatis* RNase E in *E. coli* to test its cleavage specificity in vitro. This recombinant RNase E lacked part of the predicted N-terminal scaffold domain (deletion of residues 2-145) and most of the predicted C-terminal scaffold domain (deletion of residues 825-1037), similar to RNase E variants used for in vitro work in many reports (including (18,40,44)). Our RNase E also had N-terminal 6x-His and FLAG epitope tags to facilitate purification. A variant containing the predicted catalytic site mutations D694R and D738R was purified to use as a catalytically dead control (40). The purified proteins were incubated with an in vitro-transcribed RNA substrate that contained a 106 bp duplex region and a 120 nt single-stranded region (Figure 2-5B and S2-11). Some RNA cleavage was observed in reactions with the presumed

catalytically dead RNase E, suggesting that our preps contained small amounts of an *E. coli* RNase (Figure 2-5A). We therefore focused only on bands that appeared exclusively in reactions with catalytically active RNase E. Several of these bands were subject to 5′ and 3′ RACE to map the cleavage site locations. We mapped four distinct cleavage sites, all in the single-stranded portion of the substrate (Figure 2-5A and B). Two were at positions where we previously mapped cleavage sites in vivo (17), and all four occurred at the sequence motif RN↓CNU. These data confirm the propensity of RNase E to cleave single-stranded RNAs at phosphodiester bonds 5′ of cytidines.

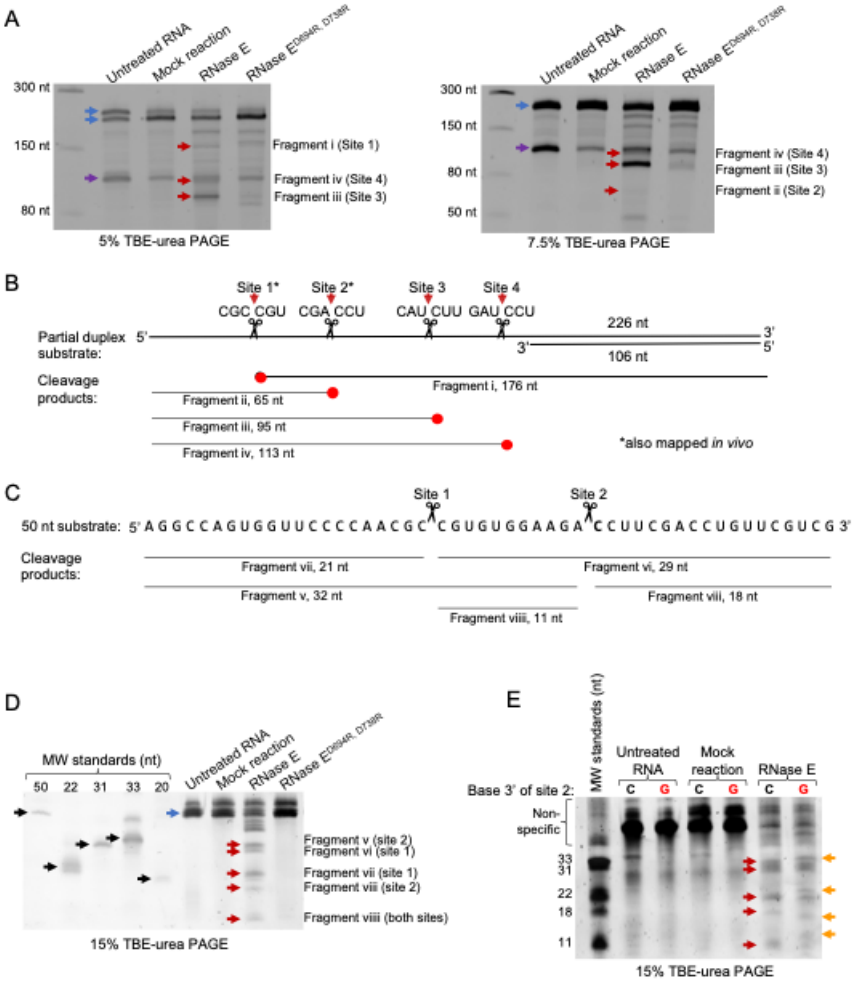



**Figure 2-5. RNase E cleaves 5′ of cytidines *in vitro*.**

**A.** SYBR-gold-stained TBE-UREA gels revealing cleavage of the RNA substrate shown in B (300 ng) upon incubation for 1 hr with 80 ng purified, recombinant *M. smegmatis* RNase E catalytic domain (residues 146-824, with an N-terminal FLAG-his tag). The D694R, D738R mutant is predicted to be catalytically dead.

Untreated RNA was not incubated with reaction buffer, while mock reactions contained RNA and buffer in the absence of enzyme. Cleavage sites are numbered 1-4 and resulting fragments visible on the gel are designated with numerals i-iv, as shown schematically in B. Red arrows denote cleavage fragments. Blue arrows indicate the longer strand of the partial-duplex substrate or the annealed substrate, and purple arrows indicate the shorter strand of the partial-duplex substrate. Note that while all samples were heated with formamide prior to loading the gels, the partial-duplex substrate did not fully denature following incubation with reaction buffer in the mock reaction or enzyme-containing reactions. **B.** Schematic (not to scale) of the partial duplex RNA substrate used in panel A. Red arrows indicate RNase E cleavage sites mapped by 5′ or 3′ RACE on cleavage products extracted from the gels shown in panel A. The thin lines below indicate the sizes of the extracted cleavage products (not to scale), with red dots indicating the ends that were mapped by RACE. **C.** A 50 nt region of the single-stranded portion of the RNA substrate shown in panel A was synthesized. The expected products from cleavage at sites 1 and/or 2 are shown below. The bolded "C" was mutated to G in panel E below. **D.** RNase E cleavage reactions using the substrate shown in panel C and enzyme that was re-purified using a more stringent wash protocol to remove contaminating *E. coli* RNases. Reactions contained 80 ng RNase E and 150 ng of RNA and were incubated for 2 hours. The expected cleavage products shown in panel C are indicated with red arrows. Black arrows indicate the positions of molecular weight standards. The blue arrow indicates the full-length substrate. **E.** Cleavage reactions were done as in panel D with the addition of a substrate with a C to G mutation at the position 3′ of cleavage site 2 (indicated with a red "G"). The indicated molecular weight standards were combined in the first lane. Bands labeled "non-specific" are unidentified byproducts of the MW standard synthesis reactions. Red arrows denote the expected cleavage products observed in panel D. Orange arrows indicate bands that appeared or shifted in position when the substrate had the C to G mutation at cleavage site 2. All gels are representative of at least three independent experiments.

To test the importance of having cytidines at the 3′ sides of RNase E cleavage sites, we synthesized a shorter substrate that represented 50 nt of the single-stranded region of the partial duplex used above, containing cleavage sites 1 and 2 (Figure 2-5C). We re-purified catalytically active and dead versions of RNase E with the addition of a 1 M NaCl wash step, and found the catalytically dead version had no detectable activity on the 50 nt substrate (Figure 2-5D). In contrast, the catalytically active RNase E cleaved the substrate into bands consistent with the sizes expected from partial cleavage at both sites (Figure 2-5C and D). We then synthesized a version of the 50 nt substrate in which the cytidine on the 3′ side of cleavage site 2 was mutated to guanosine. The two bands presumed to arise from cleavage at site 2 shifted in size (Figure 2-5E, fragments v and viii), and two new bands appeared. These results are consistent with the idea that the cleavage site cytidine is important for recognition and/or cleavage by RNase E.

## The *M. tuberculosis* transcriptome is shaped by mRNA cleavage immediately upstream of cytidines

We previously mapped *M. smegmatis* RNA cleavage sites in vivo by differential ligation (17). These data are complementary to the cleavage site analysis described above; they do not give information on the RNase responsible, but they give single-nt resolution. To determine the extent to which cleavage patterns were similar in the pathogen *M. tuberculosis*, we applied the differential ligation approach. This well-validated method distinguishes between mRNA cleavage sites and primary 5′ ends produced from transcription initiation (TSSs) based on their different chemical properties (35,45). We identified 2,983 cleavage sites with high confidence (Table S2-5A), using a filter that required the cleaved 5′ ends to pass an abundance threshold relative to nearby expression library coverage as we did previously for *M. smegmatis*. The TSSs mapped with this approach have been reported elsewhere (35). However, the relationships between the TSSs and genes, as well as operon predictions based on TSS locations, were not previously published and are therefore reported here in Table S2-5B-H.

RNA cleavage in *M. tuberculosis* occurred at a sequence motif very similar to that observed in *M. smegmatis*, with a strong bias for cleavage 5′ of cytidines (88% of high-confidence cleavage sites) and a weak bias for cleavage 3′ of purines (Figure 2-6A and (17)). Given the multiple lines of evidence shown above indicating that RNase E cleaves in this sequence context, we hypothesize that RNase E was responsible for most of the mapped *M. tuberculosis* cleavage sites. Analysis of the predicted secondary structure in the vicinity of cleavage sites revealed that cleavage occurred in regions more likely to be single-stranded (Figure 2-6B), consistent with expectations for RNase E (reviewed in (12)). We then removed one of the abundance filters used in the 5′ end data analysis

pipeline to capture a greater number of putative cleavage sites (Table S2-5I). Analysis of the sequence context of this expanded cleavage site list revealed a similar preference for cleavage immediately upstream of cytidines (85% of the 5′ ends in the dataset), with a similar but weaker signal for sequence preferences at other positions surrounding the cleavage site (Figure S2-12).

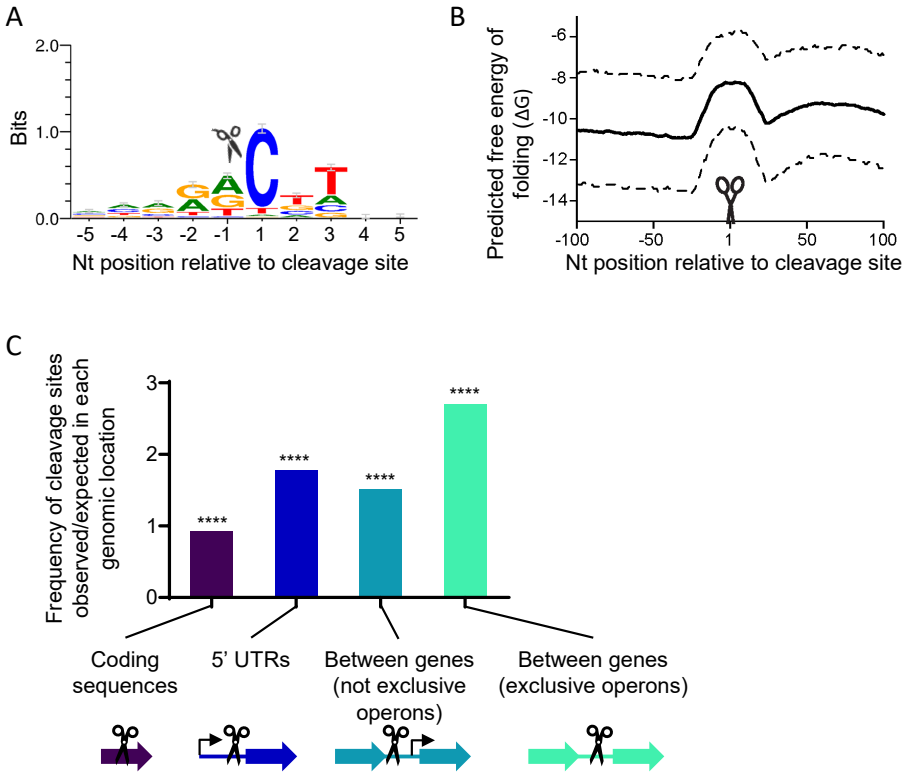



**Figure 2-6. A transcriptome-wide mRNA cleavage site map in *M. tuberculosis* reveals sequence and secondary structure preferences consistent with RNase E, and greater cleavage site frequency in 5′ UTRs and intergenic regions.**

**A.** Weblogo (3.7.4) generated from the complete set of mapped *M. tuberculosis* cleavage sites aligned by cleavage site position. Cleavage occurs between positions -1 and 1 as indicated by the scissor icon. **B.** RNA cleavage typically occurs within regions of lower secondary structure. The minimum free energy secondary structure was predicted for sliding 39 nt windows across 200 nt of sequence spanning each RNA cleavage site. For each coordinate, the mean (solid line; interquartile range, dashed lines) predicted free energy (ΔG) of secondary structure formation of all 2,983 cleaved RNAs was determined. **C.** The frequencies of RNA cleavage sites in various genomic regions were determined: coding sequences, 5′ UTRs, and between adjacent genes on the same strand. Regions between genes on the same strand were separated according to whether or not the gene pair was predicted to be transcribed in an exclusively operonic (polycistronic) fashion. Gene pairs were considered to be transcribed exclusively in operons if only the first gene had a mapped transcription start site (TSS). Gene pairs were considered to be not transcribed exclusively in operons if each gene had its own TSS. In the latter case, genes may be transcribed as a mixture of

monocistronic and polycistronic transcripts. 5′ UTRs were included only if the next upstream gene was on the opposite strand. The observed frequencies of cleavage sites in each region were compared to the frequencies that would be expected if cleavage sites were distributed among these regions without bias. ****, $p < 0.0001$ by binomial test comparing the observed vs expected frequencies.

MazF was reported to also cleave near cytidines (46), but it produces 5′ hydroxyls rather than 5′ monophosphates and its cleavage products are therefore not captured by our methodology. However, RNase J is predicted to cleave single-stranded RNAs and produce 5′ monophosphates. To determine if RNase J contributed to the mapped cleavage sites in *M. tuberculosis*, we compared the abundance of cleavage-site-derived 5′ ends in a WT strain and an RNase J deletion strain (Figure S2-13) (3). Most of the cleaved 5′ ends had similar abundance in the two strains, consistent with the hypothesis that most of them are not produced by RNase J.

## mRNA cleavage sites are disproportionately located in 5′ UTRs and intergenic regions in *M. tuberculosis*

To further investigate the contributors to RNA cleavage site selection in *M. tuberculosis*, we examined the frequencies of cleavage sites in coding sequences, 5′ UTRs, and between adjacent genes encoded on the same strand. In each case we assessed enrichment or depletion by comparing the observed number of cleavage sites to the number expected if cleavage was equally likely to occur in those various locations. Cleavage sites were present at less than the expected frequency in coding sequences, and at greater than the expected frequency in 5′ UTRs and intergenic regions (Figure 2-6C). This pattern is similar to what we previously observed in *M. smegmatis* (17). It could be the result of differential occurrence of cleavage in these locations, or could be the result of cleavage in non-coding sequences being more likely to result in products stable enough to be detected. Cleavage events that trigger very rapid degradation would be

unlikely to be detected by our methodology. Interestingly, the greatest enrichment for mapped cleavage sites occurred between genes that were transcribed exclusively as polycistrons (Figure 2-6C). In some cases there was differential abundance of the transcripts corresponding to genes upstream and downstream of the cleavage site (Table S2-6), suggesting that cleavage could lead to differential stability of segments of transcripts as has been reported in some other bacterial operons (47-53). Consistent with this idea, we found that two pairs of polycistronic *M. smegmatis* genes with intervening cleavage sites had differential stabilities upstream and downstream of the cleavage sites (Figure S2-14).

## Discussion

Here we used a combination of approaches to define the role of RNase E in mycobacterial mRNA degradation and identify its targets. The dramatic effect of *rne* knockdown on mRNA degradation rates in *M. smegmatis* is consistent with the essentiality of this enzyme in mycobacteria; it appears to play a rate-limiting step in degradation of the transcripts of almost 90% of genes. There was variability in the extent to which transcripts were stabilized upon *rne* knockdown, suggesting that while RNase E likely contributes to degradation of most mRNAs, other RNases may contribute differentially across the transcriptome. For example, the essential exoribonuclease PNPase could conceivably be the major degradation factor for those genes that were minimally affected by *rne* knockdown. An alternative explanation is that some mRNAs may be exquisitely susceptible to degradation, such that they were still efficiently degraded by the small amounts of RNase E present in the knockdown condition.

Most of our experiments were done eight hours after inducing repression of *rne* transcription, which was several hours prior to slowing of growth. While this strategy allowed us to distinguish the effects of RNase E knockdown from the effects of slowed growth to due loss of an essential function, we cannot distinguish with certainty which effects are direct and which are indirect. We have therefore made some assumptions in our analyses that should be noted. We assumed that the slowing of mRNA degradation was a direct consequence of reduced RNase E levels, because our RNAseq data did not suggest that any other RNases have reduced expression. We also assumed that the impacts of *rne* knockdown on mRNA abundance were due to a combination altered degradation (a direct effect) and altered transcription (an indirect effect). Future studies could use chemical inhibitors of RNase E (54) or degron tags (55) to fully test these assumptions and better distinguish between direct and indirect effects.

Leadered transcripts appeared to be more sensitive to RNase E levels than leaderless transcripts, suggesting that 5′ UTRs may serve as platforms for engagement with RNase E. However, there was no correlation between degree of stabilization upon *rne* knockdown and predicted secondary structure near the 5′ ends of transcripts. This suggests that the effects of 5′ UTRs on RNase E engagement cannot be explained simply by availability of 5′ ends. This finding is somewhat surprising given the reported strong effect of 5′ end engagement on RNase E activity in *E. coli* (39,56-58), and reports of 5′ end secondary structure protecting transcripts from degradation in *E. coli* (44,59). It is possible that mycobacterial RNase E is less 5′-end dependent than *E. coli* RNase E, or that other transcript features are more important determinants of sensitivity to RNase E.

Our observation of reduced transcription upon *rne* knockdown is consistent with prior work in *E. coli* showing that transcription rate was proportional to growth rate for most genes, while mRNA

degradation rates were inversely proportional to growth rate (22). We have found the same in *M. smegmatis*; a group of genes analyzed by qPCR had both slower degradation and lower steady-state abundance in carbon starvation and in hypoxia compared to log phase, indicating that transcription rate must be slower in the stress conditions than in log phase (21). mRNA degradation and transcription therefore appear to be coordinated in response to energy availability. While in the current work energy was not limiting and growth was not slowed, a coordination of mRNA degradation and transcription was evident. The mechanism of this coordination is unknown. Some mechanisms are known to regulate transcription in a widespread fashion in response to energy stress. For example, the stringent response represses transcription in response to starvation in *E. coli* and *B. subtilis* through distinct mechanisms (direct binding to RNA polymerase and depletion of GTP pools, respectively) and appears to have a similar function in mycobacteria (recently reviewed in (60)) although the mechanism is unknown. However, the stringent response didn't affect mRNA degradation rates in *M. smegmatis* (21). Another known mechanism of global transcriptional repression in mycobacteria is upregulation of a small RNA called Ms1 in *M. smegmatis* that competes with the housekeeping sigma factor for association with RNA polymerase (61). However, Ms1 abundance was not affected by *rne* knockdown.

Our data implicate RNase E as the enzyme responsible for mRNA cleavage events that produce 5' ends with monophosphorylated cytidines, which are widespread in vivo in both *M. smegmatis* (17) and *M. tuberculosis*. This cleavage sequence preference differs from what was reported in a previous study of the in vitro activity of *M. tuberculosis* RNase E (18). In that study, the presence of a single cytidine in an otherwise mono-uridine oligo was inhibitory to cleavage. However, the effects of cytidines in other sequence contexts were not tested. Our results are therefore not

inconsistent with that study, but rather expand upon it. The strong preference of mycobacterial RNase E to cleave 5′ of cytidines contrasts with the lack of strong base specificity by *E. coli* and *Synechocystis* sp. PCC 6803 RNase E at the +1 position ((62) and reviewed in (12)). Residue F67 in *E. coli* RNase E is highly conserved among the Proteobacteria and was proposed to play a key role in the catalytic mechanism by forming a binding pocket for the base one or two nt downstream of the cleavage site (40). Mutating this residue to Ala in *E. coli* abolished activity in vitro (40). However, the residue at the equivalent position in both *M. smegmatis* and *M. tuberculosis* is Val. It is tempting to speculate that differences in the key residues that position the RNA substrate in the active site are responsible for the differences in cleavage sequence preference for mycobacterial vs *E. coli* RNase E. Further work is needed to investigate this question.

Both our in vivo and in vitro data indicate that while RNase E has a strong preference for cleaving 5′ of cytidines, the impact of the surrounding sequence is weak. This could mean that the identities of the surrounding nt are unimportant for RNase E binding and cleavage, or that the identities of those nt are important but act in combinatorial ways that are not obvious from the data currently available. Interpretation of the in vivo cleavage patterns is complicated because (1) cleavage is likely affected by ribosomes and RNA-binding proteins that protect or expose particular regions and (2) cleavage products that are rapidly degraded are not detected and our methods therefore are biased towards identification of cleavage events that produce stable products. In vitro, there was a clear preference for cleavage 5′ of cytidines, and mutation of the cytidine at one cleavage site to guanosine changed the position of cleavage. However, there were many cytidines that did not produce detectable cleavage products, indicating that RNase E prefers certain positions within the test substrate. We examined secondary structure predictions of the substrate and found that

the cleaved positions did not correspond to the positions most likely to be in single-stranded loops. The in vitro cleavage pattern therefore cannot be easily explained by the predicted secondary structure. Stem-loops near cleavage sites have been shown to stimulate or direct cleavage by *E. coli* RNase E in some contexts (63-65), and therefore the sites cleaved in our study could be dictated in part by such cis-acting elements. Cis-acting unpaired regions have also been shown to affect cleavage by *E. coli* RNase E (66). The potential impact of the scaffold domains (which were partially deleted in our purified RNase E) should also be considered, as the *E. coli* RNase E scaffold domains were recently shown to affect catalytic activity (67).

Our study highlights the differences in the types of data obtained from different methods of RNA cleavage-site analysis, as well as some of the challenges in identifying RNA cleavage sites. Ligation-based methods, as we used here for *M. tuberculosis* and as we and many others have used in the past for other bacteria, precisely reveal 5′ ends generated by RNA cleavage. However, 5′ ends are only detected from cleavage events that produce relatively stable fragments with sequence and secondary structure characteristics amenable to ligation. Fragments 5′ of cleavage sites are not captured at all; these can be captured by 3-end ligation approaches, but analysis of the resulting datasets is challenging because 3′ ends generated by many RNases (including RNases E, J, and III) are chemically indistinguishable from 3′ ends generated by transcription termination. The ligation-independent method reported previously (43) and modified here, in contrast, does not identify precise cleavage site locations but may give a broader view of the ubiquity and sequence context of cleavage sites attributable to a particular RNase engineered to be induced or repressed. Ligation-based methods may be more useful for identifying cleavage

products that are stable and functional, while the ligation-independent approach may provide a more accurate view of the breadth of action of RNases of interest.

It is important to note that for both methods, there is no readily definable cutoff for identifying cleavage sites. It is therefore not possible to conclusively determine the total number of cleavage sites in a transcriptome using the combination of methods we have employed. Using read depth filters similar those we previously published for *M. smegmatis*, here we found ~3000 high-confidence *M. tuberculosis* cleavage sites with the ligation-based method. Relaxing one of the filters produced a set of ~10,000 putative cleavage sites with a similar but slightly weaker sequence context signature. Our data suggest a scenario in which the transcriptome contains many cleavage sites, some that are cleaved frequently and/or produce relative long-lived products, and others that are cleaved infrequently and/or produce relatively short-lived products. If this is true, further relaxing the filters would likely reveal still more cleavage sites, likely mixed with a greater proportion of false positives. Some sites may be cleaved so infrequently that their products are not distinguishable from noise. Together, this is consistent with (1) the underlying biology of RNases that have low sequence specificity and/or cleave at ubiquitous sequences (eg, upstream of a cytidine), (2) the fact that mRNA cleavage in vivo is affected by binding of macromolecules such as ribosomes and sRNAs, and (3) the reality that some cleavage products are extremely short-lived and difficult to detect by any method.

It is notable that RNase J, a bifunction endo/exonuclease, did not impact the abundance of most transcript 5′ ends in *M. tuberculosis*. This is consistent with the idea that RNase J has a specialized role in degradation of specific types of highly structured transcripts, as we recently reported (3), rather than a global role. It is also consistent with the idea that RNase J and RNase E may cleave

similar sequences (68,69) and therefore have partially redundant activities; however, the start contrast in phenotypes observed in mycobacterial RNase J knockout strains and RNase E knockdown strains suggest such redundancy is limited.

The cleavage sites mapped in *M. tuberculosis* were disproportionately located in untranslated regions. This may reflect the greater accessibility of such regions to RNases, as they lack protection by ribosomes. An intriguing question arising from this observation is the extent to which proteins are produced from translation of cleaved mRNAs. This has been reported in some bacteria, where there are known examples of polycistronic transcripts that are cleaved to produce fragments with different stabilities, leading in some cases to different stoichiometries of proteins encoded in operons (47-53). There is one reported example in mycobacteria but the evidence supporting it are less conclusive (70). Further studies are therefore needed to investigate the functional consequences of stable RNA cleavage products.

# Material and Methods

### Bacterial strains and culture conditions

*Mycolicibacterium smegmatis* strain $mc^2155$ and derivatives (Table 2-1) were grown in Middlebrook 7H9 liquid medium supplemented with glycerol, Tween-80, catalase, glucose, and sodium chloride as described (21) or on Middlebrook 7H10 with the same supplements except for Tween-80. *Mycobacterium tuberculosis* strain H37Rv was grown in the same way with the addition of oleic acid. *Escherichia coli* NEB-5-alpha (New England Biolabs) was used for cloning and BL21(DE3) pLysS was used for protein overexpression. *E. coli* was grown on LB. Liquid cultures were grown at 37°C with a shaker speed of 200 RPM, except for *M. tuberculosis* which was shaken

at 125 RPM. When indicated, anhydrotetracycline was used at 200 ng/mL. Antibiotic concentrations used for mycobacteria were 25 μg/mL kanamycin and 150 μg/mL hygromycin. Antibiotic concentrations used for *E. coli* were 50 μg/mL kanamycin, 150 μg/mL hygromycin, and 34 μg/mL chloramphenicol.

### *M. smegmatis* strain construction

**SS-M_0418:** The repressible *rne* strain was built by mycobacterial recombineering as described (71). A gene replacement cassette was assembled in plasmid pSS187 by NEBuilder HiFi assembly (NEB) and amplified from the plasmid as a linear fragment by PCR. The *rne* TSS is located 236 nt upstream of the translation start site (17), and the core promoter sequence is evident shortly upstream of the TSS as expected. The gene replacement cassette contained nt -846 through -347 relative to the *rne* (msmeg_4626) translation start site (a 500 bp region located upstream of the *rne* native promoter), a hygromycin resistance gene and promoter, the P766(8G) promoter which contains tet operators (tetO), the P766(8G)-associated 5′ UTR, and the first 500 bp of *rne* coding sequence. 2 μg the gene replacement cassette were dialyzed in pure water before transformation into SS-M_0078 (WT *M. smegmatis* with the recombinase plasmid pNit-recET-Kan). Correct integration of this cassette replaced the 346 nt upstream of the *rne* translation start site with the hyg resistance gene and the P766(8G) promoter and 5′ UTR, and was confirmed by sequencing. Counterselection with 15% sucrose was followed by PCR screening to identify an isolate (SS-M_0151) that lost the recombinase plasmid. SS-M_0151 was further transformed with plasmid pSS291 encoding a Tet repressor (TetR) into the L5 phage integration site.

**SS-M_0424:** A hygromycin-resistant control strain was built using the method described for SS-M_0418, the difference being that the target DNA fragment that was transformed into SS-M_0078

only contained the hygromycin resistance cassette with sequence upstream and downstream of position -346 relative to the *rne* translation start site, resulting in insertion of the hyg resistance gene and promoter without deletion of any native sequence.

### RNA extraction, RNAseq library construction, and sequencing

Cultures were grown to an OD of 0.8-0.9, with or without addition of ATc 8 hrs prior, and divided into a series of 14 ml conical tubes. RIF was added to a final concentration of 150 µg/mL and cultures were harvested after 0, 1, 2, 4, 8, 16, or 32 min by freezing in liquid nitrogen. Frozen cultures were stored at -80°C and thawed on ice for RNA extraction. RNA was extracted as in (21). Illumina libraries were constructed and sequenced by the Broad Institute Microbial 'Omics Core using the library construction procedure described in (72).

### cDNA synthesis and quantitative PCR

cDNA was synthesized as described (21) and qPCR was performed using the conditions described in (21) and the primers listed in that work and in Table S2-6.

### Gene re-annotations in *M. smegmatis* and *M. tuberculosis*

For *Mycolicibacterium smegmatis*, we used the genome sequence of *M. smegmatis* $mc^2155$ strain (NC_008596.1) from Mycobrowser Release 4 (73). For gene annotations, we combined all the annotations from PATRIC 3.6.10 (74), Mycobrowser Release 4 (73) and recently identified novel ORFs (17). The combined annotations were first updated with reannotations of 213 genes as previously described (17). Based on the assumption that transcripts starting with AUG or GUG will be translated in a leaderless fashion (35), we then further utilized the transcriptional start sites (TSS) reported in (17) to re-annotate 156 genes whose annotated 5' UTRs started with in-frame

AUG or GUG codons. In these cases the coding sequence was re-annotated to start at the TSS. The resulting annotations were scrutinized to exclude duplications and genes with frame shift errors. The reannotated CDS boundaries are listed in Table S2-8 and were used for all further analyses unless stated otherwise.

For *Mycobacterium tuberculosis*, the genome sequence and original gene annotations of *M. tuberculosis* H37Rv strain (NC_000962.3) were obtained from Mycobrowser Release 4 (73). Then for genes with only one defined TSS, we used the following procedure to determine if the coding sequence starting coordinates would be re-annotated (35). For genes with TSSs upstream of the previously annotated start codon, we re-annotated the start of the coding sequence to the TSS for those genes with in-frame AUG or GUG at the 5′ end of the transcript. For genes that had a single TSS downstream of the previously annotated start codon, the start of the coding sequence was re-annotated to the position of the TSS if the TSS was at an in-frame AUG or GUG within the first 30% of the previously annotated coding sequence. If the TSS was not at an in-frame AUG or GUG, we re-annotated the start of the coding sequence only if the next in-frame start codon (AUG, GUG, or UUG) was found in the first 30% of the previously annotated coding sequence. The reannotated CDS boundaries are listed in Table S2-9 and were used for all further analyses unless stated otherwise.

## RNAseq data analysis for differential expression analysis

The 0-min RIF-treated samples were used to measure and compare steady-state transcript abundance. Reads were aligned to *M. smegmatis* $mc^2155$ reference sequence NC_008596.1 from Mycobrowser Release 4 (73) with Bowtie v1.2.2 (75), read alignment processed by SAMtools v1.9

(76), counts determined by HTSeq v0.10.1 (77). The differential expression analysis was performed using Clipper with the gene counts normalized by qPCR normalization factors (41).

### Gene set enrichment analysis

The enrichment of KEGG pathway was tested using ClusterProfiler v4.4.4 (42), based on the gene list sorted by $\log_2$ fold changes of expected and observed abundance. FDR-adjusted *p*-values were used for multiple testing correction.

### RNAseq data analysis for expression-library-based cleavage site analysis in *M. smegmatis* and *M. tuberculosis*

This analysis was performed on the *M. smegmatis* 0-min RIF-treated samples as well as an *M. tuberculosis rne* knock down strain and corresponding control strain ((16), GEO accession GSE126286). Quality control was performed using FastQC. Reads were first scanned from 5′ end to 3′ end and cut once the average quality per base of 4-base wide sliding window dropped below 20. After such processing, reads with less than 25 bases were discarded using Trimmomatic v0.39 (78). Reads were aligned using Bowtie2 v2.4.5 (79) with the "--very-sensitive" option. We first aligned reads to tRNA and rRNA sequences only. The remaining reads were aligned to NC_008596.1 (*M. smegmatis*) or NC_000962.3 (*M. tuberculosis*). Via SAMtools v1.16.1 (76), we filtered the resulting alignments by keeping only the primary alignments with MAPQ at least 10. The aligned, filtered reads that mapped in proper pairs were split into their corresponding strands to quantify strand-specific coverage at the single-nucleotide level using BEDTools v2.30.0 (80). The coverage for each gene was then calculated by summing the single-coordinate coverage within the gene, and the average coordinate coverage for each gene was calculated by dividing the summed coverage by gene length. We only kept genes with average coordinate coverage at least

5 in all replicates and conditions. For those qualified genes, we excluded coordinates at overlapped gene regions for downstream analysis. To correct for the variability in expression level among genes and between conditions, we normalized single-coordinate coverages using the whole-gene coverages. The single-coordinate coverages were divided by the total summed coordinate coverage of each gene (excluding regions overlapping other genes) after adding one pseudocount to all coordinate positions. The final normalized coverage of each coordinate was the average of triplicates in each condition.

The coverage ratio at each qualified coordinate position between any two conditions was then calculated as the $\log_2$(Condition1/Condition2) ratio. For each group of coordinates under investigation (eg, coordinates with $\log_2$ ratios in the top 5%), we quantified the sequence context using the relative base frequency of the 20 coordinates upstream and downstream of each coordinate in the group.

### RNAseq data analysis for determination of half-lives in *M. smegmatis*

To calculate mRNA half-lives, data from all of the timepoints following RIF treatment were processed. First, reads were aligned using BWA-MEM v0.7.17 (81). Next, the resulting alignments were processed for each strand by SAMtools v1.10 (76). The raw coverage of each coordinate was calculated through BEDTools v2.29.1 (80). Then we conducted a two-step normalization of the raw coverage. First, coverage was normalized by the total number of reads in each library. Then we calculated normalization factors by performing qPCR to determine the relative expression levels of eight genes (*sigA*, *rraA*, *esxB*, *atpE*, *rne*, msmeg_4665, msmeg_5691, msmeg_6941; Table S2-7) at each sample and timepoint compared to the average of the 0-min RIF control strain (no ATc) samples. qPCR was done with cDNA made from random priming as described above, separately

from RNAseq library construction. Each qPCR reaction was performed using 400 pg of cDNA. As ribosomal rRNA depletion was not performed, the CTs obtained from the qPCR reflect the expression level of the target gene relative to the total RNA pool, which is primarily rRNA. Normalization factors were calculated separately for the region amplified by qPCR in each of the eight genes and averaged. Specifically, for a given sample $T_n$, we calculated the normalization factor $F_{Tn}$ from the qPCR target gene expression measurements as indicated below:

Calculation of the expected RNAseq coverage ($T_{n,i,RNAseq\_expected}$) for each qPCR amplicon region ($_i$) in each sample ($T_n$), where $T_0$ represents the average value for the control strain without ATc immediately after addition of RIF, and *qPCR* represents relative abundance of the amplicon determined by qPCR:

$$T_{n,i,RNAseq\_expected} = \left(\frac{T_{n,i,qPCR}}{T_{0,i,qPCR}}\right) * T_{0,i,RNAseq\_actual}$$

Calculation of a global normalization factor ($F_{Tn}$) by calculating and averaging the normalization factors for each qPCR amplicon region:

$$F_{T_n} = \frac{1}{8}\sum_{i=1}^{8} \frac{T_{n,i,RNAseq\_expected}}{T_{n,i,RNAseq\_actual}}$$

Then the final normalized coverage for each coordinate was calculated by multiplying the first step normalized coverage by the global normalization factor for each sample. The coverage for each gene was then represented by the summation of the normalized coverage of its coordinates, divided by the gene length.

**Estimation of transcription rates**

Estimated transcription rates were calculated as a function of steady-state abundance and mRNA degradation rate as described (37) and as follows:

Transcription rate = VT = (k*mRNA) + (μ*mRNA)

mRNA = steady-state mRNA abundance (taken from 0 min RIF treatment)

k = degradation rate = ln(2)/half life

μ = growth rate = ln(2)/doubling time

Doubling time = 150 min

The estimated transcription rate units are arbitrary and therefore useful only for comparison of genes or conditions within this study.

### mRNA cleavage site mapping in *M. tuberculosis*

Mapping of *M. tuberculosis* TSSs was previously described (35). The same dataset was used to identify mRNA cleavage sites. All analyses of this dataset were done using the genome annotations in NC_000962.gbk rather than the reannotations shown in Table S2-9. As described in (35), RNA 5' ends were identified, filtered based on absolute read depth and read depth relative to local expression library coverage, and subject to Gaussian mixture modeling to distinguish between TSSs and cleavage sites on the basis of relative coverage in libraries from RNA treated with RppH ("converted," capturing both TSSs and cleavage sites) and libraries from untreated RNA ("non-converted," capturing primarily cleavage sites). 5′ ends with converted/non-converted library read depth ratios less than 1.39 had a cumulative probability of ≤0.01 of belonging to the TSS population (after adjusting for multiple comparisons by the Benjamini-Hochberg procedure)

and were therefore designated RNA cleavage sites. Because cleavage may be imprecise, filtering was performed to retain the single cleavage site with the greatest converted-library read coverage in each 5 nt window. This resulted in the 2983 high-confidence cleavage sites reported in Table S2-5A. The longer list of putative cleavage sites reported in Table S2-5I was obtained by applying the same converted/non-converted ratio cutoff to a list of 5′ ends from earlier in the pipeline prior to filtering on coverage relative to local expression library coverage. Instead, only a filter requiring a minimum mean converted library read depth of 20 was applied. This resulted in 10795 putative cleavage sites.

### *M. tuberculosis* TSS analyses

TSSs from the above dataset were considered to be associated with the 5′ ends of genes if they were either (1) within 500 nt upstream of an annotated start codon or (2) within the first 25% of an annotated coding sequence. TSSs were considered to be internal within coding sequences if they were located between 25% and 80% of the way through annotated coding sequences. TSSs were considered to be associated with putative antisense transcripts if they did not meet any of the above criteria and were either (1) located on the opposite strand of an annotated coding sequence or (2) located <200 nt from the end of an annotated coding sequence on the opposite strand. TSSs were considered to be intergenic if they did not meet any of the above criteria for 5′-end associated, internal, or antisense transcripts.

Genes were assigned to operons if they were transcribed consecutively on the same strand and if both of the following criteria were met: (1) Only the first gene had an assigned TSS and (2) The downstream gene(s) were sufficiently expressed. Sufficient expression was defined as having a Reads Per Kilobase of transcript per Million mapped reads (RPKM) value in corresponding RNA-

seq expression libraries equal to the 5th percentile or above of RPKM values for all genes with TSSs. This prediction algorithm is conservative and excludes many loci that may be transcribed both polycistronically and monocistronically.

### Analysis of *M. tuberculosis* cleavage site locations relative to genes

We determined the number of coordinates in the *M. tuberculosis* genome that fell into each of the following four categories of regions: (1) coding sequences; (2) 5′ UTRs of genes with mapped TSSs and for which the next upstream gene was encoded on the opposite strand; (3) regions between the coding sequences of two consecutive genes encoded on the same strand for which both genes had mapped TSSs (not exclusive operons); and (4) regions between the coding sequences of two consecutive genes encoded on the same strand for which only the first gene had a mapped TSS (exclusive operons). We then determined the number of cleavage sites that were located within each of these regions. The **expected** frequency of cleavage sites in each region was defined as:

> (number of coordinates in region/sum of coordinates in all four regions)*total number of cleavage sites in all four regions.

The observed number of cleavage sites in each region was then divided by the expected number to obtain the values plotted if Figure 2-6C.

### Secondary structure prediction

Free energy of RNA folding and basepair probabilities for minimum free energy structure were predicted using the Vienna RNA Package utility RNAfold (82). For Figure 2-6B, the 200 nt region spanning each RNA cleavage site was extracted and the minimum free energy of secondary structure formation was predicted for 39 nt sliding windows across each such region. The data

plotted are the mean and the 25th and 75th percentile minimum free energies of 39 nt windows centered around each relative coordinate in all cleaved RNAs.

### 5′ RACE to map a putative RNase E cleavage site in the rRNA transcript

Enzymes were obtained from New England Biolabs unless otherwise specified. Five hundred ng of each RNA sample were mixed with 1 µg of oligo SSS1016 in a total volume of 9 µl, incubated at 65°C for 10 minutes, and cooled on ice for 5 minutes. Each sample was combined with 21 µl of ligation mix containing 10 µl of 50% PEG8000, 3 µl of 10X T4 RNA ligase buffer, 3 µl of 10 mM ATP, 3 µl of DMSO, 1 µl of murine RNase inhibitor, and 1 µl of T4 RNA ligase. Samples were incubated at 20°C overnight and purified with a Zymo RNA Clean & Concentrator-5 kit according to the manufacturer's instructions with the following modifications: samples were first diluted by addition of 20 µl of RNase-free water, and samples were eluted in 8 µl of RNase-free water. Three µl of each purified ligation were then subject to cDNA synthesis or mock (no-RT) cDNA synthesis. Samples were combined with 1 µl of a mix containing 50 mM Tris pH 7.5 and 500 ng/µl random primers (Invitrogen), incubated at 70°C for 10 minutes, and snap-cooled in an ice-water bath. cDNA synthesis was done as described (21). 35 ng of cDNA or the equivalent volume of the corresponding no-RT sample were mixed with 2.5 µl 10X Taq buffer, 1.25 µl each 10 µM primers SSS1017 and SSS2210, 1.25 µl DMSO, 0.5 µl of 10 mM each dNTP mix, 0.167 µl Taq polymerase, and water to a final volume of 25 µl. Cycling conditions were 5 minutes at 95°C, 35 cycles of 30 seconds at 95°C, 20 seconds at 52°C, and 25 seconds at 68°C, and a final 5 minute incubation at 68°C. PCRs were run on 1.5% agarose gels and bands that appeared in cDNA samples but not in no-RT samples were excised and sequenced with SSS2210 to identify the adapter/RNA junctions.

## Overexpression and purification of recombinant RNase E variants

Two RNase E variants were recombinantly expressed and purified for in vitro RNA cleavage assays: residues 146-824 (partial N-terminal truncation and full C-terminal truncation), and residues 146-824 with D694R and D738R mutations. pSS348, carrying the *M. smegmatis rne* coding sequence with a Δ1-145aa partial N-terminal deletion, Δ825-1037aa full C-terminal deletion, and an N-terminal addition of 6XHis tag, 3XFLAG tag, TEV protease cleavage site, and 4XGly linker sequences, was used as a template for creation of pSS420, which encodes RNase E residues 146-824 with the indicated tags in a pET38 backbone. pSS420 was then used as a template for creation of pSS421, which has the mutations D694R and D738R, predicted to abolish catalytic activity (40). All constructs were sequenced to confirm the success of point mutations and truncations.

*E. coli* strain BL21(DE3)pLysS was transformed with each of the RNase E expression plasmids and 500-1000 mL cultures were grown to an OD600 of ~0.5, then induced with 400 $\mu$M IPTG and incubated at 28°C for four hours prior to harvest. For the protein used in Figure 2-5A, Pellets were resuspended in 1X IMAC buffer (20 mM Tris-HCl pH 7.9, 150 mM NaCl, 5% glycerol, 0.01% Igepal) containing 10 mM imidazole and lysed with a BioSpec Tissue-Tearor (10 cycles of 15-30 seconds each at maximum speed, with 30-60 seconds on ice between cycles). Lysates were cleared by centrifugation, incubated for 30-60 minutes on ice with 4 ml His-Pur Ni-NTA resin 50% slurry (Thermo Scientific), washed with IMAC buffer containing 10 mM imidazole, and eluted with IMAC buffer containing 150 mM imidazole. For the proteins used in Figure 2-5D and E, the NaCl concentration in the lysate was increased to 1 M before mixing with resin pre-equilibrated in the same, and the wash buffer contained 1 M NaCl. The lysis buffer also included 1X Halt™ Protease Inhibitor Cocktail, EDTA-Free (ThermoFisher), 40 mg of lysozyme, and 16 U Turbo DNase

(Invitrogen). Eluates were concentrated with Microcon PL-30 (30,000 NMWL) protein concentrators (Millipore Sigma) and loaded onto 1 cm diameter, 38 mL Sephacryl S-200 High Resolution resin (GE Healthcare) size exclusion chromatography columns. Flow rate was regulated using a Masterflex C/L pump. The buffer was 1X IMAC with the addition of 1 mM EDTA and 1 mM DTT.

**Preparation of in vitro-transcribed RNA substrates**

Genomic DNA was used as a template to produce PCR products containing portions of the *atpB-atpE* locus downstream of the T7 Phi2.5 promoter and sequence needed for A-initiated transcription (TAATACGACTCACTATT**A**GG, where transcription initiates at the bolded "A"). One PCR product had the promoter oriented to produce the sense strand, and the other was shorter and had the promoter oriented to produce a partial antisense strand (Figure S2-11). Monophosphorylated RNA was synthesized from each of these PCR products in the presence of a 50-fold molar excess of AMP over ATP (83) with T7 RNA polymerase (NEB M0251). Each 50 µL reaction contained 1X reaction buffer, 5 mM DTT, 1 mM UTP, 1 mM CTP, 1 mM GTP, 0.5 mM ATP, 25 mM AMP, 5 units/µL T7 RNA polymerase, 1 unit/µL Murine RNase inhibitor, and 2 µg DNA template. Reactions were incubated at 37°C for 16 hours. The resulting transcripts were treated with TURBO DNase at 37°C for 30 minutes before purification with a Zymo RNA Clean & Concentrator-5 kit.

The *atpB-E* sense transcript and anti-sense transcript were combined at a 1:1 molar ratio and the mixtures were incubated in the presence of 5X annealing buffer (50 mM Tris-HCl, pH 7.9, 0.5 mM EDTA, pH 8.0, 100 mM NaCl) in a 10 µL reaction for 1 min at 90°C, then slowly cooled down to

room temperature over a period of approximately 30 min. The resulting annealed RNA mix was immediately stored at -80°C.

The 50 nt substrate (Figure 2-5C) was synthesized using the same conditions, except the in vitro transcription templates were annealed oligos (Table S2-7) rather than PCR products. Smaller molecular weight standards (Table S2-7) were also made by in vitro transcription from annealed oligos. 25 μM of each of the two DNA oligos were incubated in annealing buffer (10 mM Tris, 50 mM NaCl, and 1 mM EDTA) at 95°C for 2 minutes, followed by 47 cycles of 1.5 minutes starting at 95°C and decreasing by 1.5 degrees per cycle.

### In vitro RNase E cleavage reactions

In vitro RNase E cleavage reactions were heated at 65°C for 3 min prior to adding the enzyme, then cooled and incubated at 37°C for 1-2 hours following addition of the enzyme. The reaction buffer was composed of 20 mM Tris-HCl, pH 7.9, 100 mM NaCl, 5% Glycerol, 0.01% IGEPAL, 0.1 mM DTT, 10 mM $MgCl_2$, and each reaction containing 150-300 ng annealed RNA mix and 80 ng of purified RNase E. For the reactions shown in Figure 2-5D and E, the buffer included 10 μM $ZnCl_2$. For mock reactions, water was used instead of enzyme. Reactions were stopped by adding equal volumes of 2X Invitrogen™ Gel loading buffer II and then subjected to electrophoresis on a15%, 7.5%, or 5% polyacrylamide-8 M urea gels and visualized after 15 min staining with SYBR Gold Nucleic Acid gel stain. When indicated, bands of interest were excised, and RNA was recovered using Zymo small-RNA PAGE recovery kitfor 5′ RACE or 3′ RACE.

### 5′ RACE and 3′ RACE to map cleavage sites from in vitro RNase E cleavage reactions

For 5′ RACE, RNA extracted from bands as described above was mixed with 1 µg of RNA oligo SSS1016 in a total volume of 9 µL at 65°C for 5 min, chilled on ice and then combined with 30 U T4 RNA Ligase 1 (NEB M0437M), 40 U Murine RNase Inhibitor (NEB), 10% DMSO, 1 mM ATP, 1X T4 RNase Ligase 1 reaction buffer, and 16.7% PEG 8000 in reactions with a total volume of 30 µL. Reactions were incubated at 20°C for 18 hours followed by column purification. cDNA was synthesized using the reverse oligo SSS916 which anneals close to 3′ end of the sense strand and the cDNA synthesis protocol described above. cDNA was purified and then was used as template to perform Taq PCR with primers SSS1018 and SSS916. Purified PCR products were sequenced with oligo SSS916.

For 3′ RACE, RNA extracted from bands as described above was mixed with 1 µg RNA oligo SSS2433 (which has a 5' monophosphate and a 3' inverted deoxythymidine and was modified from (84) at 65°C for 5 min, chilled on ice and incubated at 17°C for 18 hours with the same reaction mix as used for 5′ RACE above. Following column purification, cDNA was synthesized using reverse oligo SSS2434 which anneals to the 3′ adapter, and the protocol described above. cDNA was purified and then was used as template to perform Taq PCR with primers SSS917 and SSS2434. Purified PCR products were sequenced with oligo SSS917.

**Statistical analyses and scripts.**

Statistics shown in Figures 2-1, 2-2, and 2-6 were done in Graphpad Prism version 9.2.0. The scripts for RNAseq processing, analysis and result visualization are available on Github (https://github.com/ssshell/Mycobacterial_RNase_E).

**Table 2-1. Strains and plasmids used in this study**

| Species | Strain | Plasmid | Description | Source |
|---|---|---|---|---|
| *M. smegmatis* | $mc^2155$ | None | Widely-used lab strain | ATCC |
| *M. smegmatis* | SS-M_0424 | pSS291: tetR38 driven by promoter ptb38, L5 integrating, $kan^R$ | $mc^2155$ with the $hyg^R$ gene inserted with its own promoter 347 nt upstream of, and divergent from, the *rne* translation start site. | This study |
| *M. smegmatis* | SS-M_0418 | pSS291: tetR38 driven by promoter ptb38, L5 integrating, $kan^R$ | $mc^2155$ in which the *rne* (msmeg_4626) promoter and UTR (nt -346 through -1 relative to the *rne* start codon) were replaced by the P766(8G) promoter and associated 5' UTR (19). Additionally, the $hyg^R$ gene was inserted with its own promoter upstream of, and divergent from, the P766(8G) promoter. | This study |
| *M. tuberculosis* | H37Rv | None | Widely-used lab strain | ATCC |
| *M. tuberculosis* | H37Rv Δ*rnj* | None | The *rnj* coding sequence was replaced with the $hyg^R$ coding sequence. | (3) |
| *E. coli* | BL21 DE3 pLysS | pSS420: pET38 expressing residues 146-824 of *M. smegmatis* RNase E with N-terminal 6XHis, 3XFLAG, TEV protease cleavage site, and 4XGly linker. | | This work |
| *E. coli* | BL21 DE3 pLysS | pSS421: pSS420 with mutations D694R and D738R. | | This work |

# Data Availability

All RNAseq data generated in this study are available at GSE227248. The scripts for RNAseq processing, analysis and result visualization are available on Github (https://github.com/ssshell/Mycobacterial_RNase_E).

## Acknowledgements

We thank members of the Shell and Fortune labs for helpful discussions.

## Funding

This study was funded in part by the following: NSF-CAREER award 1652756 to SSS; NIH-NIAID award P01 AI143575 to SMF, SSS, and TRI; NIH-NIAID award U19 AI107774 to SMF; and NIH-NIAID award F32 AI085911 to SSS. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or National Science Foundation.

## Supporting Information

Table S2-1-S2-9 can be accessed online:
https://www.sciencedirect.com/science/article/pii/S0021925823023402?via%3Dihub

## References

1. WHO. (2021). World Health Organization, Geneva.

2. Njire, M., Wang, N., Wang, B., Tan, Y., Cai, X., Liu, Y., Mugweru, J., Guo, J., Hameed, H.M.A., Tan, S. *et al.* (2017) Pyrazinoic Acid Inhibits a Bifunctional Enzyme in Mycobacterium tuberculosis. *Antimicrob Agents Chemother*, **61**.

3. Martini, M.C., Hicks, N.D., Xiao, J., Alonso, M.N., Barbier, T., Sixsmith, J., Fortune, S.M. and Shell, S.S. (2022) Loss of RNase J leads to multi-drug tolerance and accumulation of highly structured mRNA fragments in Mycobacterium tuberculosis. *PLoS Pathog*, **18**, e1010705.

4. He, L., Cui, P., Shi, W., Li, Q., Zhang, W., Li, M. and Zhang, Y. (2019) Pyrazinoic Acid Inhibits the Bifunctional Enzyme (Rv2783) in Mycobacterium tuberculosis by Competing with tmRNA. *Pathogens*, **8**.

5. Hicks, N.D., Yang, J., Zhang, X., Zhao, B., Grad, Y.H., Liu, L., Ou, X., Chang, Z., Xia, H., Zhou, Y. *et al.* (2018) Clinically prevalent mutations in Mycobacterium tuberculosis alter propionate metabolism and mediate multidrug tolerance. *Nat Microbiol*, **3**, 1032-1042.

6. Consortium, T.C. (2022) Genome-wide association studies of global Mycobacterium tuberculosis resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms. *PLoS Biol*, **20**, e3001755.

7. Babitzke, P. and Kushner, S.R. (1991) The Ams (altered mRNA stability) protein and ribonuclease E are encoded by the same structural gene of Escherichia coli. *Proc Natl Acad Sci U S A*, **88**, 1-5.

8. Carpousis, A.J., Van Houwe, G., Ehretsmann, C. and Krisch, H.M. (1994) Copurification of E. coli RNAase E and PNPase: evidence for a specific association between two enzymes important in RNA processing and degradation. *Cell*, **76**, 889-900.

9. Lin-Chao, S., Wong, T.T., McDowall, K.J. and Cohen, S.N. (1994) Effects of nucleotide sequence on the specificity of rne-dependent and RNase E-mediated cleavages of RNA I encoded by the pBR322 plasmid. *The Journal of biological chemistry*, **269**, 10797-10803.

10. McDowall, K.J., Lin-Chao, S. and Cohen, S.N. (1994) A+U content rather than a particular nucleotide order determines the specificity of RNase E cleavage. *J Biol Chem*, **269**, 10790-10796.

11. Py, B., Higgins, C.F., Krisch, H.M. and Carpousis, A.J. (1996) A DEAD-box RNA helicase in the Escherichia coli RNA degradosome. *Nature*, **381**, 169-172.

12. Mackie, G.A. (2012) RNase E: at the interface of bacterial RNA processing and decay. *Nature Reviews Microbiology*, **11**, 45-57.

13. Sassetti, C.M., Boyd, D.H. and Rubin, E.J. (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Molecular Microbiology*, **48**, 77-84.

14. Dejesus, M.A., Gerrick, E.R., Xu, W., Park, S.W., Long, J.E., Boutte, C.C., Rubin, E.J., Schnappinger, D., Ehrt, S., Fortune, S.M. *et al.* (2017) Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis. *mBio*, **8**.

15. Taverniti, V., Forti, F., Ghisotti, D. and Putzer, H. (2011) Mycobacterium smegmatis RNase J is a 5'-3' exo-/endoribonuclease and both RNase J and RNase E are involved in ribosomal RNA maturation. *Molecular Microbiology*, **82**, 1260-1276.

16. Plocinski, P., Macios, M., Houghton, J., Niemiec, E., Plocinska, R., Brzostek, A., Slomka, M., Dziadek, J., Young, D. and Dziembowski, A. (2019) Proteomic and transcriptomic experiments reveal an essential role of RNA degradosome complexes in shaping the transcriptome of Mycobacterium tuberculosis. *Nucleic Acids Res*, **47**, 5892-5905.

17. Martini, M.C., Zhou, Y., Sun, H. and Shell, S.S. (2019) Defining the Transcriptional and Post-transcriptional Landscapes of Mycobacterium smegmatis in Aerobic Growth and Hypoxia. *Front Microbiol*, **10**, 591.

18. Zeller, M.-E., Csanadi, A., Miczak, A., Rose, T., Bizebard, T. and Kaberdin, V.R. (2007) Quaternary structure and biochemical properties of mycobacterial RNase E/G. *The Biochemical journal*, **403**, 207.

19. Johnson, E.O., LaVerriere, E., Office, E., Stanley, M., Meyer, E., Kawate, T., Gomez, J.E., Audette, R.E., Bandyopadhyay, N., Betancourt, N. *et al.* (2019) Large-scale chemical-genetics yields new M. tuberculosis inhibitor classes. *Nature*, **571**, 72-78.

20. Wei, J.-R., Krishnamoorthy, V., Murphy, K., Kim, J.-H., Schnappinger, D., Alber, T., Sassetti, C.M., Rhee, K.Y. and Rubin, E.J. (2011) Depletion of antibiotic targets has widely varying effects on growth. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 4176-4181.

21. Vargas-Blanco, D.A., Zhou, Y., Zamalloa, L.G., Antonelli, T. and Shell, S.S. (2019) mRNA Degradation Rates Are Coupled to Metabolic Status in Mycobacterium smegmatis. *mBio*, **10**.

22. Esquerré, T., Laguerre, S., Turlan, C., Carpousis, A.J., Girbal, L. and Cocaign-Bousquet, M. (2014) Dual role of transcription and transcript stability in the regulation of gene expression in Escherichia coli cells cultured on glucose at different growth rates. *Nucleic Acids Research*, **42**, 2460-2472.

23. Rustad, T.R., Minch, K.J., Brabant, W., Winkler, J.K., Reiss, D.J., Baliga, N.S. and Sherman, D.R.J.N.a.r. (2013) Global analysis of mRNA stability in Mycobacterium tuberculosis. **41**, 509-517.

24. Chen, H., Shiroguchi, K., Ge, H. and Xie, X.S. (2015) Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli. *Molecular systems biology*, **11**, 781.

25. Nouaille, S., Mondeil, S., Finoux, A.L., Moulis, C., Girbal, L. and Cocaign-Bousquet, M. (2017) The stability of an mRNA is influenced by its concentration: a potential physical mechanism to regulate gene expression. *Nucleic Acids Res*, **45**, 11711-11724.

26. Morin, M., Enjalbert, B., Ropers, D., Girbal, L. and Cocaign-Bousquet, M. (2020) Genomewide Stabilization of mRNA during a "Feast-to-Famine" Growth Transition in Escherichia coli. *mSphere*, **5**.

27. Redon, E., Loubière, P. and Cocaign-Bousquet, M. (2005) Role of mRNA stability during genome-wide adaptation of Lactococcus lactis to carbon starvation. *The Journal of biological chemistry*, **280**, 36380-36385.

28. Dressaire, C., Picard, F., Redon, E., Loubière, P., Queinnec, I., Girbal, L. and Cocaign-Bousquet, M. (2013) Role of mRNA stability during bacterial adaptation. *PLoS ONE*, **8**, e59059.

29. Rustad, T.R., Minch, K.J., Brabant, W., Winkler, J.K., Reiss, D.J., Baliga, N.S. and Sherman, D.R. (2012) Global analysis of mRNA stability in Mycobacterium tuberculosis. *Nucleic Acids Research*, **41**, 509-517.

30. Bernstein, J.A., Khodursky, A.B., Lin, P.-H., Lin-Chao, S. and Cohen, S.N. (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proceedings of the National Academy of Sciences*, **99**, 9697-9702.

31. Esquerré, T., Moisan, A., Chiapello, H., Arike, L., Vilu, R., Gaspin, C., Cocaign-Bousquet, M. and Girbal, L. (2015) Genome-wide investigation of mRNA lifetime determinants in Escherichia coli cells cultured at different growth rates. *BMC Genomics*, **16**, 275.

32. Moffitt, J.R., Pandey, S., Boettiger, A.N., Wang, S. and Zhuang, X. (2016) Spatial organization shapes the turnover of a bacterial transcriptome. *eLife*, **5**.

33. Kristoffersen, S.M., Haase, C., Weil, M.R., Passalacqua, K.D., Niazi, F., Hutchison, S.K., Desany, B., Kolstø, A.-B., Tourasse, N.J., Read, T.D. *et al.* (2012) Global mRNA decay analysis at single nucleotide resolution reveals segmental and positional degradation patterns in a Gram-positive bacterium. *Genome Biology*, **13**, R30.

34. Cortes, T., Schubert, O.T., Rose, G., Arnvig, K.B., Comas, I., Aebersold, R. and Young, D.B. (2013) Genome-wide Mapping of Transcriptional Start Sites Defines an Extensive Leaderless Transcriptome in Mycobacterium tuberculosis. *Cell reports*.

35. Shell, S.S., Wang, J., Lapierre, P., Mir, M., Chase, M.R., Pyle, M.M., Gawande, R., Ahmad, R., Sarracino, D.A., Ioerger, T.R. *et al.* (2015) Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLoS Genetics*, **11**, e1005641.

36. Sawyer, E.B., Phelan, J.E., Clark, T.G. and Cortes, T. (2021) A snapshot of translation in Mycobacterium tuberculosis during exponential growth and nutrient starvation revealed by ribosome profiling. *Cell Rep*, **34**, 108695.

37. Nguyen, T.G., Vargas-Blanco, D.A., Roberts, L.A. and Shell, S.S. (2020) The Impact of Leadered and Leaderless Gene Structures on Translation Efficiency, Transcript Stability, and Predicted Transcription Rates in Mycobacterium smegmatis. *J Bacteriol*, **202**.

38. Grabowska, A.D., Andreu, N. and Cortes, T. (2021) Translation of a Leaderless Reporter Is Robust During Exponential Growth and Well Sustained During Stress Conditions in Mycobacterium tuberculosis. *Front Microbiol*, **12**, 746320.

39. Mackie, G.A. (1998) Ribonuclease E is a 5'-end-dependent endonuclease. *Nature*, **395**, 720-723.

40. Callaghan, A.J., Marcaida, M.J., Stead, J.A., McDowall, K.J., Scott, W.G. and Luisi, B.F. (2005) Structure of Escherichia coli RNase E catalytic domain and implications for RNA turnover. *Nature*, **437**, 1187-1191.

41. Ge, X., Chen, Y.E., Song, D., McDermott, M., Woyshner, K., Manousopoulou, A., Wang, N., Li, W., Wang, L.D. and Li, J.J. (2021) Clipper: p-value-free FDR control on high-throughput data from two conditions. *Genome Biol*, **22**, 288.

42. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L. *et al*. (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)*, **2**, 100141.

43. Culviner, P.H. and Laub, M.T. (2018) Global Analysis of the E. coli Toxin MazF Reveals Widespread Cleavage of mRNA and the Inhibition of rRNA Maturation and Ribosome Biogenesis. *Mol Cell*, **70**, 868-880 e810.

44. Richards, J. and Belasco, J.G. (2019) Obstacles to Scanning by RNase E Govern Bacterial mRNA Lifetimes by Hindering Access to Distal Cleavage Sites. *Mol Cell*, **74**, 284-295 e285.

45. Shell, S.S., Chase, M.R., Ioerger, T.R. and Fortune, S.M. (2015) RNA sequencing for transcript 5'-end mapping in mycobacteria. *Methods in molecular biology (Clifton, N.J.)*, **1285**, 31-45.

46. Schifano, J.M., Vvedenskaya, I.O., Knoblauch, J.G., Ouyang, M., Nickels, B.E. and Woychik, N.A. (2014) An RNA-seq method for defining endoribonuclease cleavage specificity identifies dual rRNA substrates for toxin MazF-mt3. *Nature communications*, **5**, 3538.

47. Båga, M., Göransson, M., Normark, S. and Uhlin, B.E. (1988) Processed mRNA with differential stability in the regulation of E. coli pilin gene expression. *Cell*, **52**, 197-206.

48. Nilsson, P. and Uhtin, B.E. (1991) Differential decay of a polycistronic Escherichia coli transcript is initiated by RNaseE-dependent endonucleolytic processing. *Molecular Microbiology*, **5**, 1791-1799.

49. Nilsson, P., Naureckiene, S. and Uhlin, B.E. (1996) Mutations affecting mRNA processing and fimbrial biogenesis in the Escherichia coli pap operon. *Journal of Bacteriology*, **178**, 683-690.

50. Lodato, P.B. and Kaper, J.B. (2009) Post-transcriptional processing of the LEE4 operon in enterohaemorrhagic Escherichia coli. *Molecular Microbiology*, **71**, 273-290.

51. Alifano, P., Fani, R., Liò, P., Lazcano, A., Bazzicalupo, M., Carlomagno, M.S. and Bruni, C.B. (1996) Histidine biosynthetic pathway and genes: structure, regulation, and evolution. *Microbiological reviews*, **60**, 44-69.

52. Ludwig, H., Homuth, G., Schmalisch, M., Dyka, F.M., Hecker, M. and Stülke, J. (2001) Transcription of glycolytic genes and operons in Bacillus subtilis: evidence for the presence of multiple levels of control of the gapA operon. *Molecular Microbiology*, **41**, 409-422.

53. Meinken, C., Blencke, H.-M., Ludwig, H. and Stülke, J. (2003) Expression of the glycolytic gapA operon in Bacillus subtilis: differential syntheses of proteins encoded by the operon. *Microbiology (Reading, England)*, **149**, 751-761.

54. Kime, L., Vincent, H.A., Gendoo, D.M.A., Jourdan, S.S., Fishwick, C.W.G., Callaghan, A.J. and McDowall, K.J. (2015) The First Small-Molecule Inhibitors of Members of the Ribonuclease E Family. *Scientific reports*, **5**, 8028.

55. Kim, J.-H., Wei, J.-R., Wallach, J.B., Robbins, R.S., Rubin, E.J. and Schnappinger, D. (2011) Protein inactivation in mycobacteria by controlled proteolysis and its application to deplete the beta subunit of RNA polymerase. *Nucleic Acids Research*, **39**, 2210-2220.

56. Celesnik, H., Deana, A. and Belasco, J.G. (2007) Initiation of RNA decay in Escherichia coli by 5' pyrophosphate removal. *Molecular cell*, **27**, 79-90.

57. Deana, A., Celesnik, H. and Belasco, J.G. (2008) The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal. *Nature*, **451**, 355-358.

58. Richards, J. and Belasco, J.G. (2016) Distinct requirements for 5'-monophosphate-assisted RNA cleavage by Escherichia coli RNase E and RNase G. *J Biol Chem*, **291**, 20825.

59. Emory, S.A., Bouvet, P. and Belasco, J.G. (1992) A 5'-terminal stem-loop structure can stabilize mRNA in Escherichia coli. *Genes & Development*, **6**, 135-148.

60. Gupta, K.R., Arora, G., Mattoo, A. and Sajid, A. (2021) Stringent Response in Mycobacteria: From Biology to Therapeutic Potential. *Pathogens*, **10**.

61. Hnilicová, J., Jirát Matějčková, J., Siková, M., Pospíšil, J., Halada, P., Pánek, J. and Krásny, L. (2014) Ms1, a novel sRNA interacting with the RNA polymerase core in mycobacteria. *Nucleic Acids Research*.

62. Hoffmann, U.A., Heyl, F., Rogh, S.N., Wallner, T., Backofen, R., Hess, W.R., Steglich, C. and Wilde, A. (2021) Transcriptome-wide in vivo mapping of cleavage sites for the compact cyanobacterial ribonuclease E reveals insights into its function and substrate recognition. *Nucleic Acids Res*, **49**, 13075-13091.

63. Bandyra, K.J., Wandzik, J.M. and Luisi, B.F. (2018) Substrate Recognition and Autoinhibition in the Central Ribonuclease RNase E. *Mol Cell*, **72**, 275-285 e274.

64. Updegrove, T.B., Kouse, A.B., Bandyra, K.J. and Storz, G. (2019) Stem-loops direct precise processing of 3' UTR-derived small RNA MicL. *Nucleic Acids Res*, **47**, 1482-1492.

65. Schuck, A., Diwa, A. and Belasco, J.G. (2009) RNase E autoregulates its synthesis in Escherichia coli by binding directly to a stem-loop in the rne 5' untranslated region. *Mol Microbiol*, **72**, 470-478.

66. Kime, L., Clarke, J.E., Romero A, D., Grasby, J.A. and McDowall, K.J. (2014) Adjacent single-stranded regions mediate processing of tRNA precursors by RNase E direct entry. *Nucleic Acids Research*, **42**, 4577-4589.

67. Ali, N. and Gowrishankar, J. (2020) Cross-subunit catalysis and a new phenomenon of recessive resurrection in Escherichia coli RNase E. *Nucleic Acids Res*, **48**, 847-861.

68. Even, S., Pellegrini, O., Zig, L., Labas, V., Vinh, J., Bréchemmier-Baey, D. and Putzer, H. (2005) Ribonucleases J1 and J2: two novel endoribonucleases in B.subtilis with functional homology to E.coli RNase E. *Nucleic Acids Research*, **33**, 2141-2152.

69. Cavaiuolo, M., Chagneau, C., Laalami, S. and Putzer, H. (2020) Impact of RNase E and RNase J on Global mRNA Metabolism in the Cyanobacterium Synechocystis PCC6803. *Front Microbiol*, **11**, 1055.

70. Sala, C., Forti, F., Magnoni, F. and Ghisotti, D. (2008) The katG mRNA of Mycobacterium tuberculosis and Mycobacterium smegmatis is processed at its 5' end and is stabilized by both a polypurine sequence and translation initiation. *BMC Molecular Biology*, **9**, 33.

71. van Kessel, J.C. and Hatfull, G.F. (2007) Recombineering in Mycobacterium tuberculosis. *Nat Methods*, **4**, 147-152.

72. Shishkin, A.A., Giannoukos, G., Kucukural, A., Ciulla, D., Busby, M., Surka, C., Chen, J., Bhattacharyya, R.P., Rudy, R.F., Patel, M.M. *et al.* (2015) Simultaneous generation of many RNA-seq libraries in a single reaction. *Nature Methods*, **12**, 323-325.

73. Kapopoulou, A., Lew, J.M. and Cole, S.T. (2011) The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis (Edinb)*, **91**, 8-13.

74. Davis, J.J., Wattam, A.R., Aziz, R.K., Brettin, T., Butler, R., Butler, R.M., Chlenski, P., Conrad, N., Dickerman, A., Dietrich, E.M. *et al.* (2020) The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res*, **48**, D606-D612.

75. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.

76. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M. *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**.

77. Anders, S., Pyl, P.T. and Huber, W. (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166-169.

78. Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114-2120.

79. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **9**, 357-359.

80. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.

81. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

82. Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) ViennaRNA Package 2.0. *Algorithms for molecular biology : AMB*, **6**, 26.

83. Luciano, D.J., Vasilyev, N., Richards, J., Serganov, A. and Belasco, J.G. (2017) A Novel RNA Phosphorylation State Enables 5' End-Dependent Degradation in Escherichia coli. *Mol Cell*, **67**, 44-54 e46.

84. Kawano, M., Reynolds, A.A., Miranda-Ríos, J. and Storz, G. (2005) Detection of 5'- and 3'-UTR-derived small RNAs and cis-encoded antisense RNAs in Escherichia coli. *Nucleic Acids Research*, **33**, 1040-1050.

# Supplemental Figures

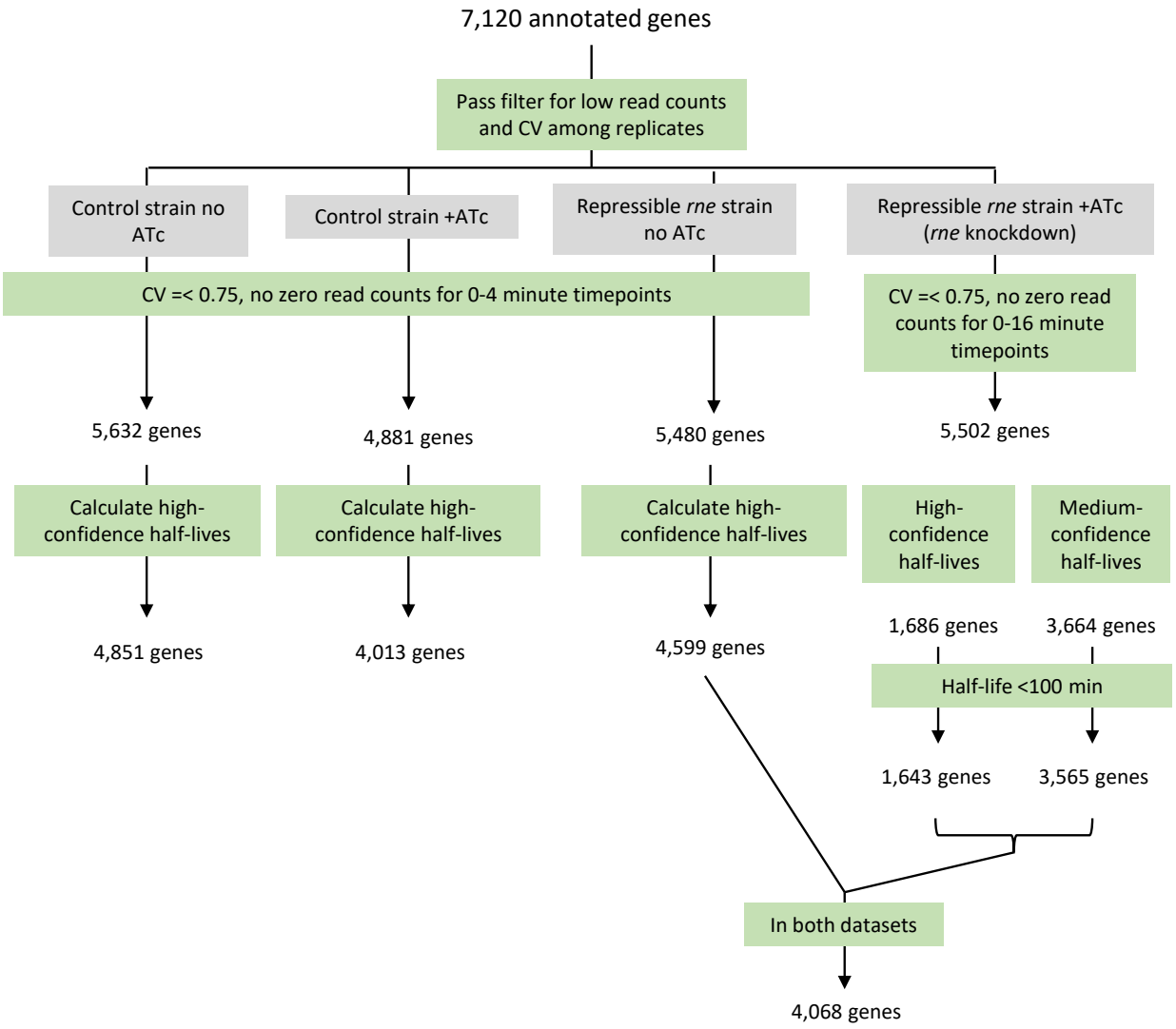



**Figure S 2-1. Overview of RNAseq data filtering for half-life calculations.**
Genes were used when they passed filters for read depth and CV among replicates. Half-life calculations are diagrammed in figures S2 and S3.

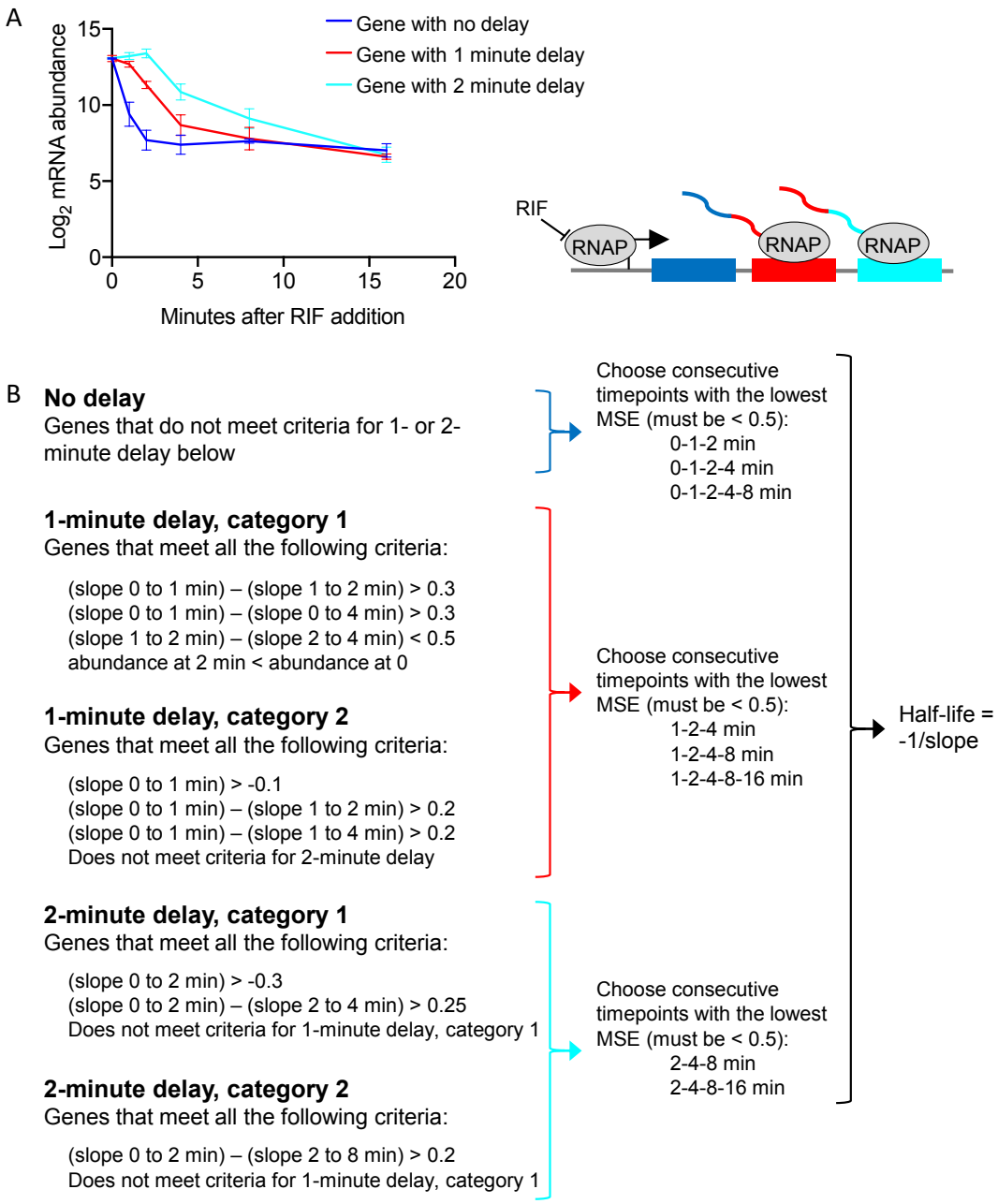



**Figure S 2-2. Half-life calculation procedure for genes in control conditions (*rne* not repressed).**

**A.** Log2-transformed mRNA abundance data show three distinct degradation patterns following addition of rifampicin to block transcription. Because rifampicin blocks transcription initiation but not transcription elongation, some genes show a delay before transcript levels decrease. The delay generally corresponds to the distance between the gene and its transcription start site. Furthermore, degradation for all genes reaches a plateau at later timepoints. We expect that the linear portion of the degradation curve between the delay (if present) and plateau is most likely to reflect the true degradation rate and therefore use this to calculate the half-life. **B.** Classification of genes into three delay categories based on linear regression fits to different sets of timepoints following addition of rifampicin, followed by half-life determination. The

slopes between the indicated timepoints were calculated and used to classify genes as having no delay, a 1-minute delay, or a 2-minute delay. After removal of early timepoints as indicated to account for the delay, the mean squared error (MSE) was used to quantify goodness of fit for linear regression using subsets of the remaining timepoints and the set of timepoints with the best fit were used to calculate the half-life.

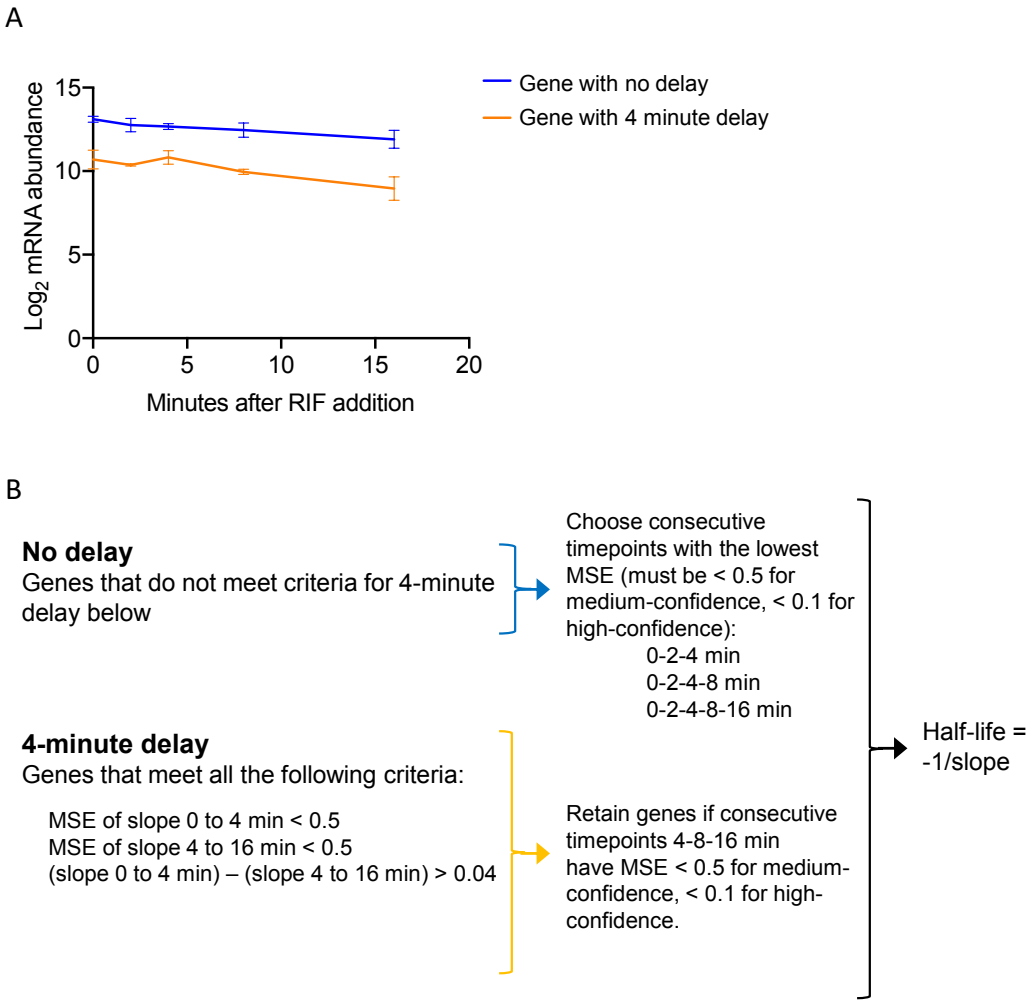



**Figure S 2-3. Half-life calculation procedure for genes in *rne* repression condition.**
**A.** Log2-transformed mRNA abundance data show two distinct degradation patterns following addition of rifampicin to block transcription, which can be best categorized as no delay or a 4-minute delay. **B.** Classification of genes into two delay categories based on linear regression fits to different sets of timepoints following addition of rifampicin, followed by half-life determination. The slopes between the indicated timepoints were calculated and used to classify genes as having no delay, or a 4-minute delay. After removal of early timepoints as indicated to account for the delay, the mean squared error (MSE) was used to quantify goodness of fit for linear regression using subsets of the remaining timepoints and the set of timepoints with the best fit were used to calculate the half-life.

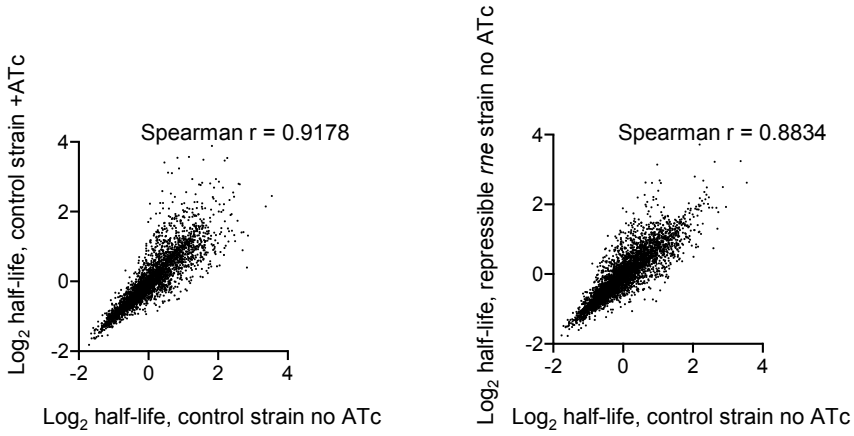



**Figure S 2-4. Correlations of half-lives between control conditions.**
Scatterplots show the half-lives calculated for genes in the control strain with and without ATc (left) and for the repressible and control strains without ATc (right).

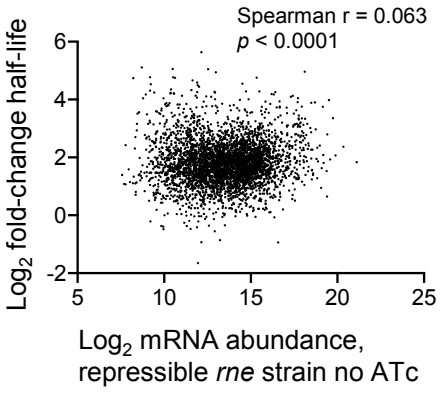



**Figure S 2-5. Fold-increase in half-life upon *rne* repression has a very weak correlation with abundance prior to repression.**
Each dot represents a gene for which half-lives were determined in the presence and absence of ATc in the repressible *rne* strain.

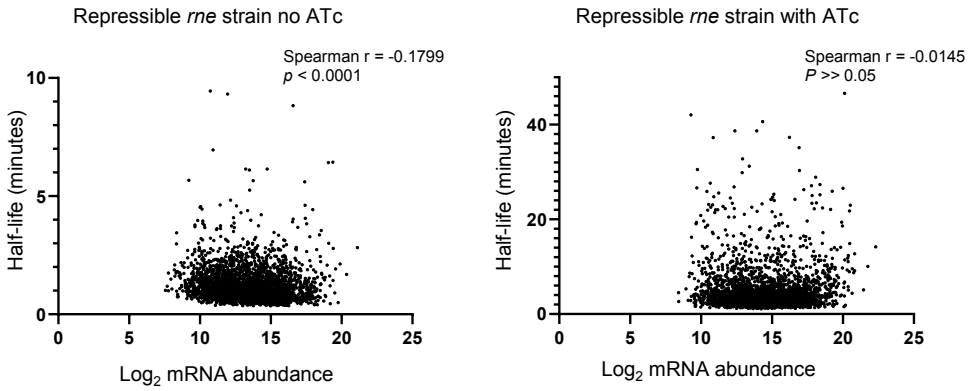



**Figure S 2-6. The relationship between mRNA abundance and mRNA half-life changes upon *rne* knockdown.**

Each dot represents a gene for which half-lives were determined in both the presence and absence of ATc in the repressible *rne* strain.

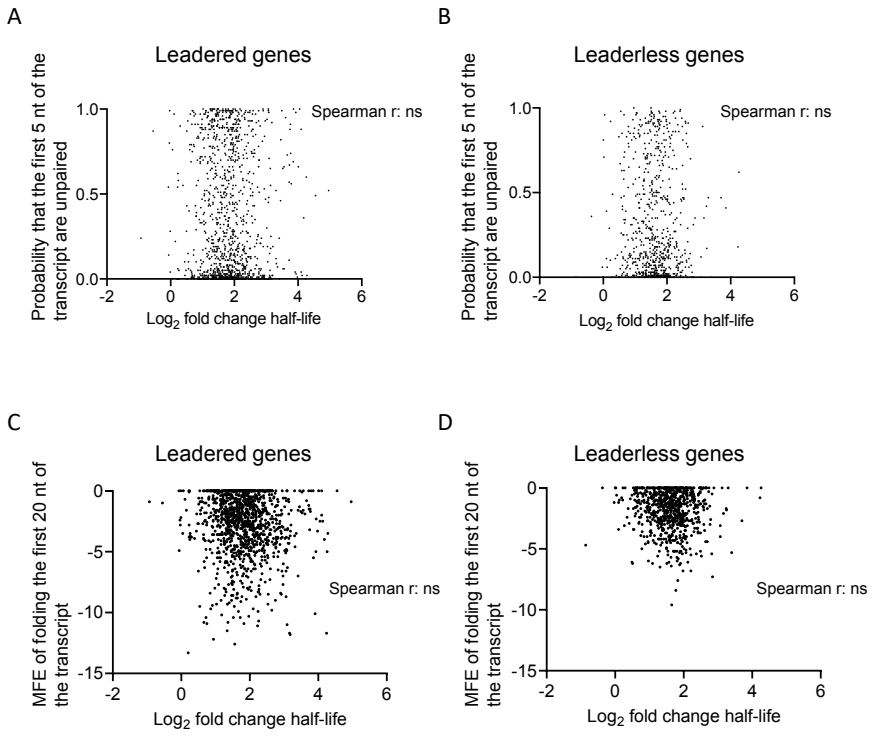



**Figure S 2-7. Predicted secondary structure near transcript 5′ ends is not correlated with degree of stabilization upon *rne* repression.**

Each dot represents a gene for which half-lives were determined in the presence and absence of ATc in the repressible *rne* strain. The MFE structure was predicted for the first 20 nt of each transcript (the 5′ 20 nt of the 5′ UTR for leadered transcripts, and the first 20 nt of the coding sequence for leaderless transcripts). **A**

and **B**, the probabilities of the first 5 nt of the transcript being unpaired given the given predicted MFE structures were determined. **C** and **D**, the MFE of folding was determined. All analyses were done in the Vienna RNAfold package (Lorenz
*et al.*, 2011). ns, $P > 0.05$.

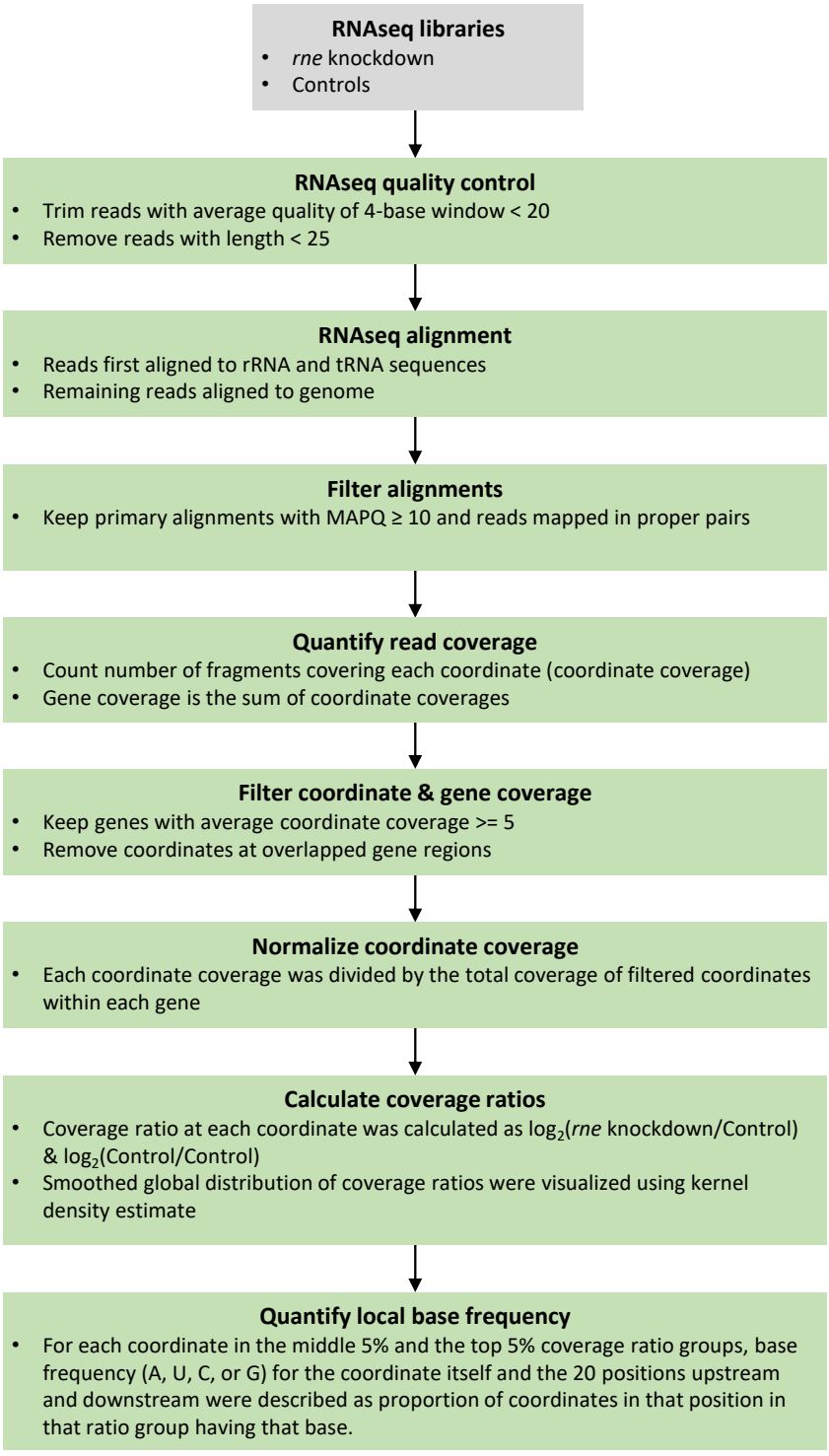



**Figure S 2-8. Pipeline for identifying RNase E cleavage sites from standard Illumina RNAseq expression libraries.**

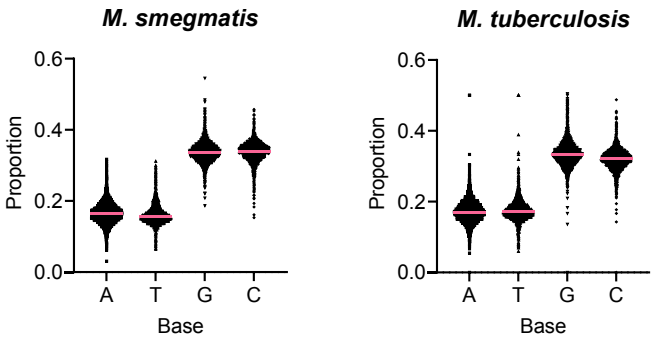



**Figure S 2-9. Base composition of coding sequences in *M. smegmatis* and *M. tuberculosis*.** Each dot represents a gene. Pink lines indicate medians.

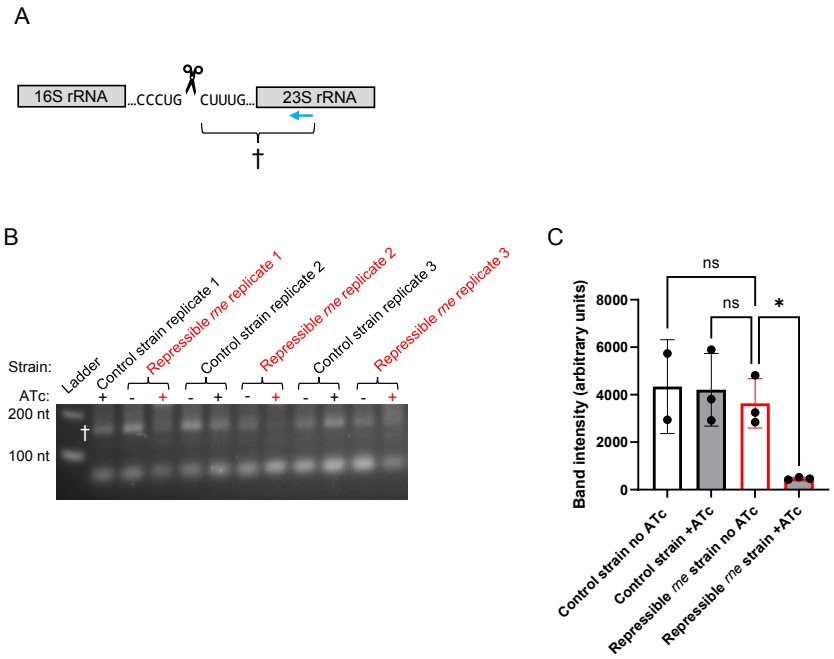



**Figure S 2-10. RNase E cleaves upstream of a cytidine during rRNA processing.**
**A.** Schematic of part of the rRNA operon with the sequence of a region reported to be cleaved by RNase E shown (Taverniti et al 2011). Graphic not to scale. The scissors indicate the exact cleavage position that we mapped by 5′ RACE, which is between positions 2041 and 2042 relative to the start of the rRNA operon (numbering as in Taverniti et al 2011). The cleavage site lies in a region predicted to be single-stranded, approximately 81 nt downstream of a predicted RNase III cleavage site and approximately 27 nt upstream of another predicted RNase III cleavage site (Taverniti et al 2011). The dagger (†) indicates the 5′ RACE PCR product shown in panel B. The blue arrow indicates the primer used for cDNA synthesis. **B.** An ethidium-bromide strained agarose gel revealing 5′ RACE PCR products. The dagger (†) indicates the PCR product shown schematically in panel A. Triplicate samples are indicated. Control strain replicate 1 in the absence of ATc was not run on this gel. The results are representative of two independent experiments. **C.** Image J was used to quantify the integrated pixel intensity of the band indicated with the dagger (†) in panel B.

Strains and conditions were compared by ANOVA and Dunnett's multiple comparisons test. * indicates $p < 0.05$.

5'-AGGGCGCUGAUCGCCAUGUUCCCCUGGUACAUCCAGUGGUUCCCCAACGC^CGUGUGGAAG-3'

5'-ACCUUCGA^CCUGUUCGUCGGCCUCAUCCAGGCCUUCAU^CUUCUCGCUGCUGACGAU^CC-3'

5'-UGUACUUCAGCCAGUCGAUGGAACUGGACCACGAGGACCACUGACGAGCAACCCUGCUGGA-3'
3'-AUGAAGUCGGUCAGCUACCUUGACCUGGUGCUCCUGGUGACUGCUCGUUGGGACGACCU-5'

5'-CCGAACAAAUCCCUACGACCCGAUCGACACGAACUCUGACGGCAACA-3'
3'-GGCUUGUUUAGGGAUGCUGGGCUAGCUGUGCUUGAGACUGCCGUUGU-5'

**Figure S 2-11. In vitro-transcribed partial duplex RNA substrate used for RNase E cleavage assays.**
Black font indicates the sense strand corresponding to the 3′ 159 nt of the *M. smegmatis atpB* coding sequence and 64 nt of the intergenic region between *atpB* and *atpE*. Blue font indicates an antisense strand used to block RNase E cleavage. Cleavage sites mapped in Fig. 5 are shown by red carets.

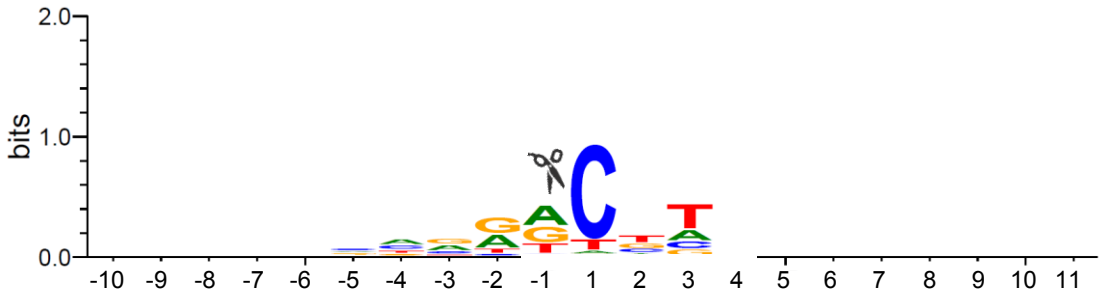



**Figure S 2-12. Sequence context of an expanded set of *M. tuberculosis* RNA cleavage sites.**
10,795 putative cleavage sites were identified using relaxed filters as described in the methods section. A weblogo (Weblogo 3.7.12) was constructed with a background frequency of 65% G+C.

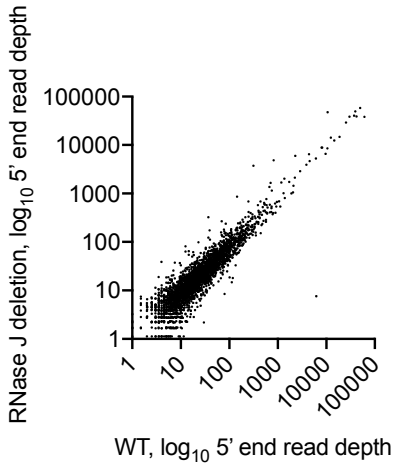



**Figure S 2-13. Most *M. tuberculosis* cleavage sites have similar abundance in WT H37Rv and an isogenic strain in which the gene encoding RNase J was deleted.**
Monophosphorylated RNA 5′ ends were mapped and quantified by adapter ligation and Illumina sequencing. Read depth for each 5′ end produced by the cleavage sites listed in Supplementary Table 5 is shown.

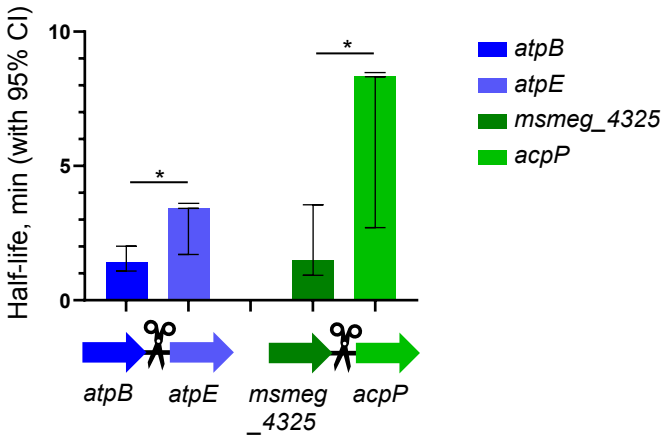



**Figure S 2-14. *M. smegmatis* gene pairs that appear to be co-transcribed and are bisected by cleavage sites display differential stabilities.**
Abundance of the four indicated transcripts was measured by qPCR 0, 4, and 8 minutes after addition of rifampicin to block transcription, and half-lives were calculated by linear regression of log2-transformed abundance. Error bars show the 95% confidence intervals of each half-life. The top error bar was truncated in two cases (*atpE* and *acpP*) where the upper 95% CI was infinity. *, $p < 0.05$ for comparison of the half-lives of the indicated genes by linear regression.

# Chapter 3 : Diverse intrinsic properties shape transcript stability and stabilization in *Mycolicibacterium smegmatis*

# Diverse intrinsic properties shape transcript stability and stabilization in *Mycolicibacterium smegmatis*

**Huaming Sun[1], Diego A. Vargas-Blanco[2], Ying Zhou[2], Catherine S. Masiello[2], Jessica M. Kelly[2], Justin K. Moy[1], Dmitry Korkin[1,*], and Scarlet S. Shell[2,*]**

[1] Program in Bioinformatics and Computational Biology, Worcester Polytechnic Institute, Worcester, Massachusetts, 01609, USA
[2] Department of Biology and Biotechnology, Worcester Polytechnic Institute, Worcester, Massachusetts, 01609, USA
* To whom correspondence should be addressed. Tel: +1 508 831 5917; Fax: +1 508 831 5936; Email: sshell@wpi.edu; dkorkin@wpi.edu.

This chapter corresponds to a manuscript that is under preparation.

## Abstract

In mycobacteria, regulation of transcript degradation is known to occur in response to environmental stress and facilitate adaptation. However, the underlying regulatory mechanisms are unknown. Here we sought to gain understanding of the mechanisms controlling mRNA stability by investigating the transcript properties associated with variance in transcript stability and stress-induced transcript stabilization. We performed transcriptome-wide mRNA degradation profiling of *Mycolicibacterium smegmatis* in both log phase growth and hypoxia-induced growth arrest. The transcriptome was globally stabilized in response to hypoxia, with all transcripts having longer half-lives but some having greater degrees of stabilization than others. The transcripts of essential genes were generally stabilized more than those of non-essential genes. We developed machine learning models that utilized a compendium of transcript properties and enabled us to identify the non-linear collective effect of diverse properties on transcript stability and stabilization. The comparisons of these properties confirmed the association of 5′ UTRs with transcript stability,

along with other differences between leadered and leaderless transcripts. Our analysis highlighted the protective effect of translation in log phase but not hypoxia-induced growth arrest. Steady-state transcript abundance had a weak negative association with transcript half-life that was stronger in hypoxia, while coding sequence length showed an unexpected correlation with half-life in hypoxia only. In summary, we found that transcript properties are differentially associated with transcript stability depending on both the transcript type and the growth condition. Our results revealed the complex interplay between transcript features and microenvironment that shape transcript stability in mycobacteria.

## Introduction

Regulation of mRNA degradation serves as a response mechanism of mycobacteria to energy-limited microenvironments. The mycobacteria include *Mycobacterium tuberculosis*, the causative agent of tuberculosis which led to over 1 million deaths in 2022 (1). Transcriptome-wide profiling of mRNA degradation in *M. tuberculosis* showed variance in transcript stability among genes and a global stabilization of the transcriptome in hypoxia, a stress condition that *M. tuberculosis* encounters within the centers of granulomas during infection (2-4). However, the regulatory mechanisms that govern transcript stability and stress-induced stabilization remain poorly understood. Further study of the regulation of mRNA degradation in mycobacteria is needed to facilitate our understanding of the stress response strategies that *M. tuberculosis* employs to adapt and persist within the host.

Global mapping of transcription start sites suggested that approximately 25% to 30% of RNA transcripts lack a 5′ untranslated region (5′ UTR) in mycobacteria (referred to as leaderless

transcripts) (5-7). Studies have shown that the presence of Shine–Dalgarno (SD) ribosome binding site sequences within the 5′ UTR is associated with higher mRNA expression levels in bacteria, measured by RNAseq expression data (5,7). However, it is also known that some leaderless transcripts have comparable translation efficiencies with leadered transcripts in *M. smegmatis* (6,8), potentially due to an alternative translation initiation mechanism and unique RNA characteristics, such as less structured start codon regions (9,10). In *M. tuberculosis*, such SD-independent translation of leaderless transcripts is also robust and less affected during adaptation to stress environments than canonical leadered transcripts (5,11). While it is likely that the 5′ ends of transcripts can contribute to the variability in mRNA half-life through either translation efficiency or degradation initiation (5,7,12), the mechanisms are not fully characterized.

Various transcript properties (features) have been reported to be associated with mRNA stability either transcriptome-wide or for individual transcripts in various organisms. Transcriptome-wide associations have been shown for growth rate in *L. lactis* and *E. coli* (13,14), transcript abundance in *E. coli* and *L. lactis* (15,16), GC content in *B. cereus*, *E. coli* and *S. cerevisiae* (15,17,18), 5′ UTR- and 3′ UTR-related features in *S. cerevisiae* (19), gene function and essentiality in *B. cereus* and *E. coli* (17,20), transcript length in *L. lactis*, *E. coli*, and *S. cerevisiae* (13,18,20), ribosome density in *S. cerevisiae* (18), and adjacent codon pair usage in *S. cerevisiae* (21). Transcript features like start codon identity and GC content in *S. cerevisiae* (18), 5′ UTR-related features in *L. lactis* (16), and transcript abundance in *E. coli* and *L. lactis* (16) were also validated experimentally with individual transcripts. In mycobacteria, the transcript features that impact mRNA degradation rates are largely unexplored, with existing analysis limited to only few transcript features and their individual broad correlations with transcript half-life in log-phase growing *M. tuberculosis* (2). It is unknown

which of the wide range of potentially associated features impact transcript stability, how the impacts of these features interact, and how they differ according to growth condition.

To model the underlying collective effect of multiple transcript features on stability, recent studies in *E. coli* and *S. cerevisiae* have used linear regression models to incorporate multiple features and quantify their contributions to variance in degradation rates (15,18,19). However, the limitation is that these models simplify the relationship between the features by assuming that they can be combined linearly to determine transcript half-life. Although more advanced sequence-based machine learning models could also be applied to predict stability, their performances rely on large amounts of data for training with the focus more on achieving accurate prediction rather than understanding the underlying mechanisms (22,23).

Given our lack of understanding of the impact of RNA features on mRNA degradation in mycobacteria, we sought to develop a comprehensive machine learning framework to identify the transcript properties that are associated with transcript half-life in the model organism *Mycolicibacterium* (nee *Mycobacterium*) *smegmatis* in both log phase growth and hypoxia-induced growth arrest. We found that in contrast to some previous reports on *M. tuberculosis* and *E. coli*, no single feature had a dominant association with mRNA half-life; rather, half-lives were best explained by the non-linear interactions of many features. The features that best explained transcript half-lives differed between log phase growth and hypoxia, and while the half-lives of most transcripts were longer in hypoxia, those of essential genes were lengthened the most. Features associated with efficient translation were generally predictive of longer half-lives in log phase but not in hypoxia, consistent with the idea that translation protects mRNA from degradation in rapidly growing cells and lower levels of translation in non-growing cells limit its

impact. mRNA secondary structure was also generally predictive of longer half-lives in cases where it did not negatively impact translation, but in ways that varied by condition and transcript leader type. 5′ UTR features were predictive of half-life in ways that appeared to extend beyond mediating translation initiation. Surprisingly, gene length was predictive of slower degradation in hypoxia, consistent with models in which diffusion of large molecules is slower in non-growing cells. Taken together, our results reveal the landscape of the collective effect of diverse transcript features on stability under different conditions in *M. smegmatis*.

# Material and Methods

## Strains and culture conditions to generate transcriptome-wide mRNA degradation datasets

Transcriptomic data were obtained from *M. smegmatis* strain SS-M_0424, a derivative of $mc^2$-155 described in (24), in which a $hyg^R$ gene was inserted upstream of, and divergent from, the *rne* gene promoter, and a $kan^R$-marked plasmid expressing tetR38 was integrated at the L5 site. This strain was constructed as a control for an *rne* knockdown strain, and its genetic modifications did not affect expression of *rne*. *M. smegmatis* was grown at 37° C with 200 rpm shaking in Middlebrook 7H9 broth supplemented with final concentrations of 0.2% glycerol, 0.05% Tween-80, 3 mg/L catalase, 2 g/L glucose, 5 g/L bovine serum albumen fraction V, and 0.85 g/L sodium chloride. RNA was extracted from cultures at defined time-points following addition of 150 µg/mL rifampicin to block transcription initiation. The log phase cultures are described in (24). Cultures for hypoxia were sealed in vials as described in (25). The volume of culture in each bottle was 13.5 mL and the OD at the time of sealing the bottles was 0.01. 19 hours after sealing the bottles,

rifampicin was injected through the rubber cap with a needle, and at the indicated timepoints (0, 3, 6, 9, 15, 30 and 60 minutes) bottles were opened, contents poured into 15 mL conical tubes, and the tubes submerged in liquid nitrogen. The elapsed time between opening the hypoxia bottles and submerging the cultures in liquid nitrogen was approximately 6 seconds. Frozen cultures were stored at -80° C. Cultures were thawed on ice and RNA extracted as in (25). The RNAs were submitted to the Broad Institute Microbial 'Omics Core where Illumina libraries were constructed as described in (26) and sequenced. The hypoxia cultures were grown and their RNA extracted and sequenced together with the log phase cultures described in (24). Separately, the RNA samples were used to synthesize cDNA and perform quantitative PCR was performed as described in (25) and the resulting data were used for normalization of the RNAseq data as described in (24).

### *M. smegmatis* genome sequence and gene annotations

The transcript features were quantified using the genome sequence of *M. smegmatis* strain $mc^2$-155 (NC_008596.1) from Mycobrowser Release 4 (27). The gene annotations were updated as previously described (24) and listed in Table S3-1. Using these annotations, we defined 1939 leadered transcripts (Table S3-2) and 960 leaderless transcripts (Table S3-3) with high confidence. For transcripts with multiple transcription start sites (TSSs), the leadered or leaderless status was determined by the TSS with the highest read coverage in log phase (7).

### RNAseq data processing and half-life calculations

RNAseq data were processed and normalized to produce transcript degradation profiles for log phase and hypoxia as previously described (24). Log phase half-lives were calculated as described

(24). To calculate half-lives in hypoxia with high-confidence, we only used genes for which linear regression of $\log_2$ transcript abundance between first 5 time points (0, 3, 6, 9, 15 mins) had a mean squared error (MSE) < 0.5 (24). Genes with zero read counts for any replicate for any of those 5 time points were also excluded. Half-life was calculated as -1/slope.

**UMAP visualization of transcript degradation profiles in log phase and hypoxia**

The $\log_2$ normalized RNAseq coverage of 7120 genes in 42 samples (3 replicates of each of the 7 time points in log phase and hypoxia) were used to visualize the transcript degradation profiles in *M. smegmatis*. UMAP plots were made in R v4.3.2 using package umap v0.2.10.0 with the parameters n_neighbors = 20, min_dist = 0.25, n_component = 2, random_state = 77. The seed value in R was set to be 7.

## Hierarchical clustering to identify transcript degradation patterns

Normalized transcript degradation profiles in log phase and hypoxia were collected as described above and in (24) for clustering analysis. To select genes with high quality degradation profiles, we conducted two preprocessing procedures. For each gene, we first calculated the coefficient of variation (CV) over three RNAseq replicates for each time point. Genes with CV > 0.75 in any of the first 4 time points (0, 1, 2, 4 minutes for log phase; 0, 3, 6, 9 minutes for hypoxia) were excluded from clustering. In order to cluster by the differences in degradation pattern rather than the absolute abundance, we then converted the profiles into relative abundance to an initial time point (0 and 1 minute for log phase; 0 minute for hypoxia). The CV-filtered genes were further selected by being required to have relative abundance at subsequent timepoints be no more than 1.5 times that of the initial time point. After preprocessing, the relative degradation profiles were

then clustered using hierarchical clustering with the Euclidean distance measure and ward.D2 agglomeration method. The degradation pattern of each cluster was represented by the mean and standard deviation of $\log_2$ mRNA abundance at each timepoint (Figure S2-2B-D and S3-2F-H). mRNA degradation is expected to follow a single exponential decay trend. For log phase, initial clustering produced a cluster of genes that exhibited a delay prior to the start of exponential decay, which is a well-established phenomenon due to rifampicin blocking transcription initiation but not elongation (28). To produce clusters unaffected by this technical issue, we removed the genes in the cluster showing the delay and re-clustered the remaining genes using degradation profiles normalized to the 1 minute timepoint rather than the 0 minute timepoint. This resulted in classification of 4972 genes into degradation pattern classes in log phase (Figure S3-2A). For hypoxia, no delays were observed, likely due to both transcription and mRNA degradation being substantially slower than in log phase, and pattern classes were directly defined for 5098 genes by the clustering of degradation profiles relative to the 0 minute timepoint (Figure S3-2E). Normalized transcript degradation profiles in log phase and hypoxia were collected as described above and in (24) for clustering analysis. To select genes with high quality degradation profiles, we conducted two preprocessing procedures. For each gene, we first calculated the coefficient of variation (CV) over three RNAseq replicates of each time point. Genes with CV larger than 0.75 in any of the first 4 time points (0, 1, 2, 4 minutes for log phase; 0, 3, 6, 9 minutes for hypoxia) were excluded for clustering. In order to run clustering by the difference of degradation pattern rather than the absolute abundance, we then converted the profiles into relative abundance to initial time point (0 and 1 minute for log phase; 0 minute for hypoxia). The CV filtered genes were further selected by being required to have the relative degradation smaller than 1.5 in all time points.

After the preprocessing, the relative degradation profiles were then clustered using hierarchical clustering with the Euclidean distance measure and ward.D2 agglomeration method. The degradation pattern of each cluster was represented by the mean and standard deviation of $\log_2$ degradation profile (Figure S3-2B-D and S3-2F-H). mRNA degradation is expected to follow a single exponential decay trend. For log phase, initial clustering produced a cluster of  genes that exhibited a delay prior to the start of exponential decay, which is a well-established phenomenon due to rifampicin blocking transcription initiation but not elongation (28). To produce clusters unaffected by this technical issue, we removed the genes in the cluster showing the delay and re-clustered the remaining genes using degradation profiles normalized to the 1 minute timepoint rather than the 0 minute timepoint. This resulted in classification of 4972 genes into degradation pattern classes in log phase (Figure S3-2A). For hypoxia, no delays were observed, likely due to both transcription and mRNA degradation being substantially slower than in log phase, and classes were directly defined for 5098 genes by the clustering of degradation profiles relative to the 0 minute timepoint (Figure S3-2E).

## Transcript property quantification

To identify the transcript properties that are associated with degradation, we quantified many potential candidate properties (Table S3-5) for each of the 7120 CDSs (Table S3-1).

*Nucleotide frequency.* This group of properties was quantified through the nucleotide frequency percentage relative to 5′ UTR or CDS region length. It contains usage of single nucleotides, adjacent dinucleotide motifs, and total G+C content. For each CDS, we quantified nt frequency for the 5′ 18 nt, the 3′ 18 nt, and the entire CDS as separate properties.

*Codon frequency*. In addition to the percentage of each nonstop codon calculated in the same manner as the nucleotide frequency for CDS, we also added binary indicators for the choice of start codon (AUG, GUG and UUG) and stop codon (UAA, UAG and UGA). To quantify the codon pair bias, we calculated the Codon Pair Bias (CPB) using the Codon Pair Score (CPS) of all codon pairs that make up the CDS (29). The CPS was calculated for each of the 3904 possible codon pairs (61 * 64), including stop codons only being used as the second codon to capture potential bias at the 3′ end of the CDS.

*Secondary structure*. These properties were quantified using the ViennaRNA v2.5.0 package (30) through the following three metrics: the ΔG of the minimum free energy (MFE) structure for a given transcript segment, the number of unpaired nucleotides at the 5′ end of the MFE structure, and the probability of specific nucleotides near the 5′ end being unpaired. The ΔGs of MFE structures were calculated using RNAfold v2.5.0 in two different ways. First, to overcome the positive correlation between ΔG and sequence length, we calculated ΔG of MFE ($\Delta G_{MFE}$) structures in a sliding window manner. Each sequence was split into subsequences by *M* nt windows, each with *M/2* nt overlap. The $\Delta G_{MFE}$ for a given sequence is the averaged $\Delta G_{MFE}$ of all its subsequences generated by a sliding window. For 5′ UTRs and CDSs, we divided each sequence into thirds and used such sliding window $\Delta G_{MFE}$s to quantify the predicted structure of the 5′ third, middle third, and 3′ third as well as the entire sequence. For the 5′ UTRs, we sought to distinguish between secondary structure directly affecting ribosome binding and other secondary structure. We therefore excluded the 3′-most 15 nt before dividing the sequence into thirds. Additionally, only 5′ UTR sequences longer than 35 nt before removing the 15 nt ribosome binding site were used, and only a 20 nt sequence window was used. For the CDS region, we calculated $\Delta G_{MFE}$s using 20,

50, and 100 nt windows. The 3′ UTRs were approximated as 60 nt after the stop codons. The MFE for 3′ UTRs were calculated using a 20 nt window only.

We also used the $\Delta G_{MFE}$ structures to measure the accessibility of the mRNA translation initiation region (TIR) for ribosome binding (10). We calculated $\Delta G_{unfold}$ separately for leadered transcripts that have 5′ UTRs at least 12 nt long (1809 transcripts) and leaderless transcripts (960 transcripts). To calculate $\Delta G_{unfold}$, $\Delta G_{mRNA}$ was first calculated using RNAfold v2.5.0 to represent the folded state of the mRNA TIR in the absence of ribosome binding. For leadered transcripts with 5′ UTRs at least 25 nt long, this region was defined as 25 nt upstream of the start codon and the first 25 nt of the coding sequence. For leadered transcripts with 5′ UTRs shorter than 25 nt and for the leaderless transcripts, this region was defined as 50 nt downstream of the transcription start sites (TSSs). Then to approximate the ribosome-bound state of the mRNA TIR, $\Delta G_{init}$, the TIR structure prediction was processed using RNAstructure v6.3 to break any base pairing within the ribosome footprint. The ribosome footprint was assumed to be 12 nt upstream of the start codon and the first 13 nt of the CDS for leadered transcripts, and the first 13 nt of the CDS for leaderless transcripts (31). Then $\Delta G_{unfold}$ was calculated as $\Delta G_{init} - \Delta G_{mRNA}$.

The number of unpaired nucleotides at each transcript 5′ end was predicted from MFE structures produced by RNAfold v2.5.0 when folding the entire CDS (leadered and leaderless genes), the 5′ UTR (leadered genes only), the 5′ UTR plus the first 18 nt of the CDS (leadered genes only), or the first 20 nt of the transcripts (leadered and leaderless genes). Separately, the probabilities of certain transcript regions being unpaired were predicted using RNAplfold v2.5.0. To assess the base-pairing status of the 5′ ends of transcripts in a different way, we folded the first 20 nt of each transcript and calculated for the first 3 nt and 5 nt (i) the probabilities that all the

nucleotides are unpaired or (ii) the averaged unpaired probability of each nucleotide. To predict accessibility of ribosome-binding regions in leadered transcripts, we folded the last 30 nt of the 5′ UTR plus either the first 20 nt of the CDS or the start codon only. We then quantified the probability of the entire start codon being unpaired as well as the averaged dinucleotide unpaired probability over either the entire folded sequence or the Shine-Dalgarno region (-6 to -14 relative to the start codon).

*Ribosome occupancy*. We used RNAseq data from GSE127827, which included libraries made from total rRNA-depleted RNA (referred to henceforth as mRNA libraries) and well as libraries made from ribosome footprints. After retrieving data using SRA Toolkit v3.0.0, we processed the sequencing data following the original methods with some modifications (32). First, quality control was performed using FastQC v0.11.9 (33). The ribosome footprint data were further processed using Trimmomatic v0.39 with the options ILLUMINACLIP:~/adaptors_SE.fa:2:30:10 SLIDINGWINDOW:4:20 MINLEN:25, which including removing adaptors, cutting reads when average quality per nucleotide was lower than 20 within a 4-nt sliding window, and discarding reads less than 25 nt long (34). This was not necessary for mRNA libraries due to their higher quality. Next, for both ribosome footprint and mRNA libraries, we performed alignment to discard reads aligned to tRNA and rRNA using Bowtie2 v2.4.5 with the option --very-sensitive (35). The remaining reads were then aligned to the genome sequence of *Mycolicibacterium smegmatis* strain $mc^2$-155 using Bowtie2 v2.4.5 with the option --sensitive-local. Reads and alignments were processed and sorted using SAMtools v1.16.1 (36). To further remove unmapped reads, PCR or optical duplicate reads, reads that are not primary alignments and alignments with MAPQ smaller than 10, we filtered the alignments using SAMtools v1.16.1 with the options -q 10 -F 1284. The

remaining alignments were then quantified in TPM for both ribosome footprint and mRNA libraries using StringTie v2.2.1 (37). Such quantification was done separately for four different transcript regions: (i) the entire CDS, (ii) the entire CDS plus the 20 nt upstream, (iii) the 5′ end of the transcript (the first 18 nt of the CDS for leaderless transcripts or the last 20 nt of the 5′ UTR plus the first 18 nt of the CDS for leadered transcripts), and (iv) the CDS excluding its first 18 nt. We also quantified the coverage for each third of every CDS (5′ third, middle third, 3′ third) to capture regional differences. Then for each of the two affinity tag swapped strains, we calculated the TPM ratios of ribosome footprint library coverage over mRNA library coverage using the averaged TPMs over two replicates. At the end, the normalized ribosome occupancy for each CDS in these transcript regions was calculated as the averaged TPM ratios over these two dual-RpsR-tagged strains.

*Shine-Dalgarno sequence*. In order to approximate the strength of the Shine-Dalgarno sequences of leadered genes, we quantified the GA percent and GA frequency within the region of -17 to -4 relative to the start codon, as well as the frequencies of 17 specific Shine-Dalgarno motif variants in the 25 nt upstream of the start codon.

*Other properties*. This group of properties includes the sequence length and steady-state transcript abundance (0 min RIF treatment; "initial abundance"). We quantified length for 5′ UTRs and CDSs based on the annotation of 7120 CDSs (Table S3-1-S3-3). CDS abundance was normalized by the CDS length. The initial abundances in log phase and hypoxia were used respectively for log phase and hypoxia model development.

## Feature selection procedure

Our complete feature set includes five different feature types (nucleotide frequency, codon frequency, secondary structure, ribosome occupancy, and others) in four transcript regions (5′ UTR, 5′ end of transcript, CDS, and 3′ UTR) (Figure 3-3A; Table S3-5). The design of this feature set was driven by our hypothesis that the transcript stability is controlled by the unknown combination of multiple transcript properties. However, the intersection of multiple feature types within and across transcript regions leads to high correlations among several features. Although those correlations might not directly affect machine learning model performance, the shared credit of correlated features contributing to the predictions could affect the importance rankings of features. Such correlations also complicate the interpretation of feature contributions.

We therefore sought a feature selection algorithm to minimize the influence of the correlated features without losing potentially important features. Commonly used selection techniques lack the capability to consider both the relationships among the features themselves and the relationships between the features and the predicted class (38). To perform feature selection in a manner suitable for our feature structure and goals, we developed the following algorithm. Our algorithm only targeted the highly correlated features ( | Spearman's $\rho$ |>= 0.6) that were of the same type and within the same transcript region. We also took into account correlations between the individual features' values and classes, measured by the Kendall rank correlation coefficient (39). For this process, we considered the 5′ UTR and 5′ end of the transcript to be the same transcript region. Our goal was to select features that have less correlations with other features, while potentially contributing the most to the model performance. The selection procedure was done separately for each of six models that used the combined feature set to predict the stability class: leadered genes and leaderless genes each in log phase, hypoxia, and fold change in hypoxia

relative to log phase. The algorithm returned a list of features to be used for the machine learning model training and evaluation (See Supplementary Materials).

## Machine learning classifier training

Given the limited number of leadered and leaderless transcripts and the imbalanced number of genes in the half-life classes, random forest emerged as an optimal choice given its fast training convergence and ability to avoid overfitting (40). To train and evaluate the classifiers, we implemented 5-fold nested cross-validation using the scikit-learn 1.2.1 package (41). Each dataset was split into 5 folds in a stratified manner for outer cross-validation. The same training and testing sets were used for random class prediction models and random forest classifiers at each fold iteration to compare their performances. To measure the performance of the classifiers given the numerically imbalanced yet equally important classes, we used the macro F1 score, *i.e.*, the unweighted mean of F1 scores for each class, for all the scoring metrics. See below formula for the F1 score and the macro F1 score.

$$F_1 = \frac{2 * TP}{2 * TP + FP + FN}$$

$$Macro\ F_1 = \frac{\sum_{i=1}^{n} F_{1\_i}}{n}$$

Where *TP* is the number of true positives, *FN* is the number of false negatives, and *FP* is the number of false positives. *n* is the number of classes.

In order to perform hyperparameter tuning for random forest classifiers, the training set was split in the same manner as before for inner cross-validation to do a randomized search on hyperparameter sets ***H*** (max_depth: [3, 5, 7], min_samples_leaf: [20, 30, 50], min_samples_split: [5, 10, 20]). The optimal hyperparameters set ***h**** selected by inner cross-validation was used for

training with the entire training set to obtain the optimal model ***RandomForestModel****, which was then evaluated using the outer testing set. To quantify the contributions of individual features, we used both the impurity-based Gini importance and the SHAP values of predictions made on outer testing sets. To reduce the bias of random sampling, the nested cross-validation was repeated 10 times with a different randomly split of training and testing sets each time. Ultimately, the outputs of the nested cross-validation include the F1 scores of all fold iterations for 10 repetitions, the averaged F1 scores of the 10 repetitions, averaged Gini importance scores of 10 repetitions, and SHAP values for each stability class across all fold iterations for 10 repetitions (See Supplementary Materials).

## Statistical comparison of machine learning models

For machine learning models developed using cross-validation, there are two potential issues in testing the statistical significance of model performance differences. First, the performance of classifiers could be driven by a specific split of training and testing sets. To better ensure that a potential significant difference between classifier F1 scores is not due to a random split, ideally the difference in F1 score should be calculated using the same fold iteration of data for each pair of compared classifiers. Second, in the case of statistically testing the significance of differences between two distributions of F1 scores or F1 score differences, the commonly used Student's t-test could provide misleading results in the context of cross-validation. The reason is that the resampled data in training and testing makes the F1 scores, and thereby the F1 score differences, dependent across iterations. This violates the independence assumption in the Student's t-test, and could lead to high Type I error due to the underestimated variance of difference (42). To address these problems, Nadeau and Bengio proposed a corrected paired t-test, which can take

into account the dependency in samples and reduce the number of fold positive errors (43). In our case, the classifiers were trained and evaluated using 5-fold nested cross-validation. The entire procedure was repeated 10 times to get averaged performance. Each time, the model was trained and tested using the subsampled training and testing sets that were overlapped in different fold iterations. We implemented Nadeau and Bengio's corrected paired t-test to evaluate the differences between our random forest classifiers and random estimators that predict class membership randomly without using any of the features. Each classifier and a random estimator are trained and tested using the same fold of the dataset through the cross-validation. At the end, 50 paired F1 scores (5 folds * 10 repetitions) of these two classifiers were collected to test for the significance of difference. For the classifiers that were trained separately for the genes, conditions or features being compared, we were not able to train and evaluate their performances with the same dataset. We therefore calculated $\Delta$F1 score as the differences between F1 scores from the random forest classifier and the random estimator for each comparison of interest and used the Wilcoxon rank-sum test to compare the $\Delta$F1 scores from 10 repetitions between conditions, gene types, and features.

## Essential gene enrichment analysis

Essentiality of 6642 *M. smegmatis* genes were defined using the CRISPR interference system (44). To statistically test the enrichment of essential genes in each stability class, only genes with essentiality designations and half-lives calculated in both log phase and hypoxia were used. This resulted in 3680 genes, of which 1327 were classified as leadered, and 793 were classified as leaderless. These genes were tested for essentiality enrichment in half-life classes using a hypergeometric test with FDR correction for multiple hypothesis testing (Figure 3-2G).

# Results

## Overview of an experimental and computational framework to unravel the intrinsic transcript properties that impact transcript stability in *M. smegmatis*

Bacterial mRNA half-lives are known to vary among transcripts and between conditions. To identify the transcript properties that contribute to the variance in transcript stability within and across conditions in *M. smegmatis*, we developed an experimental and computational framework consisting of the following four stages (Figure 3-1). We will summarize the stages here and describe them in greater detail in subsequent sections. First, we used RNAseq to quantify transcript half-lives transcriptome-wide. To characterize the impact of microenvironment on transcript stability, transcript degradation profiles were determined in log phase growth and hypoxia-induced growth arrest. We calculated transcript half-lives by linear regression of $\log_2$ transcript abundance over time for each condition. High-confidence half-lives were determined for 4,857 genes in log phase and 4,864 genes in hypoxia (Figure 3-2B). The log phase half-lives were published previously in (24). Second, transcripts were classified into quartiles based on half-life in log phase or hypoxia, or by fold-change in half-life in hypoxia compared to log phase. Third, we compiled transcript properties (features) that we hypothesized could affect half-life and developed random forest classifiers to identify properties predictive of half-life class membership. This was done separately for leadered and leaderless genes given differences in their features and the idea that their half-life determinants might differ. Fourth, the values of features identified as important were plotted by half-life class to provide an overview of the association between transcript properties and classes. We also implemented SHAP (SHapley Additive exPlanations) during classifier development to further explore the impact of features on each class (45). Together,

our pipeline reveals a comprehensive landscape of transcript half-lives and the transcript features influencing these half-lives in *M. smegmatis* in commonly studied rapid-growth and growth-arrested conditions.

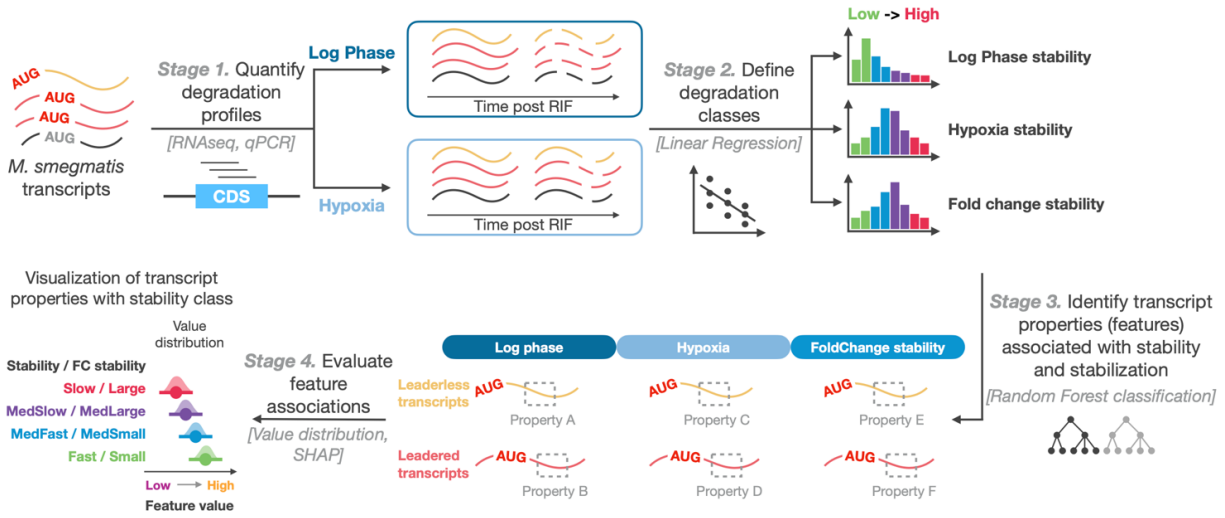



**Figure 3-1. Schematic of the framework to identify transcript properties that impact transcript stability in *M. smegmatis*.**
The framework was designed to reveal the transcript properties that were differentially associated with transcript stability depending on the transcript type and condition. Stage 1: Transcriptome-wide mRNA degradation profiles were collected in log phase and hypoxia using RNAseq followed by transcript half-life calculation. Stage 2: In each condition, transcripts were classified into four groups according to their half-lives. Stages 3 and 4: A series of random forest classifiers were trained to classify transcripts into their assigned half-life class based on the values of a set of transcript properties (features), and identify the features important for these classifications.

## Transcript degradation profiles capture variance in transcript stability both within and between growth conditions

To obtain transcriptome-wide mRNA degradation profiles in *M. smegmatis*, we inhibited transcription initiation with rifampicin (RIF) followed by RNA extraction at various time-points. Hypoxia was produced by a variation of the Wayne model in which cultures were sealed with a defined volume of headspace and incubated with shaking for 19 hours (25). RIF was injected through rubber caps with a needle to minimize introduction of oxygen, and bottles were sacrificed

at each time-point. Transcript abundance was quantified by RNAseq for samples harvested after 0, 1, 2, 4, 8, 16, and 32 minutes of RIF exposure in log phase and 0, 3, 6, 9, 15, 30, and 60 minutes of RIF exposure in hypoxia. Transcript abundance was normalized by relative abundance values determined for a set of genes by qPCR (24). A two-dimensional overview of the degradation profiles obtained by UMAP revealed a global difference between log phase and hypoxia (Figure 3-2A) (46), consistent with expectations from previous work indicating the transcript half-lives are longer in mycobacteria in response to hypoxia (2,25). Samples also clustered by time-point after addition of RIF, corresponding with the temporal changes in transcript abundance and indicating that our method successfully captured the global degradation trends in both conditions.

To further describe the transcript degradation process, we then used linear regression models to calculate transcript half-lives from the degradation profiles (Figure 3-2B; Table S3-4). The time-points used for half-life determination were carefully chosen to avoid confounding from continued elongation by RNA polymerase following addition of RIF as well as decreases in degradation rate that appear to be induced by RIF over time (24). As expected, there were wide ranges of half-lives in each condition. The half-life measurements also confirmed the expected global difference between log phase and hypoxia, with stabilization of all transcripts evident in hypoxia (Figure 3-2B). These findings are consistent with a previous assessment of global transcript stability in hypoxia-exposed *M. tuberculosis* (2), and our previous work showing stabilization of several transcripts in hypoxia-exposed *M. smegmatis* (25). However, this is the first transcriptome-wide report of mRNA half-lives in any hypoxia-exposed mycobacterial species. The observed global variance in transcript stability and transcript stabilization in response to hypoxia was maintained when we examined the transcripts of subsets of genes with defined transcription

start sites (TSSs) (Figure S3-1A, C). These subsets were composed only of genes that were monocistronic or the first in a polycistron, according to the annotations in (7) (Materials and Methods), and classified as leadered (having a 5′ UTR of 5 nt or more) or leaderless (lacking a 5′ UTR). Genes that were second or beyond in a polycistron or lacked annotated TSSs were excluded. For genes with multiple TSSs, we used the TSS with the highest read coverage in log phase to define the 5′ UTR or lack thereof (7). Direct comparison of leadered and leaderless transcripts showed a statistically significant yet biologically limited difference in half-lives, with more leadered transcripts having longer half-lives in both log phase and hypoxia (Figure S3-1E-H).

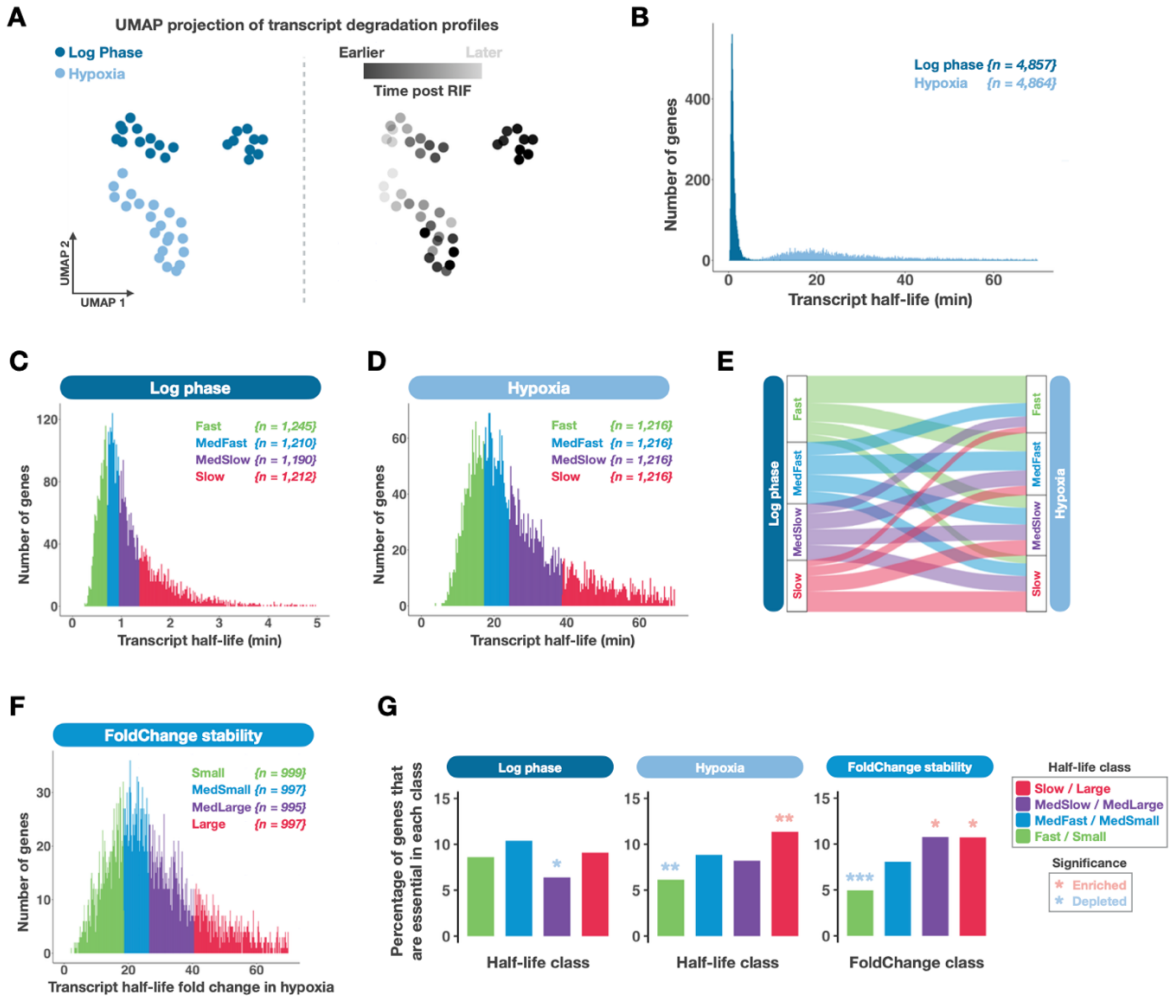



**Figure 3-2. Transcriptome-wide mRNA degradation profiles in *M. smegmatis*.**

**A.** UMAP projection showing condition differences and temporal changes in global degradation profiles (Materials and Methods). Each dot represents an RNAseq library from which normalized mRNA abundance values for each gene were obtained. The same data are shown in the two UMAP panels, colored according to condition (left, two conditions) or time after adding RIF (right, time points). **B.** Distributions of transcript half-lives in log phase and hypoxia. **C-D.** Half-life distributions with classes defined by half-life quartiles in log phase and hypoxia. **E.** Comparison of half-life class membership between log phase and hypoxia. **F.** Distribution of half-life fold changes in stabilization with classes defined by fold change quartile. **G.** Frequency of essential genes in each half-life class. Significance of enrichment and depletion of essential genes within each class were tested using a hypergeometric test with FDR correction (Materials and Methods). *p.adjust* < 0.05 *, *p.adjust* < 0.01 **, *p.adjust* < 0.001 ***.

To facilitate construction of machine learning models to identify transcript features affecting half-life, we grouped transcripts into classes based on half-life. Since the classes were split from a continuous range of half-lives, in theory one could define any number of classes. To select the number of classes that most accurately represented the degradation landscape, we first performed hierarchical clustering of the degradation profiles (Materials and Methods). The clustering produced four major classes with distinct degradation patterns (Figure S3-2A, E). We therefore decided to create four half-life classes. We chose to define classes based on half-life quartiles rather than by clustering of the complete degradation profiles to avoid confounding from continued elongation of RNA polymerase after addition of RIF as well as RIF-induced stress responses (Figure 3-2C, D). Nonetheless, the classes defined by half-lives had very similar gene composition to the clusters defined by hierarchical clustering (Figure S3-3A-F) and similarly separated genes according to transcript degradation rate (Figure S3-2).

When comparing the gene sets in each half-life class in log phase and hypoxia, we found that many genes switched classes in the two conditions (Figure 3-2E). This was true for both leadered and leaderless transcripts (Figure S3-1B, D). This suggested that the relationship between transcript features and half-life differs in different conditions. To facilitate later identification of

those features, we additionally classified genes according to the extent of stabilization in hypoxia vs log phase (defined by fold-change in half-life, Figure 3-2F).

Interestingly, we found that in hypoxia, genes classified as essential by CRISPR interference (44) were significantly enriched in the slowest degradation class while significantly depleted in the fastest degradation class (Figure 3-2G). Consistent with this, genes with a larger fold-change in stability in response to hypoxia were more likely to be essential than those with a smaller fold-change (Figure 3-2G). However, there was no consistent relationship between essentiality and half-life in log phase. This result supports the idea that global transcript stabilization in response to hypoxia is likely a regulatory mechanism as well as an energy-saving mechanism in mycobacteria. The significant stabilization of essential genes in hypoxia was only observed for leadered genes and not for leaderless genes, which suggests the possibility of different regulatory mechanisms for those two types of transcripts (Figure S3-4A, B).

## Nonlinear combinations of transcript properties (features) appear to specify half-life

We sought to identify the transcript properties that specify transcript half-life in *M. smegmatis*. To address this question as agnostically as possible, we compiled and quantified hundreds of properties, which we refer to as features. These included nucleotide and sequence features, predicted secondary structure features, and other features such as length, steady-state abundance, and ribosome occupancy from a published dataset (32). We categorized the features by type as well as by the gene region under consideration (Figure 3-3A), because we expected that some features would have different impacts depending on their location; for example, A/U-rich codons are expected to promote translation when located near the start codon due to their impact on

secondary structure (47,48), but be translated less efficiently when located elsewhere in a coding sequence due to being less preferred codons in mycobacteria.

Random forest classifiers were then trained separately for each gene type (leadered and leaderless) in each condition (log phase, hypoxia, and fold change in hypoxia relative to log phase) (Figure 3-3B). The classifiers were trained using 5-fold nested cross-validation and evaluated by the difference in F1 score compared to random prediction models (ΔF-score; see Materials and Methods). We trained classifiers using combined feature sets as well as using only features of each of the seven types (5′ UTR, CDS nucleotide, CDS secondary structure, codon, translation, and others) in order to evaluate the contribution of each feature type (Figure 3-3B). For the combined feature sets, we used a customized feature selection procedure to reduce the number of correlated features (Materials and Methods). As we predicted, classifiers that used the combined feature sets achieved the best performances, suggesting that the stability of transcripts is specified by the combination of various types of transcript properties. The ΔF-scores were low compared to those typically reported for random forest classifiers designed to distinguish between distinct clinical or physiological states (*e.g.*, diseased tissue vs healthy tissue), but were consistent with expectations for our data type, in which classes were made from continuous distributions of half-life values. A majority of the classifiers performed significantly better than random, and were strong enough to facilitate our overarching goal of identifying the features that impact half-life. Interestingly, most of the feature types could individually predict transcript stability with performance that varied depending on transcript type and condition (Figure 3-3B).


**Figure 3-3. Non-linear combinations of diverse transcript properties specify half-life in *M. smegmatis*.**

**A.** Summary of transcript features used for random forest classifiers. The features were grouped into six different types and quantified for specific transcript regions. Numbers in square brackets indicate the number of features of each type selected by our feature selection process. Numbers in the shaded regions indicate the number of features of each type in each transcript region. Asterisks indicate cases where the

total number of unique features is less than the sum of the numbers above because some features are classified as 5′ UTR type features for leadered transcripts and translation type features for leaderless transcripts (see Table S3-5). **B.** Comparisons of classifier performance to random prediction models. Random forest classifiers were trained separately for leadered and leaderless transcripts to predict stability in three conditions using various feature sets. The combined feature sets were selected by the log phase model for each transcript type (see Materials and Methods) and were used to train models in all three conditions. The 5′ UTR feature set includes both translation-related and non-translation-related features. The translation feature set includes log phase ribosome profiling. ΔF-score represents the difference in averaged F-score between random forest classifiers and random prediction estimators. Dots and bars represent mean and standard deviation of ΔF-scores for 10 repetitions of each model. The significance of the performance differences between random forest classifiers and random prediction estimators was tested using Nadeau and Bengio's corrected paired t-test (Materials and Methods). $p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***. **C-E.** Comparisons of ΔF-scores between leadered and leaderless transcript models in log phase, hypoxia, and fold change in hypoxia relative to log phase. For each condition, the combined feature sets were selected by the leaderless model and were used to train models of both transcript types. **F-G.** Comparisons of ΔF-scores between log phase and hypoxia models for leadered and leaderless transcripts. For each transcript type, the combined feature sets were selected by the log phase model and were used to train models of both log phase and hypoxia. The translation feature set excludes log phase ribosome profiling. The significance of the differences in model performance in **C-G** were tested using the Wilcoxon rank-sum test (Materials and Methods). For all panels, $p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***.

Our results confirmed the association of 5′ UTR features with transcript stability as suggested in multiple studies (12,15,19,49). We also found that translation-related features were more important in log phase than in hypoxia for both transcript types, which is further explored below. Notably, ΔF-scores resulting from the combined feature sets were far less than the sum of the ΔF-scores from the individual feature types, indicating that the collective effect was not a result of linearly accumulated contributions of each feature type. This is consistent with the idea that transcript features interact in a non-linear fashion with respect to their impact on transcript half-life. Additionally, and in contrast to some previous reports (2,15), we found that no individual feature or feature type appeared to be a dominant determinant of half-life. Rather, our results indicate that the underpinnings of mRNA stability in *M. smegmatis* are complex, arising from non-linear combinations of diverse transcript properties.

## The importance of secondary structure and translation in predicting mRNA half-life varies by transcript type and condition

To determine if some feature types were differentially important depending upon transcript type (leadered or leaderless) or condition, we further compared the performance of each classifier separately for each condition and each transcript type (Figure 3-3C-G). To rigorously compare ΔF-scores, the same feature set should be used in the models being compared. However, there were cases where the features differed between models, such as the absence of 5′ UTR features in the leaderless gene models. We therefore compared the leadered and leaderless models to each other using classifiers trained with only the features that were present in both (Figure 3-3C-E).

When considering the feature types separately, we found that CDS secondary structure features (as measured by the ΔG of minimum free energy (MFE) structures, $\Delta G_{MFE}$) were significantly more important for leaderless transcripts than for leadered transcripts in log phase (Figure 3-3C), which was exactly the opposite of the situation in hypoxia (Figure 3-3D). We also made direct comparisons between conditions separately for leadered and leaderless genes in order to use all available features for each transcript type. The trend with CDS secondary structure features was also observed in these comparisons, with these features being more important in hypoxia than log phase for the leadered genes but more important in log phase than in hypoxia for the leaderless genes (Figure 3-3F, G). These results indicate that secondary structure differentially contributes to the stability of leadered and leaderless transcripts in different conditions.

A different pattern was seen for codon features, which were more important for leadered genes than for leaderless genes in log phase only, and more important in log phase than in hypoxia for leadered genes only (Figure 3-3C, F). The impact of codon content on half-life is likely related at

least in part to translation having a greater influence on half-life in log phase, as observed for both transcript types in Figure 3-3B. However, comparisons between conditions of the impact of translation-related features were complicated by the inclusion of ribosome profiling data, which was performed only in log phase. We therefore trained classifiers excluding ribosome profiling features (Figure 3-3F, G) and directly compared their performance in log phase compared to hypoxia for each gene type. We still observed better performance of translation features in log phase than in hypoxia for both transcript types. These results suggest that translation has a larger impact on transcript half-life in log phase than in hypoxia regardless of the transcript type.

We hypothesized that translation influenced transcript half-life more in log phase because in that condition most mRNAs were being translated at rates that varied according to transcript properties, while in hypoxia most transcripts were not being actively translated. For technical reasons, we tested this experimentally using carbon starvation rather than hypoxia. In previous work, we found that carbon starvation induced transcript stabilization similar to that seen in hypoxia (25). Here we performed polysome profiling and found that indeed, while monosomes and polysomes were readily detected in log phase cells, they had much lower abundance relative to ribosomal subunits in carbon-starved cells (Figure S3-5). We furthermore collected fractions from the polysome profiling gradients and used qPCR to compare the relative abundance of four arbitrarily selected mRNAs in various fractions. For all four transcripts, the amount of transcript associated with monosomes and polysomes compared to unbound transcript decreased in carbon starvation compared to log phase (Figure S3-5). These results are consistent with the idea that in non-growing cells, a larger portion of transcripts are unassociated with ribosomes compared to in actively growing cells.

In order to compare the collective effect of all features between leadered and leaderless transcripts, we trained classifiers for both using the same set of features selected for leaderless models in each condition (Figure 3-3C-E). The combined features were better able to predict stability for leadered transcripts than for leaderless transcripts in log phase, and vice versa in hypoxia (Figure 3-3C, D). Similarly, we compared the collective effect of all features between log phase and hypoxia using the same sets of features selected for log phase models for each transcript type (Figure 3-3F, G). The result showed better performance in log phase than hypoxia for leadered transcripts and the opposite for leaderless transcripts, consistent with the results of the direct transcript type comparisons. For the classifiers predicting the extent of stabilization in response to hypoxia, we observed no significant difference between leadered and leaderless transcripts for the majority of the shared feature types except CDS nucleotide and codon content features (Figure 3-3E). However, the leadered/leaderless comparison by necessity excluded 5′ UTR features, and we noted that the 5′ UTR features were the feature type that best predicted fold change stability for leadered genes (Figure 3-3B). This contrasted with the individual log phase and hypoxia classifiers where the 5′ UTR feature group was relatively weak (Figure 3-3B). Overall, our results indicate that the specific ways that various properties contribute to transcript stability are tied to the leader status as well as the condition.

## Identification of specific features differentially predictive of half-life for leadered and leaderless transcripts

In order to identify the transcript features that were differentially important for classification of leadered vs. leaderless transcripts, we evaluated the Gini importance rankings of the same set of features, selected by leaderless models, that were used to train both leadered and leaderless

models. For each condition, we combined the top 20 most important features identified in the leadered and leaderless models and compared the relative importance levels of these features for the two gene types (Figure 3-4A, Figure S3-6A, B). We found that the most important features included features from each of the feature types in all three conditions, which further confirmed the collective effect of many features on dictating transcript stability. Furthermore, these comparisons also highlighted the different importance levels of many of the features between transcript types.

To determine the specific relationships between features of interest and half-life, we plotted the feature value distributions for each stability class (Figure 3-4B. SHAP distributions at Figure S3-10). This allowed us to better understand why these features were important for model predictions and, more interestingly, how they were associated with transcript stability. Consistent with our finding that codon frequencies were more important for leadered than leaderless transcripts in log phase, we observed a number of specific codons with higher importance levels for leadered compared to leaderless transcripts. Among them, CGC (Arg), CGG (Arg) and UUG (Leu) are examples of codons with higher importance for leadered transcripts. Their distributions exhibited inverse correlations with stability for both leadered and leaderless transcripts, suggesting that they may negatively impact transcript stability (Figure 3-4B). However, these inverse relationships were stronger for leadered transcripts than for leaderless transcripts, which may explain the differences in importance for the classifiers (Figure 3-4B). In contrast, another Arg codon, CGU, was more important for leaderless transcripts compared to leadered transcripts and had a more complex relationship with half-life class (Figure 3-4B).

For leaderless transcripts, both the frequency of CG dinucleotide motifs and extent of CDS secondary structure were positively correlated with stability (Figure 3-4B). These trends were weaker for leadered genes. These results support the conclusion that in log phase, CDS secondary structure plays a more important role for leaderless transcripts compared to leadered transcripts.

## 5′ UTRs appear to influence transcript half-life through both translation-related and translation-independent effects

The differences in stability determinants between leadered and leaderless transcripts were not limited to these shared features. Although we showed that the 5′ UTR itself was capable of predicting transcript stability (Figure 3-3B, Figure S3-6C), the mechanisms by which it impacts stability were unclear. To further explore this, we categorized 5′ UTR features as translation-related (e.g., Shine-Dalgarno sequence and predicted secondary structure in the ribosome binding region) and non-translation-related (e.g., nucleotide content and predicted secondary structure outside of the ribosome binding region) and trained models separately using these two feature groups (Figure 3-4C, Figure S3-6D). Surprisingly, our results suggest that the non-translation-related features have a larger impact on transcript stability than the translation-related features in both log phase and hypoxia. However, the model performance of translation-related features was significantly better in log phase compared to hypoxia, which is consistent with our finding that translation is more important for predicting transcript stability in log phase. Among all the translation-related features in 5′ UTR, the secondary structure seemed to be more important than the Shine-Dalgarno sequence (Figure 3-4D), which is the opposite of what was previously reported in *E. coli* (15). Such a difference could be because of the GC-richness of mycobacteria, which may

cause secondary structure to have a bigger impact on ribosome access compared to less GC-rich species.

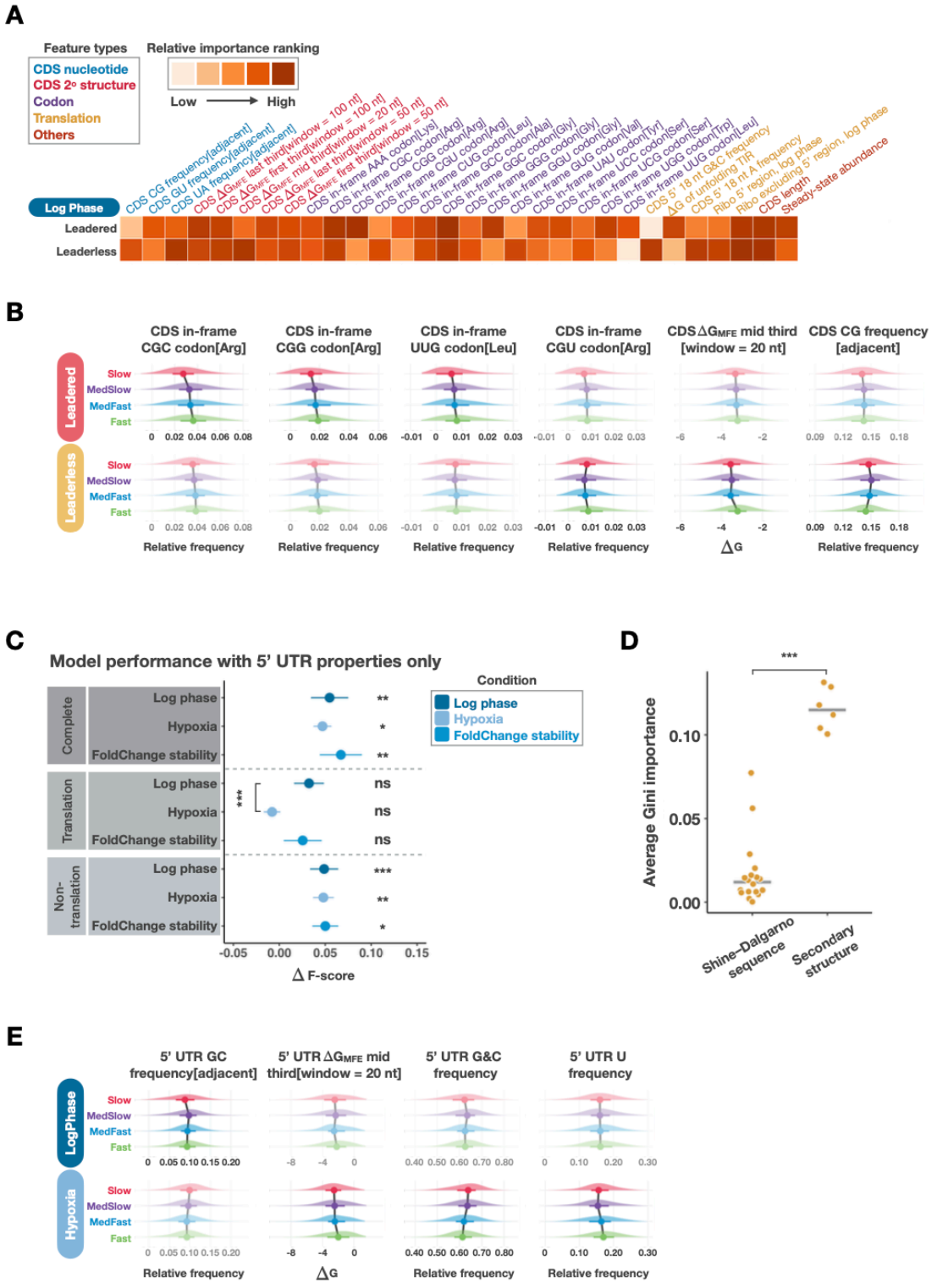


**Figure 3-4. Transcript features differentially predict half-life for leadered and leaderless transcripts in log phase.**

**A.** Summary of the most important features for the leadered and leaderless half-life class prediction models in log phase. Random forest classifiers were trained using the same set of features, selected by the leaderless model, for leadered and leaderless transcripts. The 20 features with the highest Gini importance rankings for each model were then combined and their relative importance rankings indicated by intensity of coloration in the heatmap. See Table S3-5 for feature definitions and details. **B.** Feature value distributions within each half-life class for selected features that differentially predicted half-life class for leadered and leaderless transcripts. Dimmed plots indicate that the feature was less important for that gene type. Dots and bars represent median and interquartile range. **C.** Comparisons of leadered gene models using only 5′ UTR features in three conditions. Models were trained and compared using the complete set of 5′ UTR features, translation-related 5′ UTR features only, or non-translation-related features only. See Table S3-5 for the specific features in each category. The performance differences between random forest classifiers and random prediction estimators were tested using Nadeau and Bengio's corrected paired t-test (Materials and Methods). Additionally, the log phase and hypoxia models using translation-related features were compared to each other with the Wilcoxon rank-sum test (Materials and Methods). **D.** Comparison of the importance of Shine-Dalgarno sequence features and the secondary structure features in the ribosome binding region of 5′ UTR in the log phase model for leadered transcripts. Each dot is the average Gini importance value from 10 repetitions of the model. The difference in Gini importance was tested using the Wilcoxon rank-sum test. **E.** Examples of 5′ UTR features differentially predicted half-life class between log phase and hypoxia. For all panels, $p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***.

Consistent with our finding that CDS secondary structure was more important in hypoxia for leadered transcripts, several 5′ UTR features associated with secondary structures were more predictive of transcript half-life in hypoxia. The overall 5′ UTR G+C frequency was positively correlated with stability, while 5′ UTR $\Delta G_{MFE}$ and U nucleotide frequency were negatively correlated with stability (Figure 3-4E, SHAP distributions at Figure S3-10). The frequency of the GC dinucleotide motif showed a similar trend although it was more predictive in log phase. For both the GC dinucleotide and the overall G+C content, the relationships with half-life were monotonic in hypoxia but were more complicated in log phase, with the slowest half-life class having lower frequencies than the medium-slow class. This could be a result of GC-rich sequences producing secondary structure that reduces ribosome binding in some cases. Given the greater apparent

impact of translation on half-life in log phase, we would expect that impediments to ribosome binding would negatively affect half-life in log phase more than in hypoxia.

## Leaderless gene start codons appear to affect transcription rate but not transcript half-life

Mycobacteria use both AUG and GUG start codons at high frequency. However, leaderless transcripts have more GUG start codons while leadered transcripts have more AUG start codons (Figure S3-7A, D), leading us to investigate the relationship between start codon and degradation rate. Start codon identity had low Gini importance rankings for both leadered and leaderless genes, suggesting that it may not be a major determinant of translation efficiency for either transcript type. Despite its low Gini importance, AUG-initiating leadered transcripts had slightly longer half-lives on average than GUG-initiated leadered transcripts in log phase (Figure S3-7B, E). When we examined steady-state transcript abundance as a function of start codon, we found that GUG-initiated transcripts had higher abundance on average, and that this effect was stronger for leaderless transcripts (Figure S3-7C, F). Since the relationship between start codon usage and steady-state abundance was stronger for leaderless transcripts and not explained by half-life, we considered that the identity of the first nt of a transcript may affect the efficiency of transcription initiation. It is well known that *E. coli* RNA polymerase preferentially initiates transcription with purines (50), and consistent with this, mycobacterial transcripts most often begin with purines (5,6,24). We examined the identity of the first nt of the 5′ UTRs of leadered transcripts and found that while transcripts beginning with As and Gs had equivalent half-lives, those beginning with G had higher abundance on average (Figure S3-7G-I). Together, these data suggest that mycobacterial RNA polymerase initiates transcription more efficiently with Gs than As.

## Identification of specific features differentially predictive of half-life in log phase and hypoxia

In order to identify the specific transcript features that were differentially important for classification among conditions, for each transcript type we used the same set of features, selected by log phase models, was used to train models in log phase, hypoxia and fold change in hypoxia. We then combined the top 20 features from each condition and compared the relative importance levels of these features across conditions (Figure 3-5A). Our results further confirmed the collective effect of various features on dictating transcript stability in each condition, but more importantly, revealed the ways in which the contributions of these features differed among conditions. Consistent with our results of training using 5′ UTR related features only (Figure 3-3B, 3-4C), the 5′ UTR features remained important across conditions in the combined feature models for leadered transcripts (Figure 3-5A). These results further confirmed the role of 5′ UTR in predicting transcript stability.

The ΔG of unfolding secondary structure at translation initiation regions (TIRs) is a feature that can be used to predict the ribosome accessibility (10). We found that the ΔG of unfolding TIRs was an important transcript feature associated with half-life for leadered transcripts in log phase (Figure 3-4A, 3-5B, SHAP distributions of 3-5B at Figure S3-11). Consistent with our finding that translation was more important in log phase, the ΔG of unfolding TIRs exhibited a stronger inverse correlation with half-life in log phase than in hypoxia (Figure 3-5B). This is consistent with a model in which higher accessibility of TIRs to ribosomes leads to greater translation efficiency or greater association of transcripts with ribosomes, thus protecting transcripts from degradation in log phase. The greater importance of translation in log phase was also supported by the stronger

correlation between frequencies of certain codons and half-life, such as AAA (Lys) (Figure 3-5B). However, the effects of codon frequency on transcript half-life might be a mixture of translational and non-translational effects, as suggested by the higher importance of ACG (Thr) in hypoxia (Figure 3-5B), where translation overall appears to have less impact on half-life.

In contrast, for leaderless transcripts in log phase, our results indicated a more complicated relationship between secondary structure and translation, and their correlations with transcript stability. Although it was not reflected by the $\Delta G$ of unfolding TIRs, the secondary structure at the 5′ end was still important for leaderless transcripts in log phase. Particularly, we observed a low A nucleotide frequency in the first 18 nucleotides of the CDS for transcripts in the fast half-life class and a low G+C frequency for those in the slow half-life class (Figure 3-5C, SHAP distributions at Figure S3-11). Notably, these features associated with secondary structure of the first 18 nt of CDSs were more predictive of half-life class for leaderless than leadered genes. While low secondary structure in this region is typical in many organisms (48) and was experimentally shown to increase translation efficiency for leadered transcripts in *E. coli* (47), it may have a larger influence on translation of leaderless genes because these lack the additional ribosome recruitment signals found in 5′ UTRs.

We found that the impact of secondary structure continued beyond the 5′ 18 nt of leaderless transcripts. We calculated $\Delta G_{MFE}$ with different sequence window sizes to measure the secondary structure of the 5′ third, middle third, and 3′ third of each CDS. Consistent with our previous observation that these features were collectively more predictive of half-life class for leaderless genes in log phase (Figure 3-3C), we found generally negative correlations between CDS $\Delta G_{MFE}$ and transcript half-life (Figure 3-5D, SHAP distributions at Figure S3-11). These correlations were

maintained when $\Delta G_{MFE}$ was calculated using different window sizes (Figure S3-8A). The trends of CG dinucleotide and UA dinucleotide frequency also supported this idea (Figure 3-5D). Overall, our results suggested that the CDS secondary structure is positively correlated with transcript half-life regardless of the region of CDS, consistent with the idea that secondary structure generally protects transcripts from cleavage by RNases (Figure 3-5D). These relationships were monotonic in hypoxia, but more complicated in log phase where transcripts in the slow half-life class deviated from the otherwise monotonic trend, having less secondary structure than those in the medium-slow class (Figure 3-5D). We hypothesized that stronger secondary structure might compete with ribosome binding, and since translation appears to have a strong protective effect in log phase only, transcripts in the slow class in log phase might be protected more by ribosome binding than by secondary structure. To test this, we quantified the ribosome occupancy within the 5′ third, middle third, and 3′ third of each CDS and evaluated their correlations with transcript half-life. As expected, we observed that the slow half-life class had the highest average ribosome occupancy across the entire CDS (Figure 3-5E). The idea of competition between secondary structure and ribosome binding was further supported by a positive correlation between $\Delta G_{MFE}$ and ribosome occupancy for the first third of transcripts in the slow class (Figure 3-5F). This trend was maintained when $\Delta G_{MFE}$ was calculated using a different window size, but was not observed for the middle and 3′ thirds of transcripts (Figure S3-8B-E). Together, our results highlight the potential complexity of interplay between transcript features.


**Figure 3-5. Transcript features differentially predict half-life in log phase and hypoxia.**
**A.** Summary of the most important features for the log phase, hypoxia, and log-to-hypoxia-fold-change models for leadered and leaderless transcripts. For each transcript type, random forest classifiers were trained using the same set of features, selected by the log phase model, for all three conditions. For each

transcript type, the 20 features with the highest Gini importance scores in each condition were then combined and their relative importance rankings indicated by intensity of coloration in the heatmap. See Table S3-5 for feature definitions and details. **B-D.** Feature value distributions within each half-life class for selected features that differentially predicted half-life class in different models. Dimmed plots indicate that the feature was less important for that condition and/or gene type. Dots and bars represent median and interquartile range. **B.** Selected features that were differentially important for log phase and hypoxia models for leadered genes. **C.** Selected features that were more important for log phase leaderless transcript models and are expected to impact the secondary structure of the 5′ ends of coding sequences. Plots for leadered genes are shown for comparison even though these features were not highly ranked for any leadered transcript models. **D.** Selected secondary-structure-related features that were relatively highly ranked for leaderless genes in both log phase and hypoxia models but showed different patterns of distributions across half-life classes for the two conditions. **E.** Log phase ribosome occupancy was quantified separately for each third of the CDS of each leaderless transcript. The x axes denote abundance of ribosome-bound reads mapping to the indicated transcript regions. **F.** For leaderless genes, the log phase ribosome occupancy for the first third of each CDS was plotted as a function of the $\Delta G_{MFE}$ of the first third of the CDS. The statistical significance of the Spearman correlation and slope of the linear regression fit line are shown in square brackets. $p < 0.01$ **, $p < 0.001$ ***.

## Transcript abundance and length are more predictive of half-life in hypoxia

Among the most important features, steady-state abundance and CDS length are the two identified by models across transcript types and conditions (Figure 3-5A, 3-4A, Figure S3-6A, B). The relationship between the transcript abundance and half-life has been investigated in various bacteria, yet the results are conflicting (Reviewed in (51)). Consistent with the studies in *M. tuberculosis* (2), *E. coli* (14-16,52-54) and *L. lactis* (13,16,55), we found that the distributions of transcript abundance exhibited an inverse correlation with half-life for both transcript types and conditions (Figure 3-6A, SHAP distributions at Figure S3-12). Although the correlations were weaker than what was reported for *M. tuberculosis* in log phase (2), we found that in *M. smegmatis* they were substantially stronger in hypoxia than log phase (Figure 3-6B). Comparing the correlations for leadered and leaderless transcripts did not reveal differences in log phase, but a stronger correlation was seen for leaderless transcripts compared to leadered transcripts in hypoxia (Figure 3-6B). These results indicate that, underlying the broad inverse correlation

between transcript abundance and half-life, the strength of the relationship varies depending on condition and to a lesser extent on transcript type.

5′ UTR length was an important feature in hypoxia but not in log phase (Figure 3-5A), and consistent with this, had a clear monotonic positive correlation with transcript half-life in hypoxia (Figure 3-6C, SHAP distributions at Figure S3-12). In contrast, the relationship between 5′ UTR length and transcript half-life was weaker and less straightforward in log phase (Figure 3-6C). CDS length was an important feature for both transcript types in both conditions (Figure 3-5A), exhibiting a roughly monotonic positive relationship with half-life class in hypoxia but a non-monotonic relationship in log phase (Figure 3-6D, SHAP distributions at Figure S3-12). This is consistent with the finding in *M. tuberculosis* that CDS length has little broad correlation with transcript half-life in log phase (2). In contrast, CDS length has also been shown to have negative correlation with transcript half-life in *L. lactis*, *E. coli*, and *S. cerevisiae* (13,18,20). While the strong predictive power of CDS length in *M. smegmatis* was intriguing, there were two potential confounding factors. First, RIF only inhibits transcription initiation, having no impact on elongating RNA polymerases. We attempted to control for this during the process of half-life determination by identifying transcripts with delays in degradation following the addition of RIF and excluding the 1-2 minute delay periods from the half-life calculation (see Figure S2 in (24)). However, for longer transcripts the elongating RNA polymerases may continue for longer than 2 minutes, leading to an overestimation of half-life.

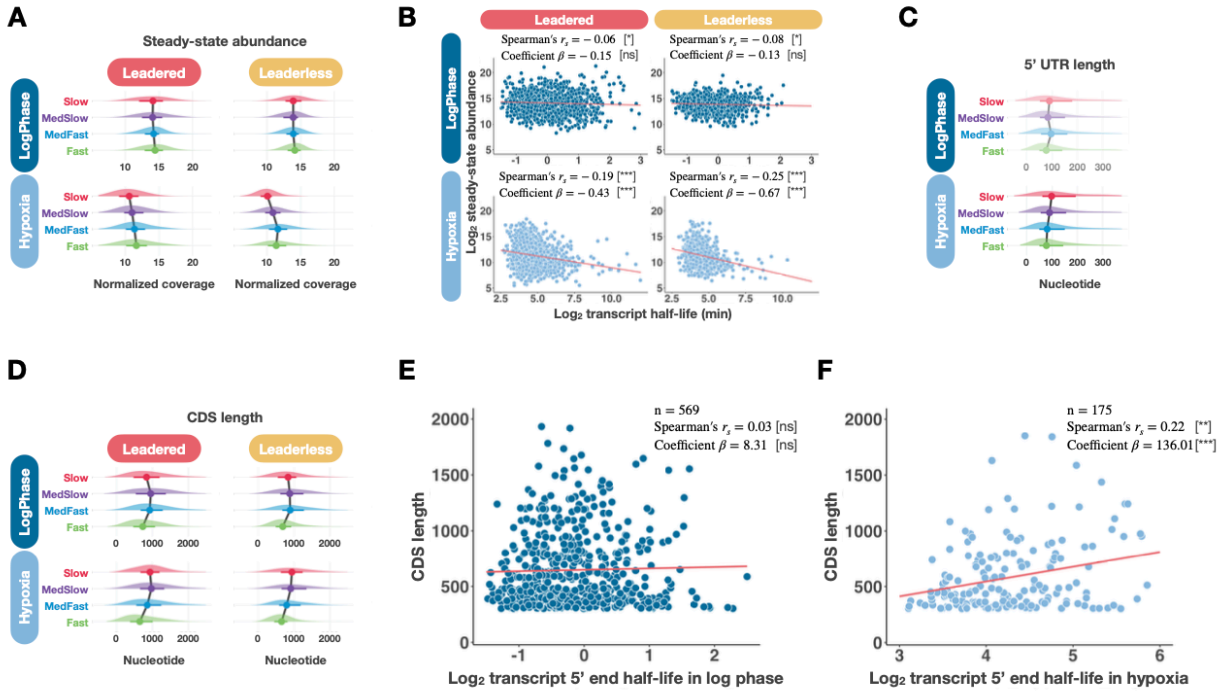



**Figure 3-6. Steady-state transcript abundance is negatively associated with half-life, while transcript length is positively correlated with mRNA half-life in hypoxia.**
**A.** Distributions of steady-state transcript abundance within each half-life class in log phase and hypoxia for leadered and leaderless transcripts. **B.** Correlations between steady-state abundance and transcript half-life. While abundance was highly ranked in all models (see Figure 3-5A), its negative correlation with half-life was stronger in hypoxia. **C.** Distributions of 5′ UTR lengths within half-life classes for leadered transcripts in log phase and hypoxia. This feature had a high importance ranking in hypoxia only. **D.** Distributions of CDS length within each half-life class in log phase and hypoxia. This feature was highly ranked in all models. **E-F.** Half-lives were calculated for only the first 300 nt of each CDS and genes were selected that had similar half-lives for the 5′ 300 nt and the whole CDS (see Figure S3-9). For these subsets of genes in log phase and hypoxia, the correlation between CDS length and 5′ 300 nt half-life are shown. In **B, E, F**, the statistical significance of the Spearman correlation and slope of the linear regression fit line are shown in square brackets. $p < 0.05$ *, $p < 0.01$ **, $p < 0.001$ ***.

Secondly, recent studies of transcript 3′ ends in *M. tuberculosis* (56) and *E. coli* (57) suggested that a sizable fraction of transcripts present in cells are degradation intermediates or incomplete transcripts resulting from premature transcription termination or paused RNA polymerases. We cannot distinguish these from complete transcripts in our RNAseq libraries, and it is possible that

longer transcripts give rise to more incomplete transcripts that are long enough to be captured in RNAseq libraries and these have different degradation kinetics than complete transcripts.

To account for these potential confounders, we calculated the half-life of only the first 300 nt of each CDS (the "5′ end half-life"). For each condition, we then divided genes into five groups according to the ratio of the 5′ end half-life to the entire gene half-life (Figure S3-9A, D). We also calculated the steady-state (0 minute RIF) coverage ratio of the 5′ 300 nt versus 3′ end 300 nt of each gene within these groups (Figure S3-9A, D). As expected, those genes with differential abundance of transcript 5′ and 3′ regions often had non-zero $\log_2$ half-life ratios, consistent with the idea that incomplete transcript fragments often have different degradation kinetics than full-length transcripts. On the other hand, genes with similar coverage of their 5′ and 3′ 300 nt generally had similar $\log_2$ half-life ratios (Figure S3-9A, D, groups 3 and 4 respectively), indicating that these genes are likely less affected by the confounders described above (Colored group in Figure S3-9A, D). For these non-confounded genes, there was no correlation between 5′ end half-life and CDS length in log phase (Figure 3-6E), but there was a significant positive correlation in hypoxia (Figure 3-6F). These relationships were maintained when leadered and leaderless transcripts were analyzed separately (Figure S3-9B, C, E, F). Consistent with the idea that the positive correlation in hypoxia was due to the condition rather than the selection of genes, we found little to no correlation between 5′ end half-life and CDS length for the genes in Figure 3-6F in log phase (Figure S3-9G-I). In summary, although being able to contribute to model predictions for both transcript types in log phase, transcript abundance and CDS length seemed to have stronger correlations with half-life in hypoxia. Consistent with this, the 5′ UTR length exhibited a

positive correlation with half-life in hypoxia, suggesting that the overall transcript length is more important for predicting half-life in hypoxia.

## Discussion

In this study, we used transcriptome-wide mRNA half-life datasets to investigate the intrinsic features that impact transcript stability in aerobically growing and hypoxia-arrested *M. smegmatis*. This led us to discover the diverse transcript features that were differentially associated with mRNA stability depending on the microenvironment. These diverse features provided evidence that translation is likely to have a larger impact on mRNA degradation in log phase than in hypoxia. We further found that, coupled with the impact of conditions, the leader status of the transcripts (leadered vs leaderless) also impacted transcript stability through various transcript features. More importantly, our results showed that it is the collective effect of diverse transcript features that shaped the transcript stability in *M. smegmatis*. Such collective impact of transcript features on mRNA half-life has been reported in other organisms as well (15,18,19). However, our study further revealed the non-linear character of such collective effect differentially in the context of transcript type and growth condition.

We developed machine learning models with the goal of associating transcript features with half-life by predicting half-life using a wide-ranging feature set, as well as to further identify likely determinants of transcript half-life by quantifying the strength of their associations. Initially, we attempted to develop regression models given the continuous nature of transcript half-life values. However, the models failed to provide accurate prediction of half-life values for us to draw reliable conclusions about feature relationships with transcript half-life. Therefore, we grouped transcript

half-life values into four half-life classes to predict half-life through classification instead. The decision to define four half-life classes was informed by the hierarchical clustering of degradation profiles to estimate the number of groups to best represent transcriptome-wide stability. Besides the innate difficulty of the four classes prediction task, the intertwined non-linear correlations, existing not only between transcript features and half-life but also among transcript features themselves, makes the classifications even more challenging. Despite the difficulties, our models achieved significantly better performance than random predictions. To compensate for the suboptimal model performance, we implemented SHAP visualization to enhance our interpretation of model predictions by showing the prediction direction along with the feature values for each half-life class (Figure S3-10-S3-12). We found that these results were consistent with the correlations we learned from the distributions of individual transcript features. Together, these computational tools provided us with enough confidence and information to draw conclusions about the associations between transcript features and half-life. Nonetheless, our current models still lack the ability to fully explain the relationships between transcript features themselves, and the mechanism of how they work together to determine transcript stability. Future studies on the relationships among those important transcript features will greatly improve model predictions and advance our understanding of the regulatory mechanism of transcript degradation.

Like *M. tuberculosis*, *M. smegmatis* exhibited variance in transcript half-lives during log phase growth and showed transcriptome-wide stabilization when exposed to hypoxia (Figure 3-2B) (2,25). Despite the potential difference in regulatory mechanisms, our study in *M. smegmatis* still provides insights to facilitate understanding of transcript stability in *M. tuberculosis*. Unlike in the

previous study in *M. tuberculosis* (2), we were able to quantify mRNA half-lives transcriptome-wide in hypoxia, showing that the extent of stabilization varied among genes and indicating that the determinants of half-life differ between the two conditions. It was reported in *M. tuberculosis* that transcript abundance was the single feature strongly correlated with transcript half-lives, while features like CDS length and G+C content showed little correlation (2). Here we greatly expanded the scope of candidate features and found diverse transcript features could contribute to predict half-life in *M. smegmatis*. Whether the collective and differential effect of the wide range of transcript features on half-life we observed in *M. smegmatis* also exists in *M. tuberculosis* awaits further investigation. Our results also suggested that the lack of broad correlations between transcript features and half-lives could be due to not only condition, but also to transcript-type-dependent regulatory mechanisms in mycobacteria.

In log phase, transcripts in both *M. tuberculosis* and *M. smegmatis* exhibited little correlation between CDS length and half-life. However, we found that the correlation became stronger in hypoxia for *M. smegmatis*. A previous study suggested that motion of large cytoplasmic components was dramatically reduced in *Caulobacter crescentus* and *E. coli* when metabolic activity was reduced, due to decreased fluidity of the cytoplasm (58). The positive correlation between CDS length and half-life is consistent with this idea, as longer transcripts would be more affected by the reported changes in diffusion rates (58), leading to reduced encounters between transcripts and RNases in the hypoxic cytoplasm. Transcript abundance was another feature whose influence was affected by growth condition. Similar to *M. tuberculosis* (2), we also observed an inverse correlation between transcript abundance and half-life in log phase, although this relationship was much weaker in *M. smegmatis*. However, we found that the strength of the

correlation was stronger in hypoxia compared to log phase. Such an association has been reported for other bacteria as well (13-16,52-55,59,60), although conflicting reports exist for *E. coli*, where some report a negative correlation (14-16,52-54) while others report no correlation or a positive correlation (59,61). The mechanistic basis of the negative correlation reported in many studies is unknown, although it has been suggested to be a function of the impact of transcript abundance on encounters with RNases (15,16).

It has been shown in *E. coli* and *S. cerevisiae* that translation efficiency is positively correlated with mRNA half-life in log phase (62,63). Our results also provided evidence to support this association in *M. smegmatis* as we observed translation-related features were more important for half-life predictions in log phase compared to hypoxia. Besides the previously identified difference in translation mechanisms (10,11,64), our results indicate that mRNA degradation mechanisms may also differ between leadered and leaderless transcripts. We first confirmed that 5′ UTR features are predictive of mRNA half-life in leadered transcripts. There were also differences in the importance of CDS features in predicting half-lives of leadered vs leaderless transcripts. For example, G+C content was particularly low in the first 18 nt of the CDS specifically for leaderless transcripts with the slowest half-lives, consistent with the idea that secondary structure in this region has a larger impact on translation efficiency for leaderless transcripts compared to leadered transcripts.

In summary, our results suggest that underlying the observed transcript stability patterns in mycobacteria lies a complex interplay between inherent transcript features and microenvironments. Additionally, our study provides a foundation to facilitate further

investigation of target transcript stability in mycobacteria, as well as an experimental and computational framework to study transcript stability more broadly in other organisms.

## Data Availability

All RNAseq data generated in this study are available at GSE227248. Other data and code generated for analysis in this study are available from the following GitHub repository, https://github.com/ssshell/mRNA_stability.

## Supplementary Materials

### Supplemental tables

Table S3-1. Complete updated gene annotations of *M. smegmatis*.

Table S3-2. Leadered gene annotations of *M. smegmatis*.

Table S3-3. Leaderless gene annotations of *M. smegmatis*.

Table S3-4. Transcript half-life values and stability classes in log phase and hypoxia of *M. smegmatis*.

Table S3-5. Complete transcript properties (features) list of *M. smegmatis*.

#### Feature selection algorithm

Below is our algorithm to reduce the number of correlated features.

**Algorithm:** Feature selection for a given feature set $\boldsymbol{F_f}$ with class labels

**Input:** Complete feature set $\boldsymbol{F_f}$

**Metrics:**

[$\rho_{ff}$], Spearman's rank correlation coefficient to quantify correlation between the values of each pair of features.

[$\tau_{fc}$], Kendall rank correlation coefficient to quantify correlation between the values of each feature and the class.

[$\boldsymbol{Ave}_{\mathbf{adjusted}P}$], Mean FDR adjusted *P* value of Kruskal-Wallis test (KW) and Kolmogorov-Smirnov test (KS). Kolmogorov-Smirnov test adjusted *P* value is taken as the minimum FDR adjusted *P* value of comparisons between each pair classes.

[$\boldsymbol{N}_{corr}$], number of correlated features with a given feature.

[$\boldsymbol{N}_{sigTest}$], number of statistically significant tests (KW, KS) of a given feature over the class.

**Procedure:**

1. Preprocessing. Removed features with zero variance.
2. Get correlated feature sets $\boldsymbol{F}_{corr}$. Each $\boldsymbol{F}_{corr}$ included features with $|\rho_{ff}| >= 0.6$. Only $\boldsymbol{F}_{corr}$ that met following criteria was further considered for selection:
    a. Correlated features have the same transcript region [*5′ UTR/5′ transcript, CDS, 3′ UTR*] and the same feature type [*Nucleotide, Codon, Secondary structure, Ribosome, Others*]
3. Determine the selection starting order of $\boldsymbol{F}_{corr}$. For all the $\boldsymbol{F}_{corr}$ that met the criteria in step 2, starting selection with the $\boldsymbol{F}_{corr}$ that included feature that has the highest $|\tau_{fc}|$ among features in all $\boldsymbol{F}_{corr}$. If tied, start with the $\boldsymbol{F}_{corr}$ that has the least amount of correlated features.
4. Select feature. For features within $\boldsymbol{F}_{corr}$:
    a. Select feature that meets criteria in the following order:
        i. Maximum $\boldsymbol{N}_{sigTest}$. If tied, go to the next metric,
        ii. Minimum $\boldsymbol{N}_{corr}$. If tied, go to the next metric,
        iii. Minimum group sum $|\rho_{ff}|$. If tied, go to the next metric,
        iv. Maximum $|\tau_{fc}|$. If tied, go to the next metric,
        v. Minimum $\boldsymbol{Ave}_{\mathbf{adjusted}P}$
    b. Update the rest of $\boldsymbol{F}_{corr}$ with features being selected and excluded.
5. Continue until all $\boldsymbol{F}_{corr}$ were selected.

**Output:** the list of selected features.

**Machine learning classifier training algorithm**

Below is the generalized classifier training algorithm. For our classifiers, $k = 5$, $n = 10$.

**Algorithm:** Classifiers training and evaluation with $k$-fold nested cross-validation for $n$ repetitions.

**Input:** Feature set $D$ with class labels, Hyperparameter set $H$

1. For $i = 1 \ldots n$ repetitions:
2. Training($D$, $k$, $H$):
3. Random stratified partition $D$ into $k$ folds $D_1 \ldots D_k$
4. For $j = 1 \ldots k$ folds:
5. *TrainSet* = $D \setminus D_j$
6. *TestSet* = $D_j$
7. Train the RandomBaselineModel on *TrainSet*
8. $h^*$ = RandomizedSearchCV[*TrainSet*, $H$, RandomForestModel]
9. $RandomForestModel^*$ = RandomForestModel trains on *TrainSet* using $h^*$
10. $Fscore_j$ = RandomBaselineModel & $RandomForestModel^*$ predict on *TestSet*
11. $FeatureImportance_j$ = Gini importance[$RandomForestModel^*$]
12. $SHAP_j$ = SHAP values of $RandomForestModel^*$ predicts on *TestSet*
13. $Fscore\_allFold_i$ = [$Fscore_j$]
14. $Fscore_i$ = Mean[$Fscore_j$]
15. $Gini_i$ = Mean[$FeatureImportance_j$]
16. $SHAP_i$ = Concatenate[$SHAP_j$]
17. $Fscore\_allFold$ = [$Fscore\_allFold_i$]
18. $Fscore\_average$ = Mean[$Fscore_i$]
19. $Gini$ = Mean[$Gini_i$]
20. $SHAP$ = Concatenate[$SHAP_i$]
21. Return $Fscore\_allFold$, $Fscore\_average$, $Gini$, $SHAP$, $Metrics\_others$

**Output:** Classifier performance of all fold across repetitions $Fscore\_allFold\ (n = 50)$, classifier performance averaged across repetitions $Fscore\_average\ (n = 10)$, impurity-based feature importance quantification averaged across repetitions $Gini$, SHAP values of individual class

***SHAP***. Output also includes other performance metrics such as precision and recall values for both individual class and averaged value across folds.

## References


1. World Health, O. (2023) *Global tuberculosis report 2023*. Geneva: World Health Organization.

2. Rustad, T.R., Minch, K.J., Brabant, W., Winkler, J.K., Reiss, D.J., Baliga, N.S. and Sherman, D.R. (2013) Global analysis of mRNA stability in Mycobacterium tuberculosis. *Nucleic Acids Res*, **41**, 509-517.

3. Via, L.E., Lin, P.L., Ray, S.M., Carrillo, J., Allen, S.S., Eum, S.Y., Taylor, K., Klein, E., Manjunatha, U., Gonzales, J. *et al.* (2008) Tuberculous granulomas are hypoxic in guinea pigs, rabbits, and nonhuman primates. *Infect Immun*, **76**, 2333-2340.

4. Belton, M., Brilha, S., Manavaki, R., Mauri, F., Nijran, K., Hong, Y.T., Patel, N.H., Dembek, M., Tezera, L., Green, J. *et al.* (2016) Hypoxia and tissue destruction in pulmonary TB. *Thorax*, **71**, 1145-1153.

5. Cortes, T., Schubert, O.T., Rose, G., Arnvig, K.B., Comas, I., Aebersold, R. and Young, D.B. (2013) Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in Mycobacterium tuberculosis. *Cell Rep*, **5**, 1121-1131.

6. Shell, S.S., Wang, J., Lapierre, P., Mir, M., Chase, M.R., Pyle, M.M., Gawande, R., Ahmad, R., Sarracino, D.A., Ioerger, T.R. *et al.* (2015) Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLoS Genet*, **11**, e1005641.

7. Martini, M.C., Zhou, Y., Sun, H. and Shell, S.S. (2019) Defining the Transcriptional and Post-transcriptional Landscapes of Mycobacterium smegmatis in Aerobic Growth and Hypoxia. *Front Microbiol*, **10**, 591.

8. Nguyen, T.G., Vargas-Blanco, D.A., Roberts, L.A. and Shell, S.S. (2020) The Impact of Leadered and Leaderless Gene Structures on Translation Efficiency, Transcript Stability, and Predicted Transcription Rates in Mycobacterium smegmatis. *J Bacteriol*, **202**.

9. Scharff, L.B., Childs, L., Walther, D. and Bock, R. (2011) Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. *PLoS Genet*, **7**, e1002155.

10. Bharmal, M.M., Gega, A. and Schrader, J.M. (2021) A combination of mRNA features influence the efficiency of leaderless mRNA translation initiation. *NAR Genom Bioinform*, **3**, lqab081.

11. Grabowska, A.D., Andreu, N. and Cortes, T. (2021) Translation of a Leaderless Reporter Is Robust During Exponential Growth and Well Sustained During Stress Conditions in Mycobacterium tuberculosis. *Front Microbiol*, **12**, 746320.

12. Richards, J. and Belasco, J.G. (2019) Obstacles to Scanning by RNase E Govern Bacterial mRNA Lifetimes by Hindering Access to Distal Cleavage Sites. *Mol Cell*, **74**, 284-295 e285.

13. Dressaire, C., Picard, F., Redon, E., Loubiere, P., Queinnec, I., Girbal, L. and Cocaign-Bousquet, M. (2013) Role of mRNA stability during bacterial adaptation. *PLoS One*, **8**, e59059.

14. Esquerre, T., Laguerre, S., Turlan, C., Carpousis, A.J., Girbal, L. and Cocaign-Bousquet, M. (2014) Dual role of transcription and transcript stability in the regulation of gene expression in Escherichia coli cells cultured on glucose at different growth rates. *Nucleic Acids Res*, **42**, 2460-2472.

15. Esquerre, T., Moisan, A., Chiapello, H., Arike, L., Vilu, R., Gaspin, C., Cocaign-Bousquet, M. and Girbal, L. (2015) Genome-wide investigation of mRNA lifetime determinants in Escherichia coli cells cultured at different growth rates. *BMC Genomics*, **16**, 275.

16. Nouaille, S., Mondeil, S., Finoux, A.L., Moulis, C., Girbal, L. and Cocaign-Bousquet, M. (2017) The stability of an mRNA is influenced by its concentration: a potential physical mechanism to regulate gene expression. *Nucleic Acids Res*, **45**, 11711-11724.

17. Kristoffersen, S.M., Haase, C., Weil, M.R., Passalacqua, K.D., Niazi, F., Hutchison, S.K., Desany, B., Kolsto, A.B., Tourasse, N.J., Read, T.D. *et al.* (2012) Global mRNA decay analysis at single nucleotide resolution reveals segmental and positional degradation patterns in a Gram-positive bacterium. *Genome Biol*, **13**, R30.

18. Neymotin, B., Ettorre, V. and Gresham, D. (2016) Multiple Transcript Properties Related to Translation Affect mRNA Degradation Rates in Saccharomyces cerevisiae. *G3 (Bethesda)*, **6**, 3475-3483.

19. Cheng, J., Maier, K.C., Avsec, Z., Rus, P. and Gagneur, J. (2017) Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast. *RNA*, **23**, 1648-1659.

20. Esquerre, T., Moisan, A., Chiapello, H., Arike, L., Vilu, R., Gaspin, C., Cocaign-Bousquet, M. and Girbal, L. (2015) Genome-wide investigation of mRNA lifetime determinants in Escherichia coli cells cultured at different growth rates. *BMC Genomics*, **16**, 275.

21. Harigaya, Y. and Parker, R. (2017) The link between adjacent codon pairs and mRNA stability. *BMC Genomics*, **18**, 364.

22. Agarwal, V. and Kelley, D.R. (2022) The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biol*, **23**, 245.

23. Yaish, O. and Orenstein, Y. (2022) Computational modeling of mRNA degradation dynamics using deep neural networks. *Bioinformatics*, **38**, 1087-1101.

24. Zhou, Y., Sun, H., Rapiejko, A.R., Vargas-Blanco, D.A., Martini, M.C., Chase, M.R., Joubran, S.R., Davis, A.B., Dainis, J.P., Kelly, J.M. *et al.* (2023) Mycobacterial RNase E cleaves with a distinct sequence preference and controls the degradation rates of most Mycolicibacterium smegmatis mRNAs. *J Biol Chem*, **299**, 105312.

25. Vargas-Blanco, D.A., Zhou, Y., Zamalloa, L.G., Antonelli, T. and Shell, S.S. (2019) mRNA Degradation Rates Are Coupled to Metabolic Status in Mycobacterium smegmatis. *mBio*, **10**.

26. Shishkin, A.A., Giannoukos, G., Kucukural, A., Ciulla, D., Busby, M., Surka, C., Chen, J., Bhattacharyya, R.P., Rudy, R.F., Patel, M.M. *et al.* (2015) Simultaneous generation of many RNA-seq libraries in a single reaction. *Nat Methods*, **12**, 323-325.

27. Kapopoulou, A., Lew, J.M. and Cole, S.T. (2011) The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis (Edinb)*, **91**, 8-13.

28. Laguerre, S., Gonzalez, I., Nouaille, S., Moisan, A., Villa-Vialaneix, N., Gaspin, C., Bouvier, M., Carpousis, A.J., Cocaign-Bousquet, M. and Girbal, L. (2018) Large-Scale Measurement of mRNA Degradation in Escherichia coli: To Delay or Not to Delay. *Methods Enzymol*, **612**, 47-66.

29. Coleman, J.R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E. and Mueller, S. (2008) Virus attenuation by genome-scale changes in codon pair bias. *Science*, **320**, 1784-1787.

30. Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol*, **6**, 26.

31. Reuter, J.S. and Mathews, D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.

32. Chen, Y.X., Xu, Z.Y., Ge, X., Sanyal, S., Lu, Z.J. and Javid, B. (2020) Selective translation by alternative bacterial ribosomes. *Proc Natl Acad Sci U S A*, **117**, 19487-19496.

33. Andrews, S. (2010).

34. Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114-2120.

35. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **9**, 357-359.

36. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M. *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**.

37. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. and Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*, **33**, 290-295.

38. Saeys, Y., Inza, I. and Larranaga, P. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507-2517.

39. KENDALL, M.G. (1938) A NEW MEASURE OF RANK CORRELATION. *Biometrika*, **30**, 81-93.

40. Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32.

41. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

42. Dietterich, T.G. (1998) Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput*, **10**, 1895-1923.

43. Nadeau, C. and Bengio, Y. (2003) Inference for the Generalization Error. *Machine Learning*, **52**, 239-281.

44. Bosch, B., DeJesus, M.A., Poulton, N.C., Zhang, W., Engelhart, C.A., Zaveri, A., Lavalette, S., Ruecker, N., Trujillo, C., Wallach, J.B. *et al.* (2021) Genome-wide gene expression tuning reveals diverse vulnerabilities of M. tuberculosis. *Cell*, **184**, 4579-4592 e4524.

45. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.I. (2020) From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell*, **2**, 56-67.

46. McInnes, L., Healy, J. and Melville, J. (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

47. Goodman, D.B., Church, G.M. and Kosuri, S. (2013) Causes and effects of N-terminal codon bias in bacterial genes. *Science*, **342**, 475-479.

48. Gu, W., Zhou, T. and Wilke, C.O. (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol*, **6**, e1000664.

49. Chen, F., Cocaign-Bousquet, M., Girbal, L. and Nouaille, S. (2022) 5'UTR sequences influence protein levels in Escherichia coli by regulating translation initiation and mRNA stability. *Front Microbiol*, **13**, 1088941.

50. Maitra, U. and Hurwitz, H. (1965) The role of DNA in RNA synthesis, IX. Nucleoside triphosphate termini in RNA polymerase products. *Proc Natl Acad Sci U S A*, **54**, 815-822.

51. Vargas-Blanco, D.A. and Shell, S.S. (2020) Regulation of mRNA Stability During Bacterial Stress Responses. *Front Microbiol*, **11**, 2111.

52. Bernstein, J.A., Khodursky, A.B., Lin, P.H., Lin-Chao, S. and Cohen, S.N. (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci U S A*, **99**, 9697-9702.

53. Esquerre, T., Bouvier, M., Turlan, C., Carpousis, A.J., Girbal, L. and Cocaign-Bousquet, M. (2016) The Csr system regulates genome-wide mRNA stability and transcription and thus gene expression in Escherichia coli. *Sci Rep*, **6**, 25057.

54. Morin, M., Enjalbert, B., Ropers, D., Girbal, L. and Cocaign-Bousquet, M. (2020) Genomewide Stabilization of mRNA during a "Feast-to-Famine" Growth Transition in Escherichia coli. *mSphere*, **5**.

55. Redon, E., Loubiere, P. and Cocaign-Bousquet, M. (2005) Role of mRNA stability during genome-wide adaptation of Lactococcus lactis to carbon starvation. *J Biol Chem*, **280**, 36380-36385.

56. Ju, X., Li, S., Froom, R., Wang, L., Lilic, M., Delbeau, M., Campbell, E.A., Rock, J.M. and Liu, S. (2024) Incomplete transcripts dominate the Mycobacterium tuberculosis transcriptome. *Nature*.

57. Herzel, L., Stanley, J.A., Yao, C.C. and Li, G.W. (2022) Ubiquitous mRNA decay fragments in E. coli redefine the functional transcriptome. *Nucleic Acids Res*, **50**, 5029-5046.

58. Parry, B.R., Surovtsev, I.V., Cabeen, M.T., O'Hern, C.S., Dufresne, E.R. and Jacobs-Wagner, C. (2014) The bacterial cytoplasm has glass-like properties and is fluidized by metabolic activity. *Cell*, **156**, 183-194.

59. Chen, H., Shiroguchi, K., Ge, H. and Xie, X.S. (2015) Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli. *Mol Syst Biol*, **11**, 808.

60. Redon, E., Loubiere, P. and Cocaign-Bousquet, M. (2005) Transcriptome analysis of the progressive adaptation of Lactococcus lactis to carbon starvation. *J Bacteriol*, **187**, 3589-3592.

61. Moffitt, J.R., Pandey, S., Boettiger, A.N., Wang, S. and Zhuang, X. (2016) Spatial organization shapes the turnover of a bacterial transcriptome. *Elife*, **5**.

62. Boel, G., Letso, R., Neely, H., Price, W.N., Wong, K.H., Su, M., Luff, J., Valecha, M., Everett, J.K., Acton, T.B. *et al.* (2016) Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature*, **529**, 358-363.

63. Presnyak, V., Alhusaini, N., Chen, Y.H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R. *et al.* (2015) Codon optimality is a major determinant of mRNA stability. *Cell*, **160**, 1111-1124.

64. Sawyer, E.B., Phelan, J.E., Clark, T.G. and Cortes, T. (2021) A snapshot of translation in Mycobacterium tuberculosis during exponential growth and nutrient starvation revealed by ribosome profiling. *Cell Rep*, **34**, 108695.

65. Conway, J.R., Lex, A. and Gehlenborg, N. (2017) UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, **33**, 2938-2940.

# Supplemental Figures

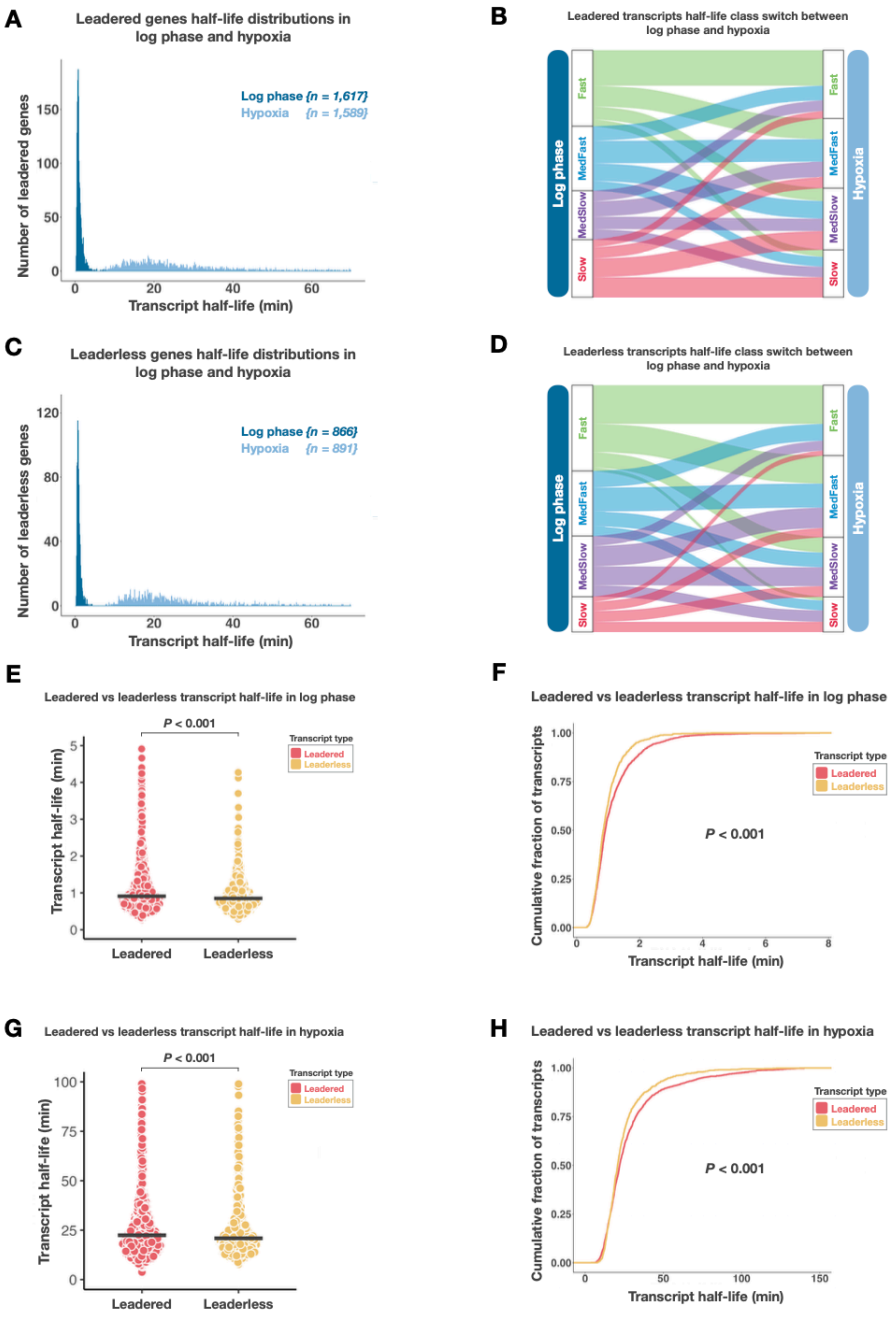



**Figure S 3-1. Leadered and leaderless transcript half-life distributions and half-life classifications.**

Half-life distributions of leadered (**A**) and leaderless (**C**) transcripts in log phase and hypoxia. Comparison of half-life class membership between log phase and hypoxia for leadered (**B**) and leaderless (**D**) transcripts. **E-H.** Comparison of half-life values between leadered and leaderless transcripts. **E** and **G**, black lines indicate

the median and transcript types were compared using the Wilcoxon rank-sum test. **F** and **H**, transcript types were compared using Kolmogorov-Smirnov test.

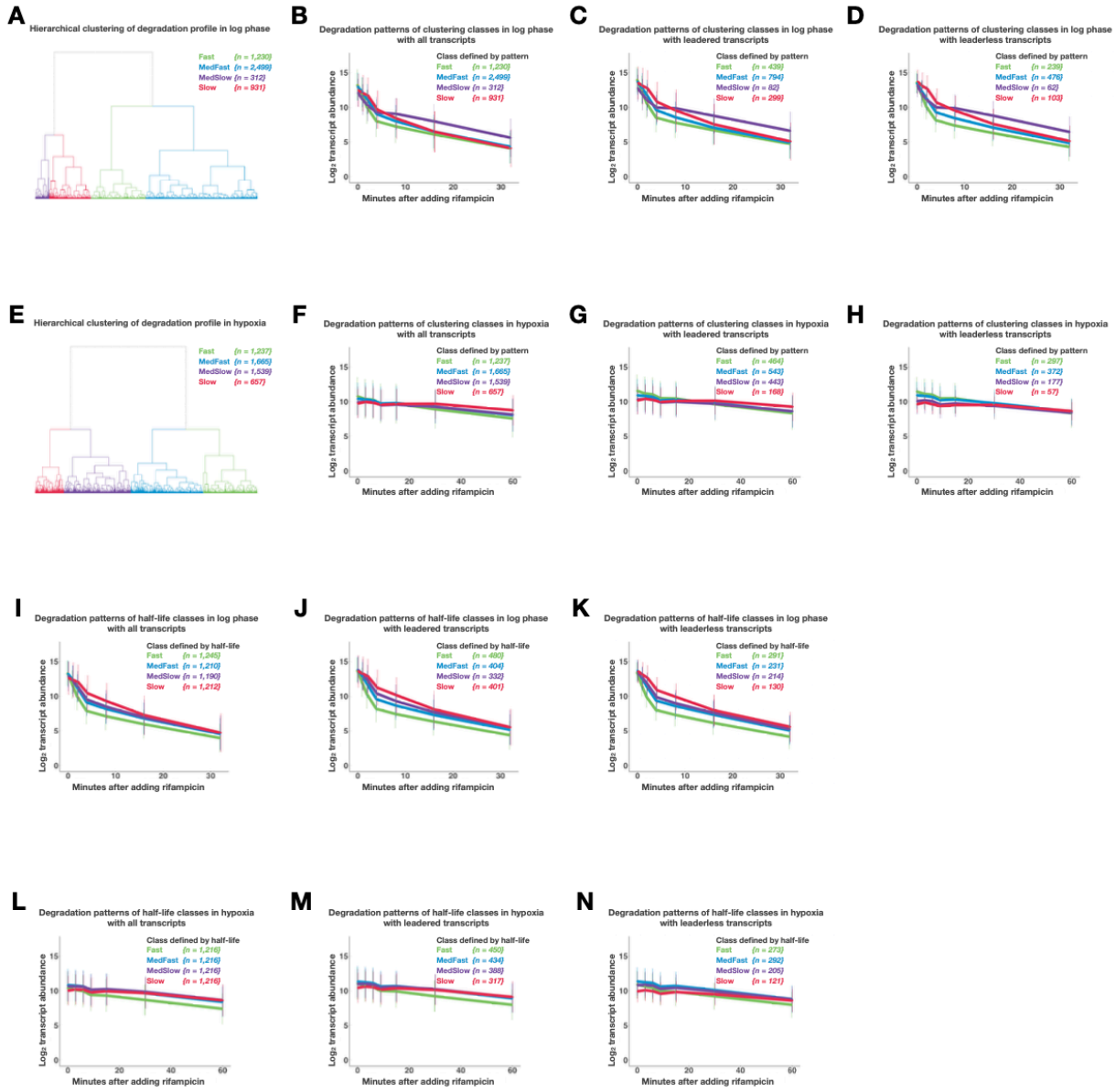



**Figure S 3-2. Classification of transcripts by hierarchical clustering of degradation patterns, and comparison with half-life classes.**

Hierarchical clustering of degradation profiles in log phase (**A**) and hypoxia (**E**) using hierarchical clustering with the Euclidean distance measure and ward.D2 agglomeration method. Degradation patterns of classes defined by hierarchical clustering for all transcripts, leadered transcripts and leaderless transcripts in log phase (**B-D**) and hypoxia (**F-H**). Degradation patterns of classes defined by half-life values (see Figure 2) for all transcripts, leadered transcripts, and leaderless transcripts in log phase (**I-K**) and hypoxia (**L-N**).

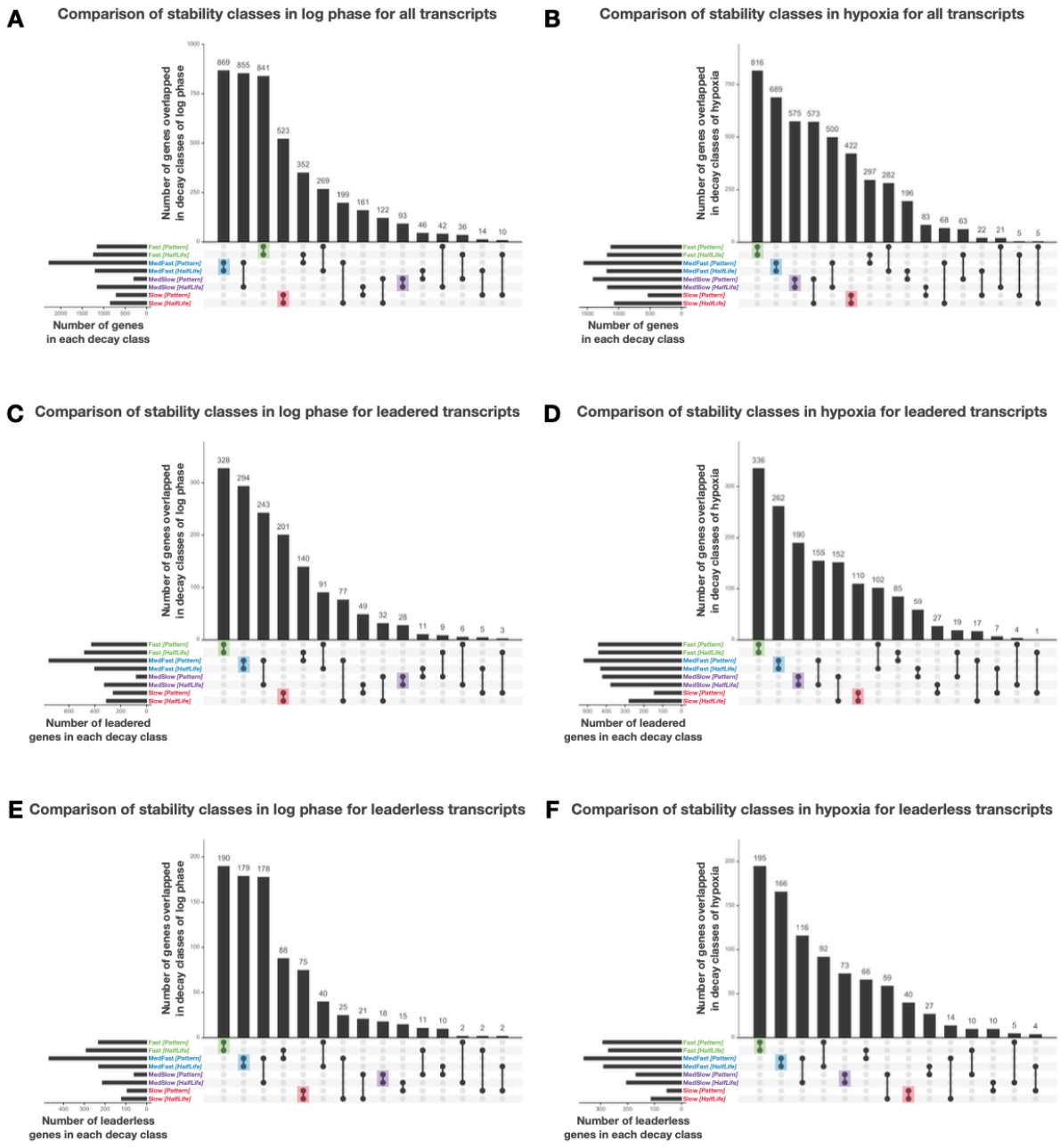



**Figure S 3-3. Comparisons of gene membership in classes defined by half-lives and classes defined by hierarchical clustering.**

Half-life classes were defined in Figure 2, and pattern classes were defined by hierarchical clustering in Figure S2. Visualizations of the number of genes overlapped in each class defined with the two metrics for all transcripts (**A, B**), leadered transcripts(**C, D**) and leaderless transcripts (**E, F**) in log phase (**A, C, E**) and hypoxia (**B, D, F**). Upset plots were made in R v4.3.2 using package UpSetR v1.4.0 (65).

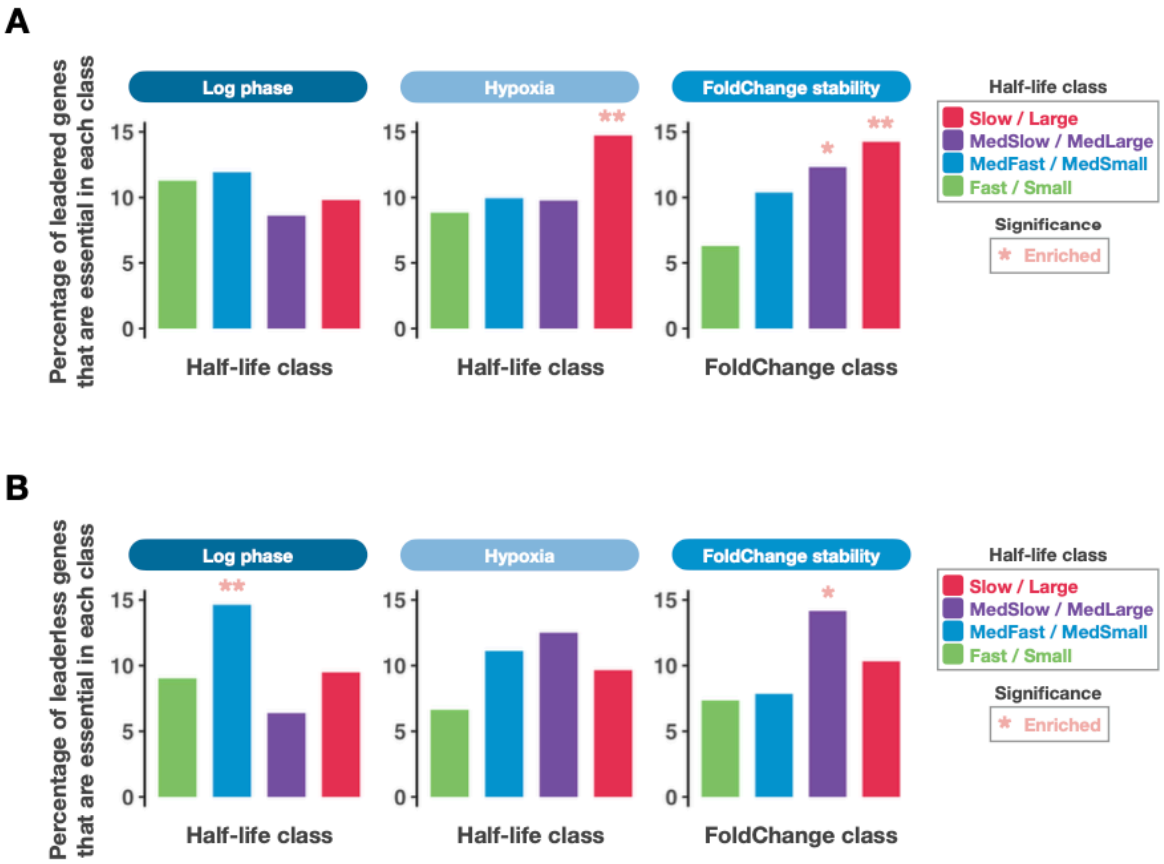



**Figure S 3-4. Frequency of essential genes in each half-life class.**
Percentage of genes that are essential in each class for leadered (**A**) and leaderless (**B**) genes in each condition. Significance of enrichment and depletion of essential genes within each class were tested using a hypergeometric test with FDR correction (Materials and Methods). *p.adjust* < 0.05 *, *p.adjust* < 0.01 **.

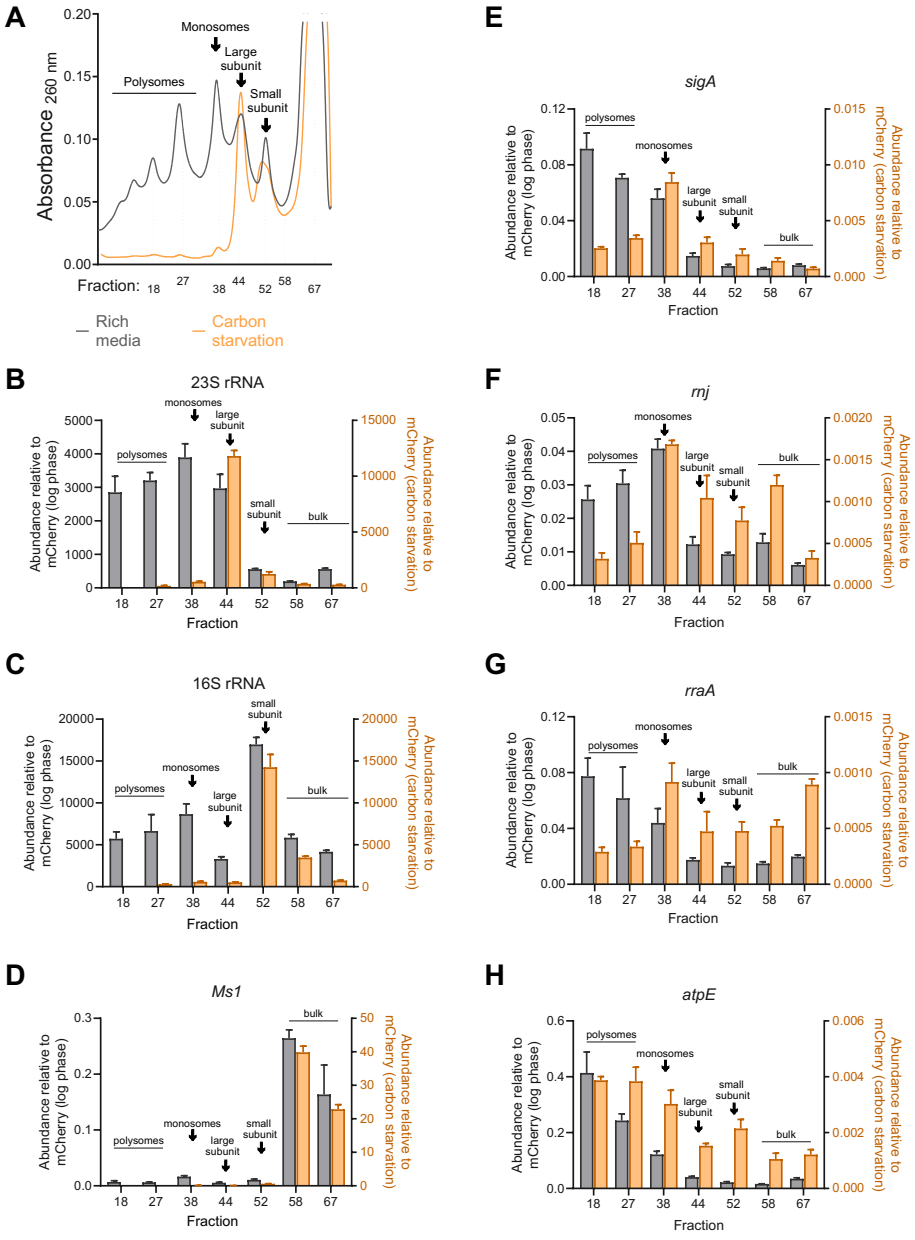



**Figure S 3-5. Carbon-starved *M. smegmatis* has fewer polysomes and a smaller proportion of its mRNA is associated with ribosomes.**

Throughout the figure, gray indicates samples from log phase growing *M. smegmatis* and orange indicates samples from *M. smegmatis* that was carbon-starved for 22 hours. **A.** Representative polysome profiling traces of sucrose-gradient-separated lysates from log phase and carbon-starved *M. smegmatis*. Selected fraction numbers are indicated. The ribosome composition of each fraction was determined by gel electrophoresis to assess relative rRNA abundance. **B-H.** Fractions from (**A**) were spiked with equal masses of in vitro-transcribed mCherry RNA and qPCR was used to quantify the abundance of the indicated transcripts relative to mCherry. Means and SD of triplicate samples from a representative experiment are shown. The entire experiment was performed twice. **B-C.** rRNA abundance. **D.** A known sRNA, Ms1, is localized in the bulk fractions consistent with expectations that it is not translated. **E-H.** The transcripts of

arbitrarily selected genes generally have relatively greater association with ribosomes during log phase growth than in carbon starvation.

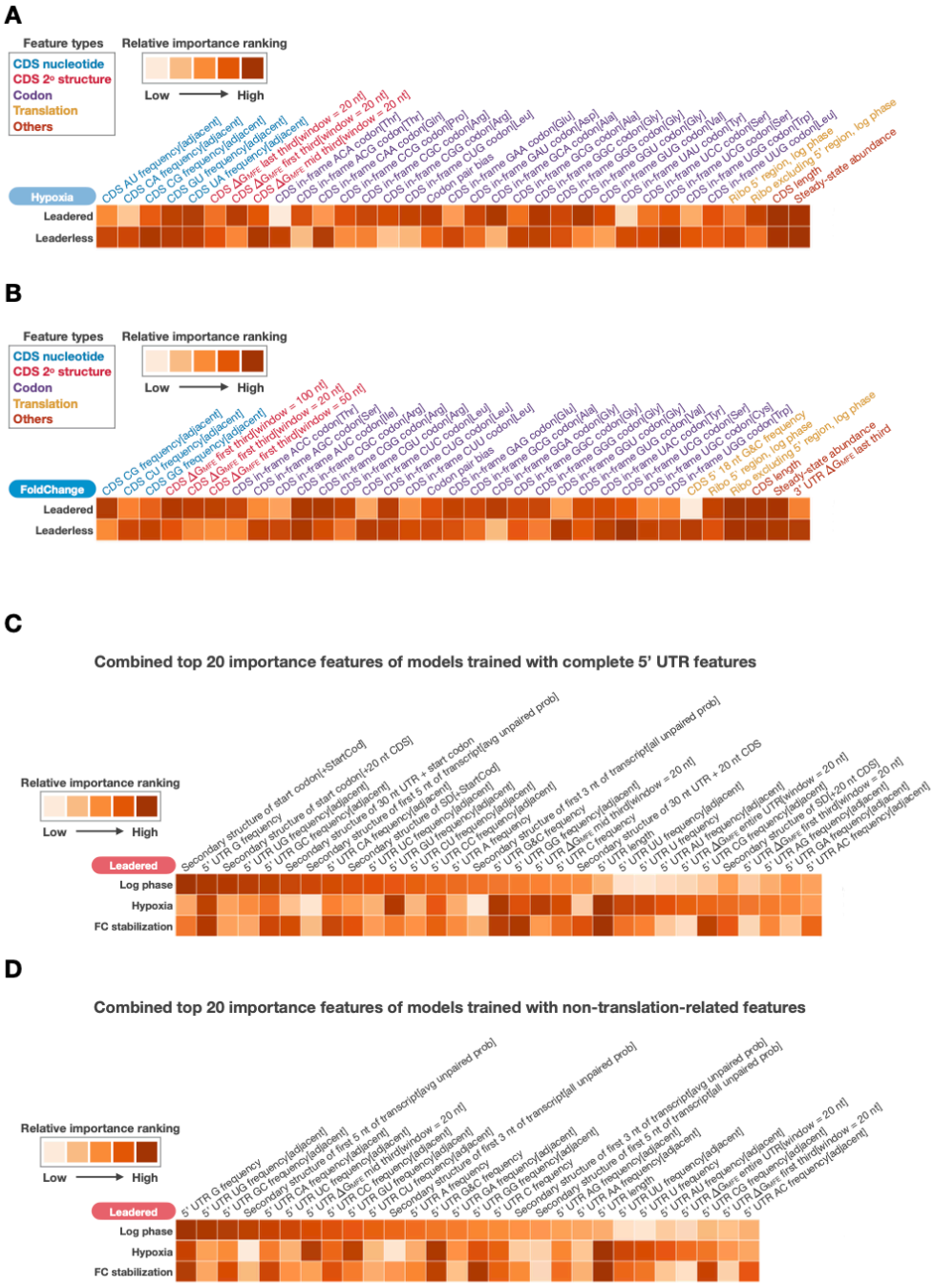
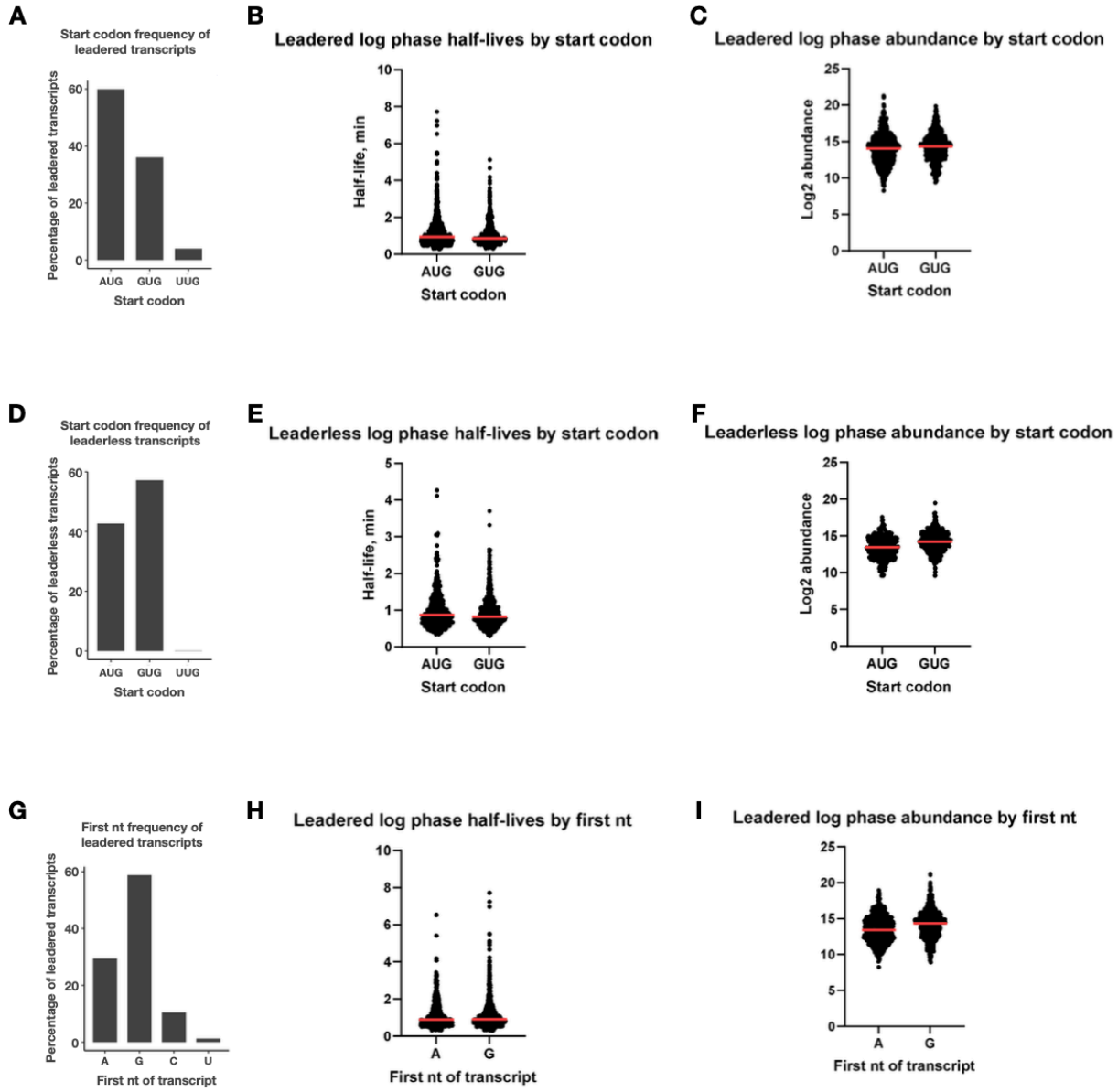



**Figure S 3-6. Comparisons of feature importance rankings.**
In all panels, the colors of the squares in the heatmaps indicate the relative importance rankings derived from the Gini importance rankings for each model. **A-B.** Random forest classifiers were trained using the same set of features, selected by the leaderless model, for leadered and leaderless transcripts in hypoxia (**A**) and using fold-change in half-life from log phase to hypoxia (**B**). For each condition, the top 20 features for the leadered and leaderless gene models were combined and shown in the heatmap. **C-D.** Random forest

classifiers were trained using the complete 5′ UTR features list (**C**) or the non-translation-related 5′ UTR features (**D**) for log phase, hypoxia, and fold-change half-life from log phase to hypoxia. The top 20 features for each condition were combined for the heatmaps shown.




## Figure S 3-7. Transcripts beginning with G appear to be transcribed at higher rates than those beginning with A.

**A** and **D.** Start codon frequencies of leadered (**A**) and leaderless (**D**) transcripts. **B** and **E.** Log phase half-life distributions of transcripts binned by start codons for leadered (**B**) and leaderless (**E**) transcripts. **D** and **F.** Log phase transcript abundance distributions of transcript binned by start codon for leadered (**C**) and leaderless (**F**) transcripts. **G.** Frequencies of leadered transcripts starting with each of the four nucleotides. **H.** Log phase half-life distributions for leadered transcripts starting with A and G. **I.** Los phase transcript abundance distributions for leadered transcripts starting with A and G. Transcripts from genes with UUG

start codons were not included in panels **B, C, E,** and **F** because their frequencies were too low to draw generalizable conclusions.

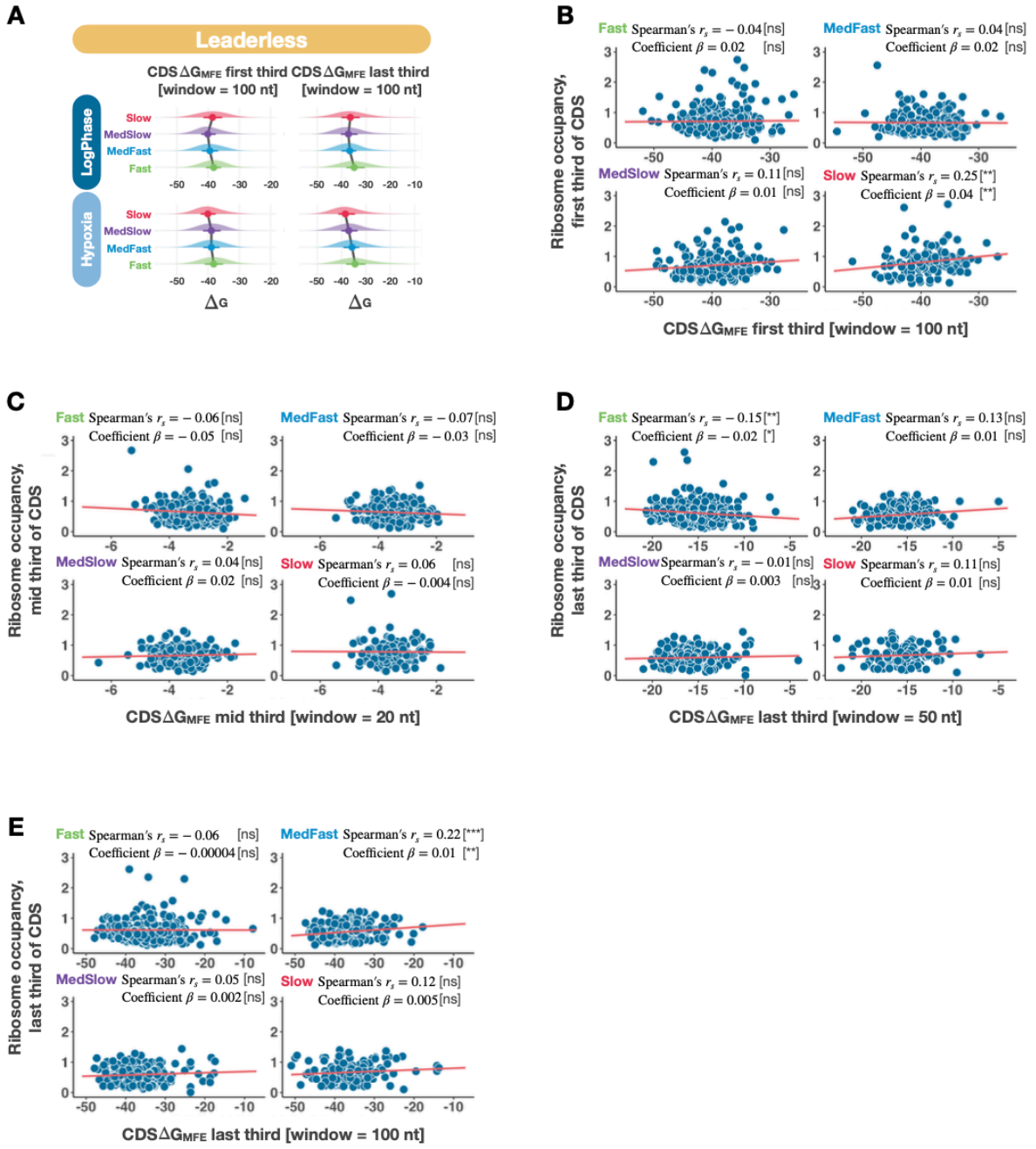



**Figure S 3-8. Correlations between secondary structure, half-life and ribosome occupancy for leaderless transcripts.**

**A.** The $\Delta G_{MFE}$ for the 5′ third and 3′ third of each CDS was calculated using a 100 nt window size. The distributions for each half-life class in log phase and hypoxia are shown. **B-E.** Correlations between $\Delta G_{MFE}$ and ribosome occupancy of the 5′ third, middle third, and 3′ third of each CDS with different window sizes used for $\Delta G_{MFE}$ calculation.

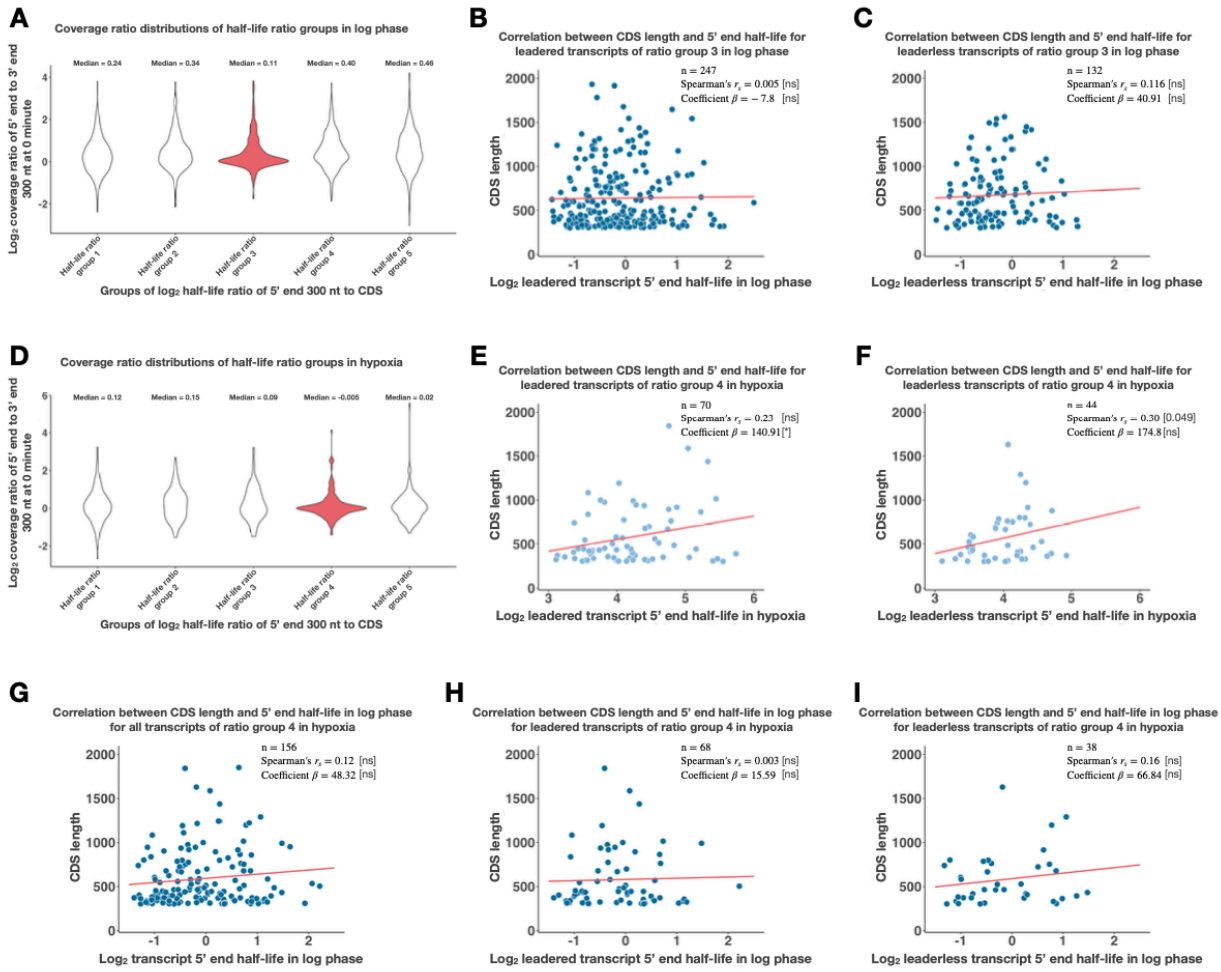



**Figure S 3-9. Correlations between CDS length and 5′ end half-life for selected groups of genes.**

Half-lives were calculated using all reads aligning to each gene (CDS half-life) or using only reads aligning to the 5′ 300 nt of each gene (5′ end half-life). The $\log_2$ ratios of the 5′ end half-life to the CDS half-life were determined, and genes were divided into quintiles based on these ratios. The quintiles are shown in the x axes of **A** and **D**. Additionally, steady-state abundance of reads aligning to the first 300 nt and last 300 nt of each CDS was determined and the $\log_2$ ratio of these abundance values was plotted on the y axes of **A** and **D**. The groups of transcripts highlighted in pink had $\log_2$ ratios of 5′ end to 3′ end abundance near zero and were selected for further analysis. We expect that these groups contain the largest portion of full-length transcripts that cover the entire CDS. **B-C.** Correlations between CDS length and 5′ end half-life for leadered and leaderless transcripts from panel **A** group 3 in log phase. **E-F.** Correlations between CDS length and 5′ end half-life for leadered and leaderless transcripts from panel **D** group 4 in hypoxia. **G-I.** Correlations between CDS length and 5′ end half-life in log phase for transcripts from panel **D** group 4, which was selected based on hypoxia data. Correlations were visualized separately for all transcripts (**G**), leadered transcripts only (**H**), and leaderless transcripts only (**I**) that had half-lives measured in log phase.

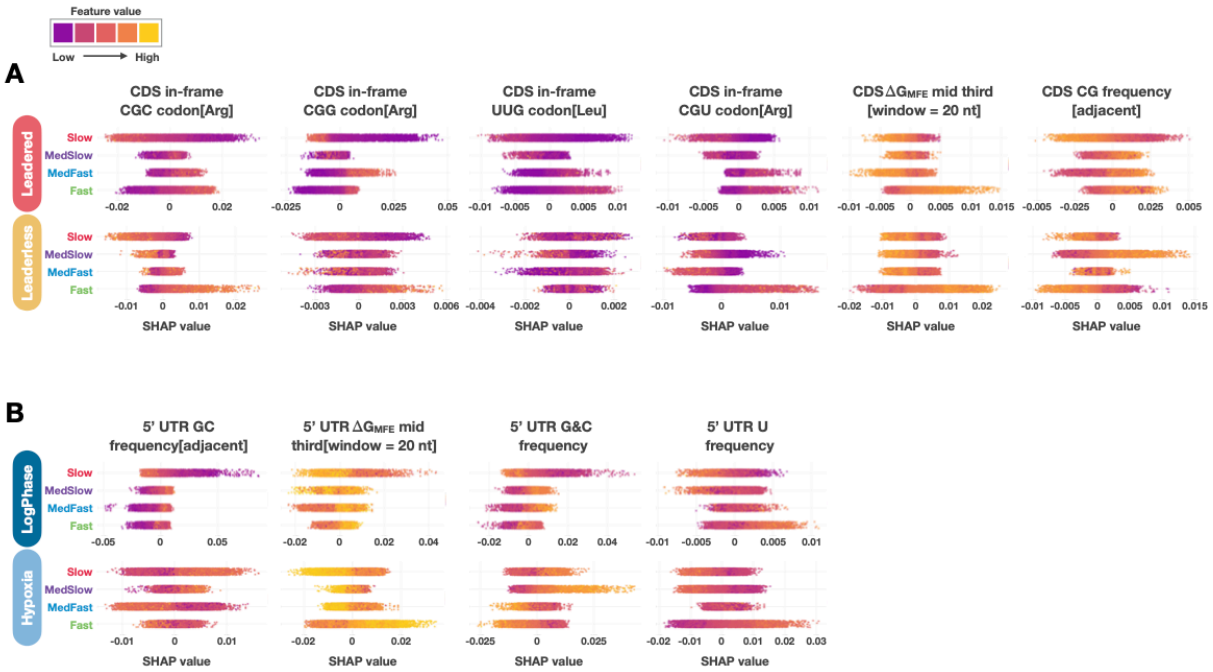



**Figure S 3-10. SHAP value distributions for features in Figure 3-4 B, E.**

SHAP values were extracted and visualized for each half-life class. Feature values were encoded by color. Each dot represents a prediction of transcript half-life class on test set genes during classifier training and evaluation (Materials and Methods). A positive SHAP value corresponds to a positive prediction of the class, while a negative SHAP value means negative prediction of the class.

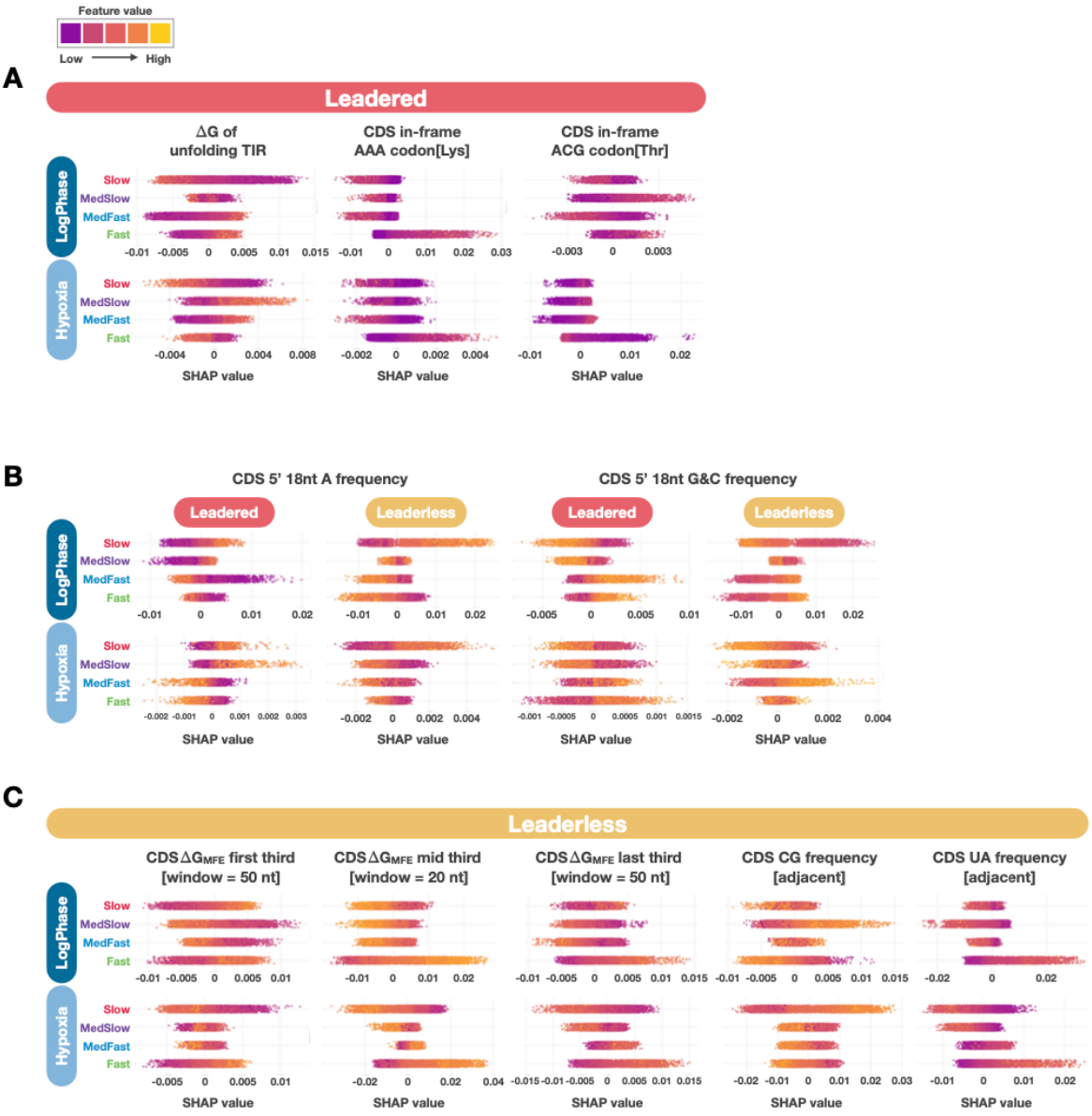



**Figure S 3-11. SHAP value distributions for features in Figure 3-5 B, C, D.**
SHAP values were extracted and visualized for each half-life class. Feature values were encoded by color. Each dot represents a prediction of transcript half-life class on test set genes during classifier training and evaluation (Materials and Methods). A positive SHAP value corresponds to a positive prediction of the class, while a negative SHAP value means negative prediction of the class.

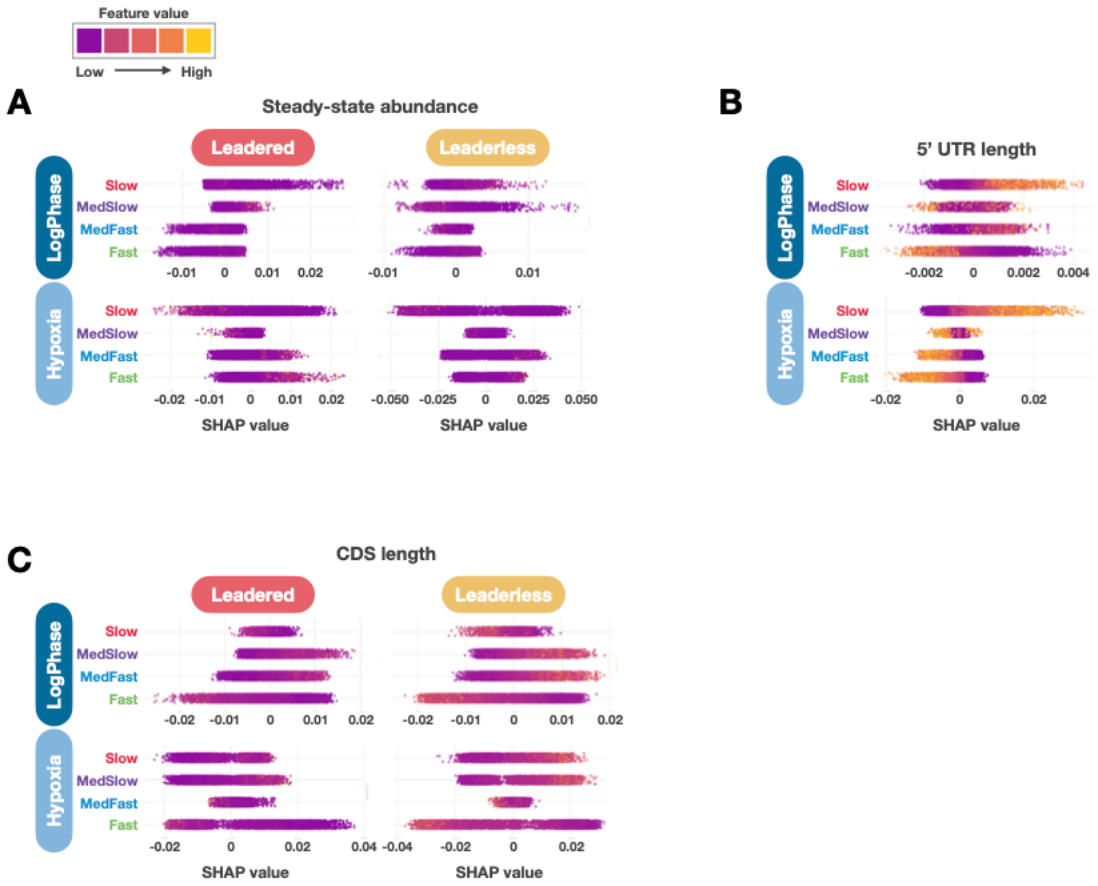



**Figure S 3-12. SHAP value distributions for features in Figure 3-6 A, C, D.**
SHAP values were extracted and visualized for each half-life class. Feature values were encoded by color. Each dot represents a prediction of transcript half-life class on test set genes during classifier training and evaluation (Materials and Methods). A positive SHAP value corresponds to a positive prediction of the class, while a negative SHAP value means negative prediction of the class.

# Chapter 4 : Conclusions and future directions

# Mycobacterial RNase E cleaves with a distinct sequence preference and controls the degradation rates of most *Mycolicibacterium smegmatis* mRNAs

In Chapter 2, we aimed to establish and characterize the broad impact of RNase E on mRNA degradation by investigating the effects of RNase E knockdown in mycobacteria. Among all the ribonucleases encoded in mycobacteria, RNase E is an endoribonuclease that was predicted to play a major role in transcriptome-wide mRNA degradation based on its essentiality (1-3) and known role in *E. coli* (4-8). However, the function of RNase E and its impact on mRNA degradation had not been clearly defined in mycobacteria. To investigate the role of RNase E in mRNA degradation, we constructed a repressible *rne* strain in *Mycolicibacterium smegmatis*. By comparing the transcriptome-wide mRNA half-lives of this strain with control strain, we observed a global stabilization of mRNA in response to *rne* knockdown with the extent differing among mRNAs. The variance in stabilization could be related to activities of other ribonucleases, and their interactions with RNase E itself, which requires further investigation. Interestingly, the amount of increase in half-live seemed to be larger for leadered transcripts than leaderless transcripts, suggesting that degradation mechanisms for leadered vs leaderless transcripts may differ. To identify and characterize the sites of RNA cleavage by RNase E, we adapted a cutting-edge method for analysis of RNA cleavage from standard RNAseq expression libraries. Through a customized pipeline to compare patterns of read coverage between *rne* knockdown and control strains, we found that the RNase E cleavage regions were enriched for cytidines. This allowed us to determine that RNase E was responsible for a set of thousands of mRNA cleavage sites that were previously mapped with high resolution *in vivo* (9) and found to occur primarily with cytidine

in the +1 position. Together, these results indicate that RNase E is responsible for producing most of the cleaved RNAs with monophosphorylated 5′ ends that are present in the *M. smegmatis* transcriptome. We then harnessed existing datasets to confirm that RNase E has the same role and cleavage site preference in *M. tuberculosis* (10). Interestingly, the cleavage site specificity of mycobacterial RNase E differs dramatically from that of *E. coli* RNase E, highlighting the differences in RNA processing and degradation mechanisms in these two organisms.

While the importance of RNase E in mycobacterial mRNA degradation is clear, it is not known how the role of RNase E relates to the roles of other enzymes involved in mRNA degradation. As an endonuclease, RNase E cleaves transcripts into pieces. Exonucleases are then responsible for degrading these pieces into individual nucleotides. The RNAseq libraries we use for measuring mRNA half-lives capture full-length RNAs and fragments larger than roughly 100 nt; thus, we cannot track the fate of smaller RNA fragments through our methods. We do not know if RNase E cleaves mRNAs in one or a few places, triggering rapid degradation by other RNases, or if RNase E itself is responsible for cleaving transcripts into pieces too small to be captured in the RNAseq libraries. We hypothesize that a 3′ to 5′ exoribonuclease called PNPase also makes major contributions to mRNA degradation, based on unpublished preliminary data. Further work is needed to understand the respective roles of RNase E and other RNases in the mRNA degradation process. Such knowledge would be useful for directing future study of the mechanisms by which mRNA degradation is regulated in response to stress.

## Diverse intrinsic properties shape transcript stability and stabilization in *Mycolicibacterium smegmatis*

In Chapter 3, we sought to gain a better understanding of the mechanisms that control mRNA degradation by investigating the influence of transcript properties and microenvironments on transcript half-lives in *Mycolicibacterium smegmatis*. We and others found that transcriptome-wide mRNA degradation profiles in mycobacteria exhibit variance in transcript stability among genes (11,12), and here we extended these findings to show variance in stability of thousands of transcripts within and between growth conditions. However, the transcript features that contribute to such variance was largely unexplored. To investigate the roles and impact of transcript features on mRNA degradation in *M. smegmatis*, we developed an experimental and computational framework combining RNAseq and machine learning. Through quantifying mRNA half-lives in both aerobically growing and hypoxia-arrested *M. smegmatis*, we confirmed the variance in transcript stability among genes and between conditions. This is consistent with the transcriptome-wide half-life profiles previously shown for *M. tuberculosis* during log phase growth. We found that mRNA half-lives were longer for most transcripts in hypoxia-induced growth arrest, consistent with previous reports by us and others of broad mRNA stabilization in energy-stressed bacteria (12-15). Interestingly, the transcripts of essential genes tended to be stabilized more than those of non-essential genes, suggesting that transcript stabilization may be biased towards the most important transcripts. We specifically quantified the associations between transcript features and half-life under the combinations of leader status and environmental conditions. We found that in contrast to previously reports for *M. tuberculosis* and *E. coli* suggesting that transcription rate was the dominant determinant of mRNA degradation rates (12,16-18), a wide range of

transcript features contribute to observed variance in degradation rates. The interactions of these features was non-linear, and the importance of various features for model performance differed between conditions as well as between leadered and leaderless genes. While some of the broad themes that emerged were easily explained, such as translation protecting transcripts from degradation specifically in log phase growth, others were more complicated or surprising. For example, we found that codon content had influences that differed by leader status and condition and did not appear to be fully explained by translation efficiency. We also found that coding sequence length positively correlated with mRNA half-life specifically in hypoxia, which to our knowledge has not been previously reported for any bacteria.

## Proposed future direction based on the results derived here

We demonstrated the influence of diverse transcript features on mRNA stability in Chapter 3, as well as the broad impact of RNase E on mRNA degradation rates in Chapter 2. Our machine learning models were powerful for revealing the importance of specific mRNA features for degradation, but fell short of fully explaining the observed mRNA degradation rates. We wondered if transcript stability can be better explained by a model combining the effects of transcript features with the known RNase E cleavage sequence preferences. We hypothesize that the additional information of RNase E cleavage preferences could potentially improve the current suboptimal performance of machine learning models to predict transcript stability. Through our work in Chapter 2, we have established the transcriptome-wide map of RNase E cleavage sites. However, there are still unresolved questions. Although our results suggested that the impact of sequence surrounding cleavage sites was limited, we also observed a greater than expected

frequency of cleavage sites in untranslated regions. These results indicate that there could be a potential preference of these cleavage events determined by other more complicated sequence signatures of the target transcripts and/or by binding of ribosomes. Those sequence signatures could also influence interactions and synergies between RNase E and other RNases, which are largely unknown in mycobacteria. All of these could potentially affect the predictions of transcript stability.

To address these problems, we propose in future work to first quantify the time course profile of cleavage events. This can be achieved by using our current half-life RNAseq libraries to quantify the abundance of local transcript regions, such as each 5% of each transcript, over the degradation time course. By comparing the abundance of each of these transcript regions over the degradation time course, we can potentially identify the regions in which cleavage events occurred first. These cleavage orders can be further confirmed through the time course profile of RNase E or other ribonuclease knockdown strains. Then we can create groups of transcripts with similar cleavage event orders. Finally, we can develop machine learning models using our current compendium of mRNA properties to predict transcript stability within each of these groups. Eventually these results could enhance our understanding of mRNA degradation mechanisms by elucidating the activities of ribonucleases using more complicated transcript features.

# Reference


1. Sassetti, C.M., Boyd, D.H. and Rubin, E.J. (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol*, **48**, 77-84.

2. Sassetti, C.M. and Rubin, E.J. (2003) Genetic requirements for mycobacterial survival during infection. *Proc Natl Acad Sci U S A*, **100**, 12989-12994.

3. Taverniti, V., Forti, F., Ghisotti, D. and Putzer, H. (2011) Mycobacterium smegmatis RNase J is a 5'-3' exo-/endoribonuclease and both RNase J and RNase E are involved in ribosomal RNA maturation. *Mol Microbiol*, **82**, 1260-1276.

4. Babitzke, P. and Kushner, S.R. (1991) The Ams (altered mRNA stability) protein and ribonuclease E are encoded by the same structural gene of Escherichia coli. *Proc Natl Acad Sci U S A*, **88**, 1-5.

5. Carpousis, A.J., Van Houwe, G., Ehretsmann, C. and Krisch, H.M. (1994) Copurification of E. coli RNAase E and PNPase: evidence for a specific association between two enzymes important in RNA processing and degradation. *Cell*, **76**, 889-900.

6. Lin-Chao, S., Wong, T.T., McDowall, K.J. and Cohen, S.N. (1994) Effects of nucleotide sequence on the specificity of rne-dependent and RNase E-mediated cleavages of RNA I encoded by the pBR322 plasmid. *J Biol Chem*, **269**, 10797-10803.

7. McDowall, K.J., Lin-Chao, S. and Cohen, S.N. (1994) A+U content rather than a particular nucleotide order determines the specificity of RNase E cleavage. *Journal of Biological Chemistry*, **269**, 10790-10796.

8. Py, B., Higgins, C.F., Krisch, H.M. and Carpousis, A.J. (1996) A DEAD-box RNA helicase in the Escherichia coli RNA degradosome. *Nature*, **381**, 169-172.

9. Martini, M.C., Zhou, Y., Sun, H. and Shell, S.S. (2019) Defining the Transcriptional and Post-transcriptional Landscapes of Mycobacterium smegmatis in Aerobic Growth and Hypoxia. *Front Microbiol*, **10**, 591.

10. Plocinski, P., Macios, M., Houghton, J., Niemiec, E., Plocinska, R., Brzostek, A., Slomka, M., Dziadek, J., Young, D. and Dziembowski, A. (2019) Proteomic and transcriptomic experiments reveal an essential role of RNA degradosome complexes in shaping the transcriptome of Mycobacterium tuberculosis. *Nucleic Acids Res*, **47**, 5892-5905.

11. Zhou, Y., Sun, H., Rapiejko, A.R., Vargas-Blanco, D.A., Martini, M.C., Chase, M.R., Joubran, S.R., Davis, A.B., Dainis, J.P., Kelly, J.M. *et al.* (2023) Mycobacterial RNase E cleaves with a distinct sequence preference and controls the degradation rates of most Mycolicibacterium smegmatis mRNAs. *J Biol Chem*, **299**, 105312.

12. Rustad, T.R., Minch, K.J., Brabant, W., Winkler, J.K., Reiss, D.J., Baliga, N.S. and Sherman, D.R. (2013) Global analysis of mRNA stability in Mycobacterium tuberculosis. *Nucleic Acids Res*, **41**, 509-517.

13. Vargas-Blanco, D.A., Zhou, Y., Zamalloa, L.G., Antonelli, T. and Shell, S.S. (2019) mRNA Degradation Rates Are Coupled to Metabolic Status in Mycobacterium smegmatis. *mBio*, **10**.

14. Morin, M., Enjalbert, B., Ropers, D., Girbal, L. and Cocaign-Bousquet, M. (2020) Genomewide Stabilization of mRNA during a "Feast-to-Famine" Growth Transition in Escherichia coli. *mSphere*, **5**.

15. Anderson, K.L., Roberts, C., Disz, T., Vonstein, V., Hwang, K., Overbeek, R., Olson, P.D., Projan, S.J. and Dunman, P.M. (2006) Characterization of the Staphylococcus aureus heat shock, cold shock, stringent, and SOS responses and their effects on log-phase mRNA turnover. *J Bacteriol*, **188**, 6739-6756.

16. Bernstein, J.A., Khodursky, A.B., Lin, P.H., Lin-Chao, S. and Cohen, S.N. (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci U S A*, **99**, 9697-9702.

17. Esquerre, T., Moisan, A., Chiapello, H., Arike, L., Vilu, R., Gaspin, C., Cocaign-Bousquet, M. and Girbal, L. (2015) Genome-wide investigation of mRNA lifetime determinants in Escherichia coli cells cultured at different growth rates. *BMC Genomics*, **16**, 275.

18. Nouaille, S., Mondeil, S., Finoux, A.L., Moulis, C., Girbal, L. and Cocaign-Bousquet, M. (2017) The stability of an mRNA is influenced by its concentration: a potential physical mechanism to regulate gene expression. *Nucleic Acids Res*, **45**, 11711-11724.