bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Title:

A combinatorial model predicts leaderless mRNA start codon selection in *C. crescentus*

Authors:

Mohammed-Husain M. Bharmal (orcID: 0000-0002-2365-3034), Jared M. Schrader*

(orcID:0000-0002-5728-5882)

Affiliations:

Department of Biological Sciences, Wayne State University, Detroit, MI, 48202, USA

*Corresponding Author (Schrader@wayne.edu)

Abstract

Bacterial translation is thought to initiate by base-pairing of the 16S rRNA and the Shine-Dalgarno sequence in the mRNA's 5′ UTR. However, transcriptomics has revealed that leaderless mRNAs, which completely lack any 5′ UTR, are broadly distributed across bacteria and can initiate translation in the absence of the Shine-Dalgarno sequence. To investigate the mechanism of leaderless mRNA translation initiation, synthetic *in vivo* translation reporters were designed that systematically tested the effects of start codon accessibility, leader length, and start codon identity on leaderless mRNA translation initiation. Using this data, a simple computational model was built based on the combinatorial relationship of these mRNA features which can accurately classify leaderless mRNAs and predict the translation initiation efficiency of leaderless mRNAs. Thus, start codon accessibility, leader length, and start codon identity combine to define leaderless mRNA translation initiation in bacteria.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Translation initiation is a critical step for fidelity of gene expression in which the ribosome initiation complex is formed on the start codon of the mRNA. Since the canonical start codon, AUG, compliments both initiator and elongator methionyl-tRNAs, the ribosome must distinguish the start AUG codon from elongator AUG codons. Incorrect initiation at an elongator AUG can lead to non-functional products that can be detrimental to cellular fitness (1-3). Canonical start codon selection is thought to occur by the base-pairing of the 16S rRNA with a Shine-Dalgarno (SD) sequence in the mRNA located 5nt upstream of the start codon (4-6). The base pairing between the 16S rRNA and mRNA was shown to be critical for initiation since mutation of the anti-SD (aSD) in the 16S rRNA is lethal (7), and translation of a gene lacking a canonical SD sequence could be restored when the 16S of the rRNA were mutated to a complimentary sequence (8). While the SD-aSD pairing clearly impacts translation initiation efficiency (TIE) in *E. coli*, other studies have found that the SD:aSD interaction is not essential for correct selection of the start codon (9,10). Indeed, "orthogonal" ribosomes with altered 16S rRNA aSD sequences were found to initiate at the normal start codons throughout the transcriptome (11). Interestingly, *E. coli* lacks SD sites within its genome in approximately 30% of its translation initiation regions (TIRs) with other species of bacteria containing SD sites in as few as 8% of their TIRs (12,13). Indeed, RNA-seq based transcription mapping experiments have found that many bacterial mRNAs are "leaderless" and begin directly at the AUG start codon (14-16), and that these mRNAs are abundant in pathogens such as *M. tuberculosis* and in the mammalian mitochondria (17)..

To account for the lack of essentiality of the SD site, a "Unique accessibility model" was proposed which posited that start codon selection occurs due to the TIR being accessible to

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

initiating ribosomes, while elongator AUGs are physically inaccessible due to RNA secondary structures (18). This model was based upon a strong negative correlation observed between mRNA secondary structure content in the TIR and TIE (19-21). This model is further supported by genomic analysis of RNA secondary structure prediction of mRNA TIRs in which there's a lower amount of secondary structure in the TIR compared to elongator regions, which is conserved across all domains of life (22). While the unique accessibility model is overly simplistic, more advanced computational approaches have been able to combine TIR accessibility with SD strength, spacing, and standby sites to more accurately predict TIE of leadered mRNAs (23). While TIR accessibility has been shown to be critical in many leadered mRNAs, it has not yet been systematically tested for leaderless mRNAs.

Genome-wide RNA-seq transcript mapping experiments have revealed that leaderless mRNAs are widespread across bacteria (14), yet little is known about their mechanism of translation initiation. While very few leaderless mRNAs has been identified in *E. coli* (0.7% leaderless mRNAs (24)), other bacteria and archaea contain a large majority of their transcripts as leaderless mRNAs (up to 72% leaderless mRNAs (14,25)). Additionally, sizeable proportions of leaderless mRNAs have been identified in bacteria of clinical significance, such as *Mycobacterium tuberculosis*, and of industrial significance like *Corynebacterium glutamicum* (15,26). In the model bacterium *Caulobacter crescentus* approximately 17% of mRNAs are leaderless (27), with the fastest doubling time of any bacterium with large numbers of leaderless mRNAs. In addition, *C. crescentus* has good genetic tools, making it an ideal model to study translation initiation of leaderless mRNAs.

Importantly, the role of TIR accessibility has not been systematically tested for leaderless mRNAs, however, some aspects of their initiation have been identified which are distinct from

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

leadered mRNAs. Mitochondrial leaderless mRNAs have been found to lack 5′ secondary structure (28), in support of a TIR accessibility model. In opposition to the canonical initiation mechanism, leaderless mRNAs can initiate with 70S ribosomes where IF2 is known to stimulate their translation, and IF3 can inhibit leaderless translation (29,30). Additionally, AUG is the most efficient start codon in leaderless mRNAs, and near-cognate start codons GUG and UUG are initiated with much lower efficiency (31-35). Suppressor tRNAs could restore initiation on non-AUG codons for leadered RNAs, but not for leaderless RNAs (31), suggesting that for leaderless mRNAs an AUG start codon has unique initiation properties independent of perfect codon-anticodon base-pairing. Indeed, genomic prediction of leaderless mRNAs suggests a very high preference of AUG (79%) at the 5′ end of leaderless mRNAs; with a smaller percentage of GUG (10%), UUG (6%) and others (3%) (13). In addition to the start codon identity, TIE of mRNAs with short leaders (<5nt) is significantly lower as compared to their fully leaderless counterparts (33,34,36-38). Altogether, this suggests that leaderless mRNAs strongly prefer AUG and are inhibited by having short leaders.

In order to understand the mRNA sequence features needed for leaderless translation initiation, we systematically measured the effect of TIR accessibility, start codon identity, and leader length on leaderless mRNA translation initiation in *C. crescentus.* Using synthetic *in vivo* translation initiation reporters, we show that TIR accessibility, start codon identity, and leader length all dramatically affect leaderless mRNA TIE. The dependencies of each mRNA feature on TIE were then built into a simple computational model ($TIE_{leaderless}$ model) that accurately predicts which RNAs in the *C. crescentus* transcriptome would be initiated as leaderless RNAs with an AUC of a Receiver Operator Characteristic (ROC) curve of 0.99. The $TIE_{leaderless}$ model also accurately predicts the translation initiation efficiency of *in vivo* leaderless mRNA reporters

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(R^2=0.87) and to a lesser extent the overall translation efficiency of leaderless mRNAs measured by ribosome profiling (R^2=0.44). This therefore provides the first systematic analysis of mRNA features required for leaderless initiation and the *C. crescentus* $TIE_{leaderless}$ model will likely provide a foundation for our understanding of leaderless mRNA translation initiation across bacteria.

Materials and Methods

Computational predictions of start codon accessibility

Retrieving transcript sequences

All the RNA sequences were retrieved from transcription start sites and translation start site data available from RNA-seq and ribosome profiling respectively (27,40) using the *C. crescentus* NA1000 genome sequence (39). The TIR sequences were then extracted from all ORFS using 50 bases (25 bases upstream of start codon and 25 bases downstream from start codon). If the 5′ upstream untranslated region (UTR) was less than 25 bases, then 50 bases from transcription start site was used for all TIR calculations.

Calculation of ΔG_{unfold}

Start codon accessibility was computed similar to (41) by comparing the native TIR RNA structure (ΔG_{mRNA}) to that of the same TIR bound by an initiating ribosome (ΔG_{init}). Since ribosome binding requires a single-stranded region of the mRNA we approximated this by forcing the TIR to be single stranded. The overall calculation was performed in three steps:

1. Calculation of ΔG_{mRNA}:

   The minimum free energy (mfe) labelled as ΔG_{mRNA} was calculated using RNAfold web server of the Vienna RNA websuite (42) at 28°C by inputting all the TIR sequences in a

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

text file using command line function 'RNAfold --temp 28 <input_sequences.txt >output.txt'. The output file was in the default RNAfold format with each new sequence on one line followed by dot-bracket notation (Vienna format) in the next line. RNAstructure (43) was used to generate ct files for each of the mfe structures predicted in RNAfold which contained all the base pair indexes for each sequence.

2. Calculation of ΔG_{init}:

   The base pairs in the TIR (from up to 12 bases upstream of the start codon to 13 bases downstream of the start codon) were broken and forced to be single stranded including any pairs formed from the TIR and outside. If the 5′UTR length was more than or equal to 25 bases, then the RBS was selected from -12 to +13 bases (25 bases). If the 5′UTR length was less than 25, then the TIR comprised of the entire 5′UTR to +13 bases. A new dot bracket file with these base-pairing constraints was then used in the RNAfold program (42) with the same RNA sequence to calculate the ΔG_{init}.

3. Calculation of ΔG_{unfold}:

   Lastly, ΔG_{unfold} was calculated by subtracting ΔG_{mRNA} (mfe of mRNA in native state) from ΔG_{init} (mfe of mRNA after ribosome binding) (eq 1. $\Delta G_{unfold} = \Delta G_{init} - \Delta G_{mRNA}$).

Cell growth and media

E. coli culture

For cloning, plasmids with the reporter gene were transformed in *E. coli* top10 competent cells using heat shock method for 50-55 secs at 42°C. Luria-Bertani (LB) liquid media was used for outgrowth and the colonies were plated on LB/kanamycin (50 µg/mL) agar plates.

For miniprep, the *E.coli* cultures were inoculated overnight(O/N) in liquid LB/kanamycin (30 µg/mL).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

C. crescentus culture

For cloning, plasmids were transformed in NA1000 *C. crescentus* cells after sequence verification using electroporation. The *C. crescentus* NA1000 cells were grown in Peptone Yeast Extract (PYE) liquid medium. After transformation, for the outgrowth liquid PYE medium was used (2mL) and then plated on PYE/kanamycin (25 μg/mL) agar plates. For imaging, the *C. crescentus* culture were grown O/N at different dilutions in liquid PYE/kanamycin (5 μg/mL). Next day, the cultures growing in log phase were diluted and induced in liquid PYE with kanamycin (5 μg/mL) and Xylose (final concentration of 0.2%) such that the optical density (OD) was around 0.05 to 0.1.

Design and generation of translation reporters

Oligos and plasmid design

For the design and generation of reporter assay, a plasmid with a reporter gene (yellow fluorescent protein (YFP)), under the control of an inducible xylose promoter was used. The pBYFPC-2 plasmid containing the kanamycin resistant gene was originally generated from (44). A list of oligos used for generating plasmids with different 5′ UTRs of YFP is attached as a supplementary table (Table S4).

Inverse PCR mutagenesis and Ligation

The 5′UTR region and start codon of the YFP reporter protein was replaced with other TIR sequences. This was done by inverse PCR, in which the leaderless TIR is attached to the reverse primer as an overhang. Initial denaturation was done at 98°C for 5 mins. Followed by 30 cycles of denaturation at 98°C for 10 secs, annealing at 60°C for 10 secs and extension at 72°C for 7 mins and 20 secs. After 30 cycles, final extension was done at 72°C for 5 mins. The polymerase used was Phusion (Thermoscientific 2 U/μL). The PCR product was then DPNI

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

treated to cut the template DNA using DPNI enzyme (Thermoscientific 10 U/µL). The DPNI treated sample was then purified using Thermo fisher GeneJET PCR Purification kit. The purified sample (50 ng) was then used for blunt end ligation using T4 DNA Ligase (Thermoscientific 1 WeissU/µL).

Transformation in *E. coli* cells

5 µL of the ligation reaction was then added to 50 µL of *E. coli* top10 competent cells. Then the mixture was incubated in ice for 30 mins. Then heat shocked for 50-55 secs in the water bath at 42°C. Then immediately kept in ice for 5 mins, after which 750 µL of LB liquid medium was added to the cells for outgrowth and kept for incubation at 37°C for 1 hr at 200 rpm. After this, 200-250 µL of the culture was plated on LB/kanamycin (50 µg/mL) agar plates.

Colony screening and sequence verification

The colonies grown on LB/kanamycin plates were screened by colony PCR to first screen for the presence of the new TIR insert. The cloning results in the replacement of the larger 5′UTR region of YFP with a smaller region containing a leaderless TIR, thus distinguished easily on an analytical gel. The forward and reverse primer used for the screening results in approximately 180 base pairs, whereas the original fragment amplified with the same oligos is 245 bases. The forward oligo used was pxyl-for: cccacatgttagcgctaccaagtgc and reverse oligo is eGYC1: gtttacgtcgccgtccagctcgac. Upon verification, a small aliquot (4 µL) of the colony saved in Taq polymerase buffer was inoculated in 5 mL of liquid LB/kanamycin (30 µg/mL) and incubated overnight at 37°C at 200 rpm. Next day, the culture was miniprepped using Thermo fisher GeneJET Plasmid Miniprep kit. The concentration of DNA in the miniprepped samples were measured using Nanodrop 2000C from Thermoscientific. DNA samples were sent to

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genewiz for sanger sequencing to verify the correct insert DNA sequences using the DNA primer eGYC1: gtttacgtcgccgtccagctcgac (44).

Transformation in *C. crescentus* NA1000 cells

After the sequences were verified, the plasmids were transformed in *C. crescentus* NA1000 cells. For transformation, the NA1000 cells were grown overnight at 28°C in PYE liquid medium at 200rpm. The next day, 5 mL of cells were harvested for each transformation, centrifuged and washed three times with autoclaved milliQ water. Then, 1 µL of sequence verified plasmid DNA was mixed with the cells and electroporated using Bio-Rad Micropulser (program Ec1 set at voltage of 1.8 kV). Then, the electroporated cells were immediately inoculated in 2 mL of PYE for 3 hours at 28°C at 200rpm. Then 10-20 µL of culture was plated on PYE/ kanamycin agar plates. Kanamycin-resistant colonies were grown in PYE/kanamycin media overnight and then stored as a freezer stock in the -80°C freezer

Cellular assay of translation reporters

C. crescentus cells harboring reporter plasmids were serially diluted and grown overnight in liquid PYE/kanamycin medium (5 µg/mL). The next day, cells in the log phase were diluted with fresh liquid PYE/kanamycin (5 µg/mL) to have an optical density (OD) of 0.05-0.1. The inducer xylose was then added in the medium such that the final concentration of xylose is 0.2%. The cells were grown for 6 hours at 28°C at 200 rpm. After this, 2-5 µL of the cultures were spotted on M2G 1.5% agarose pads on a glass slide. After the spots soaked into the pad, a coverslip was placed on the pads and the YFP level was measured using fluorescence microscopy using a Nikon eclipse NI-E with CoolSNAP MYO-CCD camera and 100x Oil CFI Plan Fluor (Nikon) objective. Image was captured using Nikon elements software with a YFP filter cube with exposure times of 30ms for phase-contrast images and 300 ms for YFP images

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

respectively. The images were then analyzed using a plugin of software ImageJ (45) called MicrobeJ (46).

Three component model calculations and leader length/identity analysis

For all RNA transcripts in the *C. crescentus* genome identified in (27,40), we computed their capacity to initiate as a leaderless mRNA using equation 2: ($TIE_{Leaderless\ mRNA(k)}$ = Max TIE (1) - (1-$TIE_{\Delta Gunfold}$) - (1-$TIE_{start\ codon\ identity(j)}$) – (1-$TIE_{leader\ length(i)}$) where k= a given RNA transcript, j=start codon identity, and i=leader length(nt). To identify putative leaderless mRNA TIRs, we first asked if the 5′ end contained an AUG or near cognate start codon, and if not we scanned successively from the 5′ end for AUG trinucleotides within the first 8 nt. Near cognate start codons were omitted from positions containing leader nucleotides since AUG codons yielded higher TIE values even in the presence of a leader. We next asked if there is an AUG or near cognate start codon further downstream by scanning 5′ to 3’ through the first 18 nt. If found, we calculated $TIE_{leaderless\ mRNA}$ with all different possible cognate/near-cognate start codons along the TIR. Then of all the different possibilities, the one having the highest $TIE_{leaderless}$ score was selected for further analysis (Fig 7A).

To utilize $TIE_{leaderless\ mRNA}$ for classification, each RNA was then categorized into two different classes based on 5′ end sequencing data and ribosome profiling based global assays ((27,40))**:** true leaderless – RNAs that are known to initiate directly at a 5′ start codon, and false leaderless – RNAs that are not initiated at a 5′ start codon. A small subset was classified as “unknown”, as they contain very short leaders and lack SD sites, making their mode of translation initiation ambiguous. Using these $TIE_{leaderless\ mRNA}$ values, a ROC curve was plotted using scikit-learn library in python (47) with the “true leaderless” and “false leaderless” RNAs ($TIE_{leaderless\ mRNA}$ values for the *C. crescentus* transcriptome can be found in Table S1).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To utilize $TIE_{leaderless\ mRNA}$ for prediction of translation initiation reporter levels, we first converted all negative $TIE_{leaderless\ mRNA}$ scores to zero. Next, we compared the $TIE_{leaderless\ mRNA}$ scores to the YFP levels of the translation initiation and performed a linear regression calculation using the linest function in microsoft excel and libreoffice calc. For prediction of native leaderless mRNA translation levels, TE measurements from ribosome profiling experiments (27) were compared to the $TIE_{leaderless\ mRNA}$ scores.

Results

Computational prediction of *C. crescentus* start codon Accessibility

To assess the role of mRNA accessibility across mRNA types, ΔG_{unfold} calculations were performed on all *C. crescentus* translation initiation regions (TIRs). ΔG_{unfold} represents the amount of energy required by the ribosome to unfold the mRNA at the translation initiation region (TIR) and has been identified as a metric that correlates with translation efficiency in *E. coli* (41). ΔG_{unfold} was calculated for all TIRs by first predicting the minimum free energy of the 50 nt region of the mRNA (ΔG_{mRNA}) around the start codon using RNAfold (42). ΔG_{init} was then calculated in which the TIR (25nt surround the start codon), roughly equivalent to a ribosome footprint, was constrained to be single stranded to approximate accessibility for the ribosome to initiation. ΔG_{unfold} was then calculated using equation 1 (eq 1. $\Delta G_{unfold} = \Delta G_{init} - \Delta G_{mRNA}$) which represents the energy required to open the TIR to facilitate translation initiation (Fig 1A). ΔG_{unfold} calculations were performed on all the CDSs in the genome (Fig 1B) and classified into mRNA types based on transcriptome and ribosome profiling maps of the *C. crescentus* genome (27). The transcripts were categorized into two major classes: leaderless (no 5′ UTR) and

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

leadered (those containing a 5′ UTR). Leadered mRNAs were further categorized into subclasses based upon the presence of the Shine-Dalgarno (SD) sequence (27). Shine-Dalgarno (SD) (containing a SD sequence in the 5′ UTR) and nonSD (lacking an SD sequence in the 5′ UTR). Since it is also known that some polycistronic operons reinitiate translation between CDSs without dissociation of the ribosomal subunits, we also examined the ΔG_{unfold} of TIRs occurring downstream of the first CDS in polycistronic mRNAs (Operons). The average ΔG_{unfold} value of leaderless mRNAs (5.6 kcal/mol) was significantly lower than SD (11.9 kcal/mol, p= 1.5E-105), nonSD (10.3 kcal/mol, p= 6.9E-71) and internal operon TIRs (13.2 kcal/mol,p= 1.1E-143) as calculated by pairwise 2-sided T-tests with unequal variance (Fig 1B). The lower ΔG_{unfold} values of nonSD TIRs may be due to the loss of stabilization of TIRs from base pairing between the anti-SD site in the 16S rRNA and the SD site in the mRNA. We also observed that average ΔG_{unfold} of nonSD TIRs was significantly lower than SD TIRs (p= 1.8E-14) and operon TIRs (p=1.4E-44). The difference between the average ΔG_{unfold} of SD and operon genes was also significant (p=2.1E-09). Since the ribosome is an efficient RNA helicase, it is possible that the increased ΔG_{unfold} of operon TIRs may be tolerated by the ribosome's ability to unwind such structures when terminating on the previous CDSs. Since leaderless mRNAs showed a significantly lower ΔG_{unfold}, and lack complexities associated with leadered mRNAs, such as SD or standby sites which are important for leadered initiation (23), we further explored the functional role of ΔG_{unfold} in leaderless mRNAs.

Systematic analysis of *C. crescentus* leaderless mRNA TIR determinants using *in vivo* translation reporters

Leaderless mRNAs initiation is known to be strongly influenced by addition of nucleotides prior to the start codon (leader nts) and by start codon identity (31-38); however, the

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

role of TIR accessibility has been poorly described in this class of mRNAs. To understand the role of these three mRNA features we systematically tested each feature using *in vivo* leaderless mRNA translation initiation reporters. Translation initiation reporters were designed in which the start codon of plasmid pBXYFPC-2 was replaced with an AUG fused directly to the +1 nt of the xylose promoter (44). An additional 15-24 nt after the 5′ AUG was added to allow complete replacement of the 5′ leader and start codon in pBXYFPC-2 with a leaderless TIR. Since only the first 6-9 codons are altered across leaderless mRNA mutants, and the vast majority of the YFP CDS is unaltered, this allows a sensitive system to measure changes in translation initiation. As leaderless TIR mutants may also alter the amino acid sequence, additional care was also taken to ensure that mutations would not alter the N-end rule amino acid preferences of the resulting proteins (48). Using this *in vivo* translation initiation system, we generated three different sets of leaderless TIR reporters to test the effect of ΔG_{unfold}, start codon identity, and additional leader length on *C. crescentus* translation initiation.

As leaderless mRNAs were predicted to have TIRs with low ΔG_{unfold} values, we engineered several RNA hairpins in the TIR to assess the role of ΔG_{unfold} on translation initiation (Fig 2A). Since very few natural *C. crescentus* mRNAs contained RNA structure content in their TIRs (Fig 1B), six synthetic hairpins were designed, varying in stem and loop sizes (Table S1). Into each construct, we also introduced synonymous codon mutations designed to alter the secondary structure content, yielding a range of ΔG_{unfold} values without altering the amino acid sequence within a given hairpin (Table S2). Importantly, the entire range of of ΔG_{unfold} values across the synthetic hairpins spans the entire range calculated for natural leaderless mRNAs (Fig 1, Table S2). For all hairpins, we observed that lowering ΔG_{unfold} and thereby increasing the accessibility of the start AUG led to an increase in the level of YFP production (Fig 2B). Since,

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6/7 of the hairpin mutant sets showed a relationship in which hairpin codon usage frequency positively correlated with ΔG_{unfold} (Table S2), it is most likely that the observed reduction in YFP reporter levels is a result of increased structure content and is not likely to be caused by faster elongation of common codons in the TIR. Additionally, across all mutant hairpins sets generated, we observed a strong negative correlation between the YFP reporter level and the ΔG_{unfold} across a vast range of values with a linear correlation R^2 value of 0.84 (Fig 2B). These data suggest that accessibility of the start codon is a critical feature for leaderless mRNA translation initiation.

Next, we systematically tested the effect of the start codon identity on the *in vivo* translation initiation reporters. In *C. crescentus*, natural leaderless mRNAs initiate with an AUG, GUG, or UUG start codon (27,40). Since its well established that start codon identity can affect leaderless mRNA translation initiation (31-35) we generated variants with different start codon identities. Here, AUG was mutated to other near cognate start codons GUG, CUG, UUG, AUC, AUU, AUA which are known to be the start codons of other leadered mRNAs in *C. crescentus* (27). We also included a non-cognate GGG codon as a negative control since no GGG start codons are known to occur in *C. crescentus*. The results showed that replacing the original AUG codon with any of the other near cognate codons drastically decreased the translation initiation reporter levels, while the GGG codon yielded the lowest translation initiation reporter levels (Fig 3). These data show that the AUG triplet is by far the preferred start codon for *C. crescentus* leaderless mRNAs.

Finally, we systematically tested the role of additional leader length on *C. crescentus* leaderless mRNAs. In *E. coli*, even a single nucleotide before the AUG is known to inhibit initiation of leaderless mRNAs (36). To test if *C. crescentus* leaderless mRNAs were negatively impacted by leader nucleotides we generated a set of reporters with 0, 1, 2, 3, 5, 10, or 20 5′

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Adenosines before the AUG start codon (Fig 4). An A-rich sequence was chosen as it lacks any possible SD sites and is unlikely to form secondary structure, and ΔG_{unfold} values were not altered upon addition of these 5′ bases to the leaderless translation initiation reporter (Table S1). Across this set of mutants, additional nucleotides showed a strong decrease in translation initiation reporter levels with increasing leader length (Fig 4). The translation initiation reporter levels dropped by approximately 2-fold for each additional A that was added to the 5′ end ($TIE_{leader\ length} = 0.45 \times i^{-0.91}$, R^2=0.92, i=leader length (nt)). This confirms that even a short leader can lead to a significant reduction in translation initiation of *C. crescentus* leaderless mRNAs.

Leaderless mRNA TIR determinants affect translation efficiency of natural leaderless mRNAs

Because the *in vivo* translation initiation reporters were all synthetic constructs, we explored the extent to which each mRNA feature (ΔG_{unfold}, start codon identity, and leader length) occur in natural *C. crescentus* leaderless mRNAs. As noted previously, ΔG_{unfold} is significantly lower for leaderless mRNAs than for other mRNA types (Fig 1B). To analyze the role of start codon selection, we calculated the fraction of AUGs at the 5′ end of all *C. crescentus* leaderless mRNAs and of the random chance of finding each start codon based on the genomes' GC percentage. This analysis revealed a strong enrichment of AUGs at the 5′ end of *C. crescentus* leaderless mRNAs as compared to random, and a slight enrichment of the GUG near cognate start codons (Fig 5A). Of all the leaderless mRNAs, only 4.4% (17/385) are initiating with non-AUG start codons as compared to the leadered mRNAs of which 27.23% (989/3632) of genes initiate with non-AUG start codons (Table S3). Since these near cognate start codons were translated much more poorly than AUG in our translation initiation reporters, it's possible that for leaderless mRNAs there's a positive selection for the AUG start codon and a negative

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

selection for near-cognate start codons. Additionally, by exploring the length of mRNAs, we noticed that there was a much greater occurrence of leaderless mRNAs than mRNAs with short leaders <10nt (Fig 5B). Additional leader nucleotides were strongly inhibitory of leaderless translation, and only 8 contain SD motifs, suggesting some of these short-leadered mRNAs may be poorly initiated.

To estimate the effects of each mRNA feature (ΔG_{unfold}, start codon identity, and leader length) on natural leaderless mRNA translation, we next analyzed ribosome profiling data of the *C. crescentus* mRNAs (27). Here, we utilized translation efficiency measurements which approximate the relative number of ribosome footprints to mRNA fragments from the same cell samples (49). In total, translation efficiency data for 191 leaderless mRNAs and 38 short leadered mRNAs (1-10 leader length) were obtained for cells grown in PYE media (27). We separated leaderless mRNAs into three groups based upon their ΔG_{unfold} values (0-5, 5-10, and >10 kcal/mol) and compared their translation efficiency. The median translation efficiency was reduced as the ΔG_{unfold} increased (Fig 5C) (median= 1.2 for 0-5 kcal/mol, median= 0.89 for 5-10 kcal/mol, median= 0.54 for >10 kcal/mol), similar to the dependence observed in the synthetic translation reporters (Fig 2B). For start codon identity, we noticed that a majority of leaderless mRNAs with near-cognate start codons had translation efficiencies that were not measurable. However, for the 7 GUG mRNAs whose translation efficiency was measured, the median (0.70) was lower than that of the AUG initiated leaderless mRNAs median (0.97) (Fig 5D), in line with the findings of the synthetic reporters (Fig 3). Finally, we compared the translation efficiency of leaderless mRNAs with those with very short leaders (Fig 5E). Since 8 of these mRNAs with short leaders contain SD sequences in the leader, we removed these RNAs from the analysis because we expect them to initiate translation by the canonical mechanism. As leader length

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

increases, we generally observed that the TE tends to decrease (Fig 5E), again in line with the synthetic reporters (Fig 4). Overall these data suggest that the effects of ΔG_{unfold}, start codon identity, and leader length observed in the synthetic translation initiation reporters are also observed across natural *C. crescentus* leaderless mRNAs.

Many RNAs present in the *C. crescentus* transcriptome are not initiated as leaderless mRNAs, so we explored the relative fraction of 5′ AUG trinucleotides in all classes of RNAs (Fig 6A). As noted previously, leaderless mRNAs are highly enriched in AUG codons (Fig 5A). Surprisingly, leadered mRNAs contain a similar fraction of 5′ AUGs as would be predicted from the genome's GC%, which is also observed in small non-coding RNAs (sRNAs), and anti-sense RNAs (asRNAs). Conversely, tRNAs and rRNAs contain zero cases with a 5′ AUG. To explore why these RNAs are not initiated as leaderless mRNAs, we calculated the ΔG_{unfold} of each class of 5′ AUG containing RNA (Fig 6B). If these 5′ AUGs found in non-leaderless RNAs were inaccessible to ribosomes, it would be permissible for this sequence to be present at the 5′ end without causing aberrant initiation. Indeed, for the RNAs with 5′ AUGs, we observe that leaderless mRNAs have a low ΔG_{unfold} (median= 5.0), while leadered mRNAs (median= 9.5), sRNAs (median = 14), and asRNAs (median = 9.0) all contain a significantly higher ΔG_{unfold} values. This suggests that RNAs with inaccessible 5′ AUGs are blocked from leaderless mRNA initiation.

Three component model describes leaderless mRNA start codon selection

In order to understand the mRNA determinants that dictate leaderless mRNA translation, we built a computational model based upon the three features (ΔG_{unfold}, start codon identity, and leader length) which were found to affect leaderless translation initiation. From our synthetic *in vivo* translation initiation reporters, we performed curve fitting to assess the relative effect of

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

each feature on TIE. For each feature (ΔG_{unfold}, start codon identity, and leader length) the highest reporter level measured in each mutant set was normalized to 1 before curve fitting. ΔG_{unfold} data was fit to an exponential equation ($TIE_{\Delta Gunfold} = e^{(-t*0.354)}$) where t is ΔG_{unfold}(kcal.mol), $R^2 = 0.78$), leader length data was fit to a power equation ($TIE_{leader\ length} = 0.45\times(i^{-0.92})$ where i is leader length >0, $R^2 = 0.92$, and $TIE_{leader\ length}$=1 for i=0), and $TIE_{start\ codon}$ was based directly on reporter levels for each near-cognate start codon (Fig 3) and all other codons were given a value of 0 (Table S1). For each mRNA feature, we therefore generated a function that could calculate the relative TIE of any RNA in *C. crescentus* based upon the mRNA sequence. We then built a computational model in which the three features were assumed to be independent from each other to calculate a summed TIE. In this model, we set the maximum TIE to 1, and then subtracted the effects of the sequence feature as measured from the *in vivo* translation reporters in equation 2 ($TIE_{Leaderless\ mRNA(k)}$ = Max TIE (1) - (1-$TIE_{\Delta Gunfold}$) - (1-$TIE_{start\ codon\ identity(j)}$) – (1-$TIE_{leader\ length(i)}$) where k= a given RNA transcript, j=start codon identity, and i=leader length(nt). Using equation 2 we predicted the TIE for each RNA in the *C. crescentus* transcriptome (Fig 7A). For all RNAs, we successively scanned for the closest AUG or near cognate start codon to the 5′ end and used this for the TIE calculation. RNAs known to be initiated as leaderless mRNAs (27,40) yielded higher TIE scores (median = 0.15, $\sigma = 0.35$), while TIE scores for all other RNAs were typically lower (median = -0.95, $\sigma = 0.45$). To estimate the utility of this model at classifying leaderless mRNAs, we used a ROC analysis (Fig 7B). The ROC area under the curve was equal to 0.99, suggesting this simple model can accurately classify those RNAs that are initiated as leaderless mRNAs with high accuracy and precision. The success of this simple $TIE_{leaderless}$ model to classify leaderless mRNAs based on

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the combinations of ΔG_{unfold}, start codon identity, and leader length suggests that these mRNA features combinatorically control translation initiation on leaderless mRNAs.

In addition to the classification of RNAs as leaderless mRNAs we also explored how well the $TIE_{leaderless}$ model predicted translation initiation efficiency. Here, the translation initiation reporters generated were all scored with the $TIE_{leaderless}$ model and compared to their YFP fluorescence. Since $TIE_{leaderless}$ scores below zero are not physically possible, those with negative $TIE_{leaderless}$ values were set to zero to signify they are not predicted to be translated. Overall, the $TIE_{leaderless}$ score correlates strongly to the YFP reporter levels (R^2=0.87) with a slope of 2050 AU. We then compared the $TIE_{leaderless}$ scores to the TE as measured by ribosome profiling of the natural leaderless mRNAs. Since natural leaderless mRNAs encode many genes with diverse codon usages, a poorer correlation was obtained with TE (R^2=0.44, slope=2.4 A.U.) than with the TIE reporters (Fig 7D). Since the TIE reporters all code for YFP with near-identical codon usage, and the natural mRNAs have variable codon usage frequencies, it is possible that translation elongation differences between natural ORFs also impact translation efficiency. Indeed, translation elongation rates have been estimated to be rate limiting *in vivo* in other bacteria (50,51). While it is objectively harder to quantitatively predict translation levels, the $TIE_{leaderless}$ model performs rather well.

Discussion

Here we provide the first systematic analysis of mRNA structure content, start codon identity, and leader length on the initiation of leaderless mRNAs (Fig 7E). Importantly, this study was performed using the bacterium *C. crescentus* which is adapted to efficient leaderless mRNA initiation (27). As has been observed for leadered mRNAs (19,41), mRNA structure content at the leaderless TIR hinders leaderless mRNA translation initiation, suggesting that

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ribosome accessibility is a key feature for leaderless mRNAs. As previously observed in *E. coli*, changes in start codon identity from the preferred "AUG" and presence of leader nucleotides leads to a significant reduction of TIE for *C. crescentus* leaderless mRNAs. Using these quantitative data, we generated a combinatorial $TIE_{leaderless}$ model that predicts the ability of an RNA to initiate as a leaderless mRNA from the individual effects of these features which can be computed for any RNA in the transcriptome. This $TIE_{leaderless}$ model both accurately and sensitively predicts the ability of all RNAs in the *C. crescentus* transcriptome to initiate as leaderless mRNAs. While 5′ AUG is highly enriched in leaderless mRNAs and only rarely observed in non-coding RNAs (Fig 6A), non-coding RNAs containing 5′ AUGs utilize a high ΔG_{unfold} to prevent aberrant translation initiation (Fig 6B). Additionally, very short leaders which were found to inhibit leaderless mRNA initiation, are selected against in leaderless mRNAs and are common in 5′ regions of non-coding RNAs containing non-initiating AUGs. Finally, leaderless mRNAs are much more selective for AUG start codons than are leadered mRNAs, suggesting that the additional stabilization of the translation initiation complex provided by the SD-aSD base pairing helps facilitate initiation on near-cognate start codons.

Leaderless mRNAs have been found to initiate translation in bacterial, archaeal, and both cytoplasmic and mitochondrial eukaryotic ribosomes (17,28,52) suggesting that leaderless initiation is an ancestral initiation mechanism. It is therefore possible that the $TIE_{leaderless}$ model generated here in *C. crescentus* may also perform well across organisms. Indeed, even a few nucleotides preceding the AUG inhibit leaderless mRNA translation initiation in *C. crescentus*, *E. coli*, and mammalian mitochondria (36,38). The strong inhibition of leaderless mRNA translation by TIR secondary structure is likely why leaderless mRNAs in mitochondria have been found to lack 5′ secondary structures (28). *C. crescentus* shares a similar preference for 5′

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AUGs to *E. coli* for leaderless mRNA initiation (32). Interestingly, in the *Mycobacteria*, GUG start codons are much more abundant in leaderless mRNAs and tend to be initiated more similarly to AUG codons in this organism (16). *Mycobacterium* GUG initiated leaderless mRNAs tend to code for short regulatory ORFs (16), as opposed to ORFs encoding functional genes in *C. crescentus* . This suggests that there are likely to be some species-specific differences in leaderless mRNA features arising from the differences in the translation initiation machinery. Indeed, across prokaryotes, 79% of predicted leaderless genes contain AUG as the start codon, whereas GUG, UUG and others are found with an average of 10%, 6% and 3% respectively (13). Surprisingly, leaderless mRNAs across organisms appear to initiate with assembled 70S/80S ribosomes (30,53-55), further suggesting a conserved mechanism of initiation. Therefore, an important goal moving forward will be to determine how broadly across organisms this $TIE_{leaderless}$ model might apply.

Funding

Research reported in this publication was supported by NIGMS of the National Institutes of Health under award numbers R35GM124733 to JMS and Wayne State University start-up funds.

Acknowledgements

We thank members of the Schrader lab for critical feedback.

Conflict of Interests

No conflict of interests exist.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Drummond, D.A. and Wilke, C.O. (2009) The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet*, **10**, 715-724.
2. Kurland, C.G. and Ehrenberg, M. (1987) Growth-optimizing accuracy of gene expression. *Annu Rev Biophys Biophys Chem*, **16**, 291-317.
3. Rodnina, M.V. and Wintermeyer, W. (2001) Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanisms. *Annu Rev Biochem*, **70**, 415-435.
4. Steitz, J.A. and Jakes, K. (1975) How Ribosomes Select Initiator Regions in Messenger-Rna - Base Pair Formation between 3' Terminus of 16s Ribosomal-Rna and Messenger-Rna during Initiation of Protein-Synthesis in Escherichia-Coli. *Proceedings of the National Academy of Sciences of the United States of America*, **72**, 4734-4738.
5. Chen, H., Bjerknes, M., Kumar, R. and Jay, E. (1994) Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. *Nucleic Acids Res*, **22**, 4953-4957.
6. Shine, J. and Dalgarno, L. (1974) The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A*, **71**, 1342-1346.
7. Jacob, W.F., Santer, M. and Dahlberg, A.E. (1987) A single base change in the Shine-Dalgarno region of 16S rRNA of Escherichia coli affects translation of many proteins. *Proc Natl Acad Sci U S A*, **84**, 4757-4761.
8. Hui, A. and de Boer, H.A. (1987) Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in Escherichia coli. *Proc Natl Acad Sci U S A*, **84**, 4762-4766.
9. Calogero, R.A., Pon, C.L., Canonaco, M.A. and Gualerzi, C.O. (1988) Selection of the mRNA translation initiation region by Escherichia coli ribosomes. *Proc Natl Acad Sci U S A*, **85**, 6427-6431.
10. Melancon, P., Leclerc, D., Destroismaisons, N. and Brakiergingras, L. (1990) The Anti-Shine-Dalgarno Region in Escherichia-Coli 16s Ribosomal-Rna Is Not Essential for the Correct Selection of Translational Starts. *Biochemistry*, **29**, 3402-3407.
11. Saito, K., Green, R. and Buskirk, A.R. (2020) Translational initiation in E. coli occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing. *Elife*, **9**.
12. Chang, B., Halgamuge, S. and Tang, S.L. (2006) Analysis of SD sequences in completed microbial genomes: non-SD-led genes are as common as SD-led genes. *Gene*, **373**, 90-99.
13. Srivastava, A., Gogoi, P., Deka, B., Goswami, S. and Kanaujia, S.P. (2016) In silico analysis of 5′-UTRs highlights the prevalence of Shine-Dalgarno and leaderless-dependent mechanisms of translation initiation in bacteria and archaea, respectively. *J Theor Biol*, **402**, 54-61.
14. Beck, H.J. and Moll, I. (2018) Leaderless mRNAs in the Spotlight: Ancient but Not Outdated! *Microbiol Spectr*, **6**.
15. 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.
16. 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.
17. Montoya, J., Ojala, D. and Attardi, G. (1981) Distinctive features of the 5′-terminal sequences of the human mitochondrial mRNAs. *Nature*, **290**, 465-470.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

18. Nakamoto, T. (2006) A unified view of the initiation of protein synthesis. *Biochem Biophys Res Commun*, **341**, 675-678.
19. de Smit, M.H. and van Duin, J. (1990) Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc Natl Acad Sci U S A*, **87**, 7668-7672.
20. de Smit, M.H. and van Duin, J. (1994) Control of translation by mRNA secondary structure in Escherichia coli. A quantitative analysis of literature data. *J Mol Biol*, **244**, 144-150.
21. Skripkin, E.A., Adhin, M.R., de Smit, M.H. and van Duin, J. (1990) Secondary structure of the central region of bacteriophage MS2 RNA. Conservation and biological significance. *J Mol Biol*, **211**, 447-463.
22. 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.
23. Salis, H.M. (2011) The ribosome binding site calculator. *Methods Enzymol*, **498**, 19-42.
24. Romero, D.A., Hasan, A.H., Lin, Y.F., Kime, L., Ruiz-Larrabeiti, O., Urem, M., Bucca, G., Mamanova, L., Laing, E.E., van Wezel, G.P. *et al.* (2014) A comparison of key aspects of gene regulation in Streptomyces coelicolor and Escherichia coli using nucleotide-resolution transcription maps produced in parallel by global and differential RNA sequencing. *Mol Microbiol*.
25. Babski, J., Haas, K.A., Nather-Schindler, D., Pfeiffer, F., Forstner, K.U., Hammelmann, M., Hilker, R., Becker, A., Sharma, C.M., Marchfelder, A. *et al.* (2016) Genome-wide identification of transcriptional start sites in the haloarchaeon Haloferax volcanii based on differential RNA-Seq (dRNA-Seq). *BMC Genomics*, **17**, 629.
26. Pfeifer-Sancar, K., Mentz, A., Ruckert, C. and Kalinowski, J. (2013) Comprehensive analysis of the Corynebacterium glutamicum transcriptome using an improved RNAseq technique. *BMC Genomics*, **14**, 888.
27. Schrader, J.M., Zhou, B., Li, G.W., Lasker, K., Childers, W.S., Williams, B., Long, T., Crosson, S., McAdams, H.H., Weissman, J.S. *et al.* (2014) The coding and noncoding architecture of the Caulobacter crescentus genome. *PLoS Genet*, **10**, e1004463.
28. Jones, C.N., Wilkinson, K.A., Hung, K.T., Weeks, K.M. and Spremulli, L.L. (2008) Lack of secondary structure characterizes the 5′ ends of mammalian mitochondrial mRNAs. *RNA*, **14**, 862-871.
29. Tedin, K., Moll, I., Grill, S., Resch, A., Graschopf, A., Gualerzi, C.O. and Blasi, U. (1999) Translation initiation factor 3 antagonizes authentic start codon selection on leaderless mRNAs. *Mol Microbiol*, **31**, 67-77.
30. O'Donnell, S.A. and Janssen, G.R. (2002) Leaderless mRNAs bind 70S ribosomes more strongly than 30S ribosomal subunits in Escherichia coli. *Journal of Bacteriology*, **184**, 6730-6733.
31. Van Etten, W.J. and Janssen, G.R. (1998) An AUG initiation codon, not codon-anticodon complementarity, is required for the translation of unleadered mRNA in Escherichia coli. *Molecular Microbiology*, **27**, 987-1001.
32. O'Donnell, S.M. and Janssen, G.R. (2001) The initiation codon affects ribosome binding and translational efficiency in Escherichia coli of cI mRNA with or without the 5′ untranslated leader. *J Bacteriol*, **183**, 1277-1283.
33. Hering, O., Brenneis, M., Beer, J., Suess, B. and Soppa, J. (2009) A novel mechanism for translation initiation operates in haloarchaea. *Mol Microbiol*, **71**, 1451-1463.
34. Chen, W.C., Yang, G.P., He, Y., Zhang, S.M., Chen, H.Y., Shen, P., Chen, X.D. and Huang, Y.P. (2015) Nucleotides Flanking the Start Codon in hsp70 mRNAs with Very Short 5 '-UTRs Greatly Affect Gene Expression in Haloarchaea. *Plos One*, **10**.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

35. Brock, J.E., Pourshahian, S., Giliberti, J., Limbach, P.A. and Janssen, G.R. (2008) Ribosomes bind leaderless mRNA in Escherichia coli through recognition of their 5 '-terminal AUG. *Rna-a Publication of the Rna Society*, **14**, 2159-2169.
36. Krishnan, K.M., Van Etten, W.J., 3rd and Janssen, G.R. (2010) Proximity of the start codon to a leaderless mRNA's 5 terminus is a strong positive determinant of ribosome binding and expression in Escherichia coli. *J Bacteriol*, **192**, 6482-6485.
37. Jones, R.L., 3rd, Jaskula, J.C. and Janssen, G.R. (1992) In vivo translational start site selection on leaderless mRNA transcribed from the Streptomyces fradiae aph gene. *J Bacteriol*, **174**, 4753-4760.
38. Christian, B.E. and Spremulli, L.L. (2010) Preferential Selection of the 5 '-Terminal Start Codon on Leaderless mRNAs by Mammalian Mitochondrial Ribosomes. *Journal of Biological Chemistry*, **285**, 28379-28386.
39. Marks, M.E., Castro-Rojas, C.M., Teiling, C., Du, L., Kapatral, V., Walunas, T.L. and Crosson, S. (2010) The genetic basis of laboratory adaptation in Caulobacter crescentus. *J Bacteriol*, **192**, 3678-3688.
40. Zhou, B., Schrader, J.M., Kalogeraki, V.S., Abeliuk, E., Dinh, C.B., Pham, J.Q., Cui, Z.Z., Dill, D.L., McAdams, H.H. and Shapiro, L. (2015) The global regulatory architecture of transcription during the Caulobacter cell cycle. *PLoS Genet*, **11**, e1004831.
41. Mustoe, A.M., Corley, M., Laederach, A. and Weeks, K.M. (2018) Messenger RNA Structure Regulates Translation Initiation: A Mechanism Exploited from Bacteria to Humans. *Biochemistry*, **57**, 3537-3539.
42. Gruber, A.R., Lorenz, R., Bernhart, S.H., Neubock, R. and Hofacker, I.L. (2008) The Vienna RNA websuite. *Nucleic Acids Res*, **36**, W70-74.
43. Reuter, J.S. and Mathews, D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.
44. Thanbichler, M., Iniesta, A.A. and Shapiro, L. (2007) A comprehensive set of plasmids for vanillate- and xylose-inducible gene expression in Caulobacter crescentus. *Nucleic Acids Res*, **35**, e137.
45. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B. *et al.* (2012) Fiji: an open-source platform for biological-image analysis. *Nat Methods*, **9**, 676-682.
46. Ducret, A., Quardokus, E.M. and Brun, Y.V. (2016) MicrobeJ, a tool for high throughput bacterial cell detection and quantitative analysis. *Nat Microbiol*, **1**, 16077.
47. 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. *Journal of Machine Learning Research*, **12**, 2825-2830.
48. Tobias, J.W., Shrader, T.E., Rocap, G. and Varshavsky, A. (1991) The N-end rule in bacteria. *Science*, **254**, 1374-1377.
49. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S. and Weissman, J.S. (2009) Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*, **324**, 218-223.
50. Racle, J., Picard, F., Girbal, L., Cocaign-Bousquet, M. and Hatzimanikatis, V. (2013) A genome-scale integration and analysis of Lactococcus lactis translation data. *PLoS Comput Biol*, **9**, e1003240.
51. Vieira, J.P., Racle, J. and Hatzimanikatis, V. (2016) Analysis of Translation Elongation Dynamics in the Context of an Escherichia coli Cell. *Biophys J*, **110**, 2120-2131.
52. Baltz, R.H., Hegeman, G. and Skatrud, P.L. (1993) *Industrial microorganisms: basic and applied molecular genetics*. American Society for Microbiology.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

53. Moll, I., Hirokawa, G., Kiel, M.C., Kaji, A. and Blasi, U. (2004) Translation initiation with 70S ribosomes: an alternative pathway for leaderless mRNAs. *Nucleic Acids Res*, **32**, 3354-3363.
54. Udagawa, T., Shimizu, Y. and Ueda, T. (2004) Evidence for the translation initiation of leaderless mRNAs by the intact 70 S ribosome without its dissociation into subunits in eubacteria. *J Biol Chem*, **279**, 8539-8546.
55. Andreev, D.E., Terenin, I.M., Dunaevsky, Y.E., Dmitriev, S.E. and Shatsky, I.N. (2006) A leaderless mRNA can bind to mammalian 80S ribosomes and direct polypeptide synthesis in the absence of translation initiation factors. *Mol Cell Biol*, **26**, 3164-3169.

Figure Legends

Figure Legends

Figure 1. Leaderless mRNA translation initiation regions are more accessible than leadered mRNAs. A.) Predicted unfolding energy of mRNAs. The predicted mRNA minimum free energy (ΔG_{mRNA}) is represented on the left. The orange translation initiation region indicates a ribosome footprint surrounding the start codon (pink). The image on the right represents the mRNA upon initiation (ΔG_{init}) where the orange initiation region is unfolded. The ΔG_{unfold} represents the amount of energy required by the ribosome to unfold the translation initiation region of the mRNA. B.) Violin plots of ΔG_{unfold} (right) calculated for all the mRNAs of each class (left) in the *Caulobacter crescentus* genome based on the transcript architecture(27,40). P-values were calculated based on t-test (two tailed, unequal variance).

Figure 2. ΔG_{unfold} strongly influences leaderless mRNA translation. A.) Synonymous codon mutations in synthetic leaderless constructs were generated to alter the ΔG_{unfold} of the translation initiation region. The bases in the start codon are colored pink, red bases highlight where mutations were introduced to disrupt base pairing. B.) *In vivo* translation reporter levels the various leaderless RNA mutants. Each hairpin and its synonymous codon mutant set are shown with the same color (Raw data can be found in Table S1). The natural log of the average YFP

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

intensity per cell is shown and error bars represent the standard deviation of three biological replicates. The dotted blue line represents a linear curve fit with an R^2 value of 0.84 and a slope of -0.3.

Figure 3. Leaderless mRNAs have a strong preference for AUG start codons. Leaderless mRNA *in vivo* translation reporters were generated with the start codons listed on the X-axis and their average YFP intensity per cell were measured. On the right, is a zoomed in view of all non-AUG codons tested. Error bars represent the standard deviation from three biological replicates.

Figure 4. Leaderless mRNAs are inhibited by additional upstream nucleotides. Leaderless mRNA *in vivo* translation reporters were generated with variable number of leading nucleotides on the X-axis and their average YFP intensity per cell were measured (Raw data can be found in Table S1). Error bars represent three biological replicates.

Figure 5. ΔG_{unfold}, start codon identity, and leader length correlate with translation efficiency (TE) across native leaderless mRNAs.

A.) Bar graph showing the fraction of leaderless mRNAs starting with AUG, GUG, UUG and CUG start codons. Also shown are the random chances of trinucleotides being AUG, GUG, UUG and CUG calculated based on GC content (67%) of *C. crescentus* genome. P-values were calculated based on a two-tailed Z-test. B.) Bar graph showing the fraction of leaderless mRNAs and mRNAs with 5′ untranslated region (UTR) of length 1 to 10. mRNAs containing Shine-Dalagarno sites were excluded from this analysis. P-values were calculated based on a two-tailed Z-test of each leader length compared to leader length 0. C.) Violin plot of translation efficiency (TE) as measured by ribosome profiling(49) of natural leaderless mRNAs binned in three groups depending on ΔG_{unfold} values (0-5, 5-10, and >10 kcal/mol). P-values based on t-test (two tailed, unequal variances). D.) Violin plot of TE as measured by ribosome profiling(49) of natural

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

leaderless mRNAs starting with AUG and GUG. P-values were calculated based on a t-test (2-tailed, unequal variance). E.) Violin plot showing the TE as measured by ribosome profiling(49) on the Y-axis of leaderless mRNAs (green) and with leaders of varying length (1-10) shown in grey. P-values were calculated based on t-test (2-tailed, unequal variance).

Figure 6. Non-coding RNAs with 5′ AUGs are rare and have higher ΔG_{unfold}.

A.) Bar graph showing the fraction of natural leaderless mRNAs starting with trinucleotide AUG and other types of RNAs starting with trinucleotide AUG, but not initiated at that AUG (leadered mRNAs, sRNAs, rRNAs, tRNAs and asRNAs). Also shown is the random chance of trinucleotide being AUG out of 10000 nucleotides; calculated based on GC content of *C. crescentus* genome. P-values were calculated using a two-tailed Z-test with each RNA class compared to the random probability of 5′ AUG. B.) Violin plot showing ΔG_{unfold} of natural leaderless mRNAs starting with AUG (green) and other types of RNAs starting with AUG, but not initiated at that AUG (leadered mRNAs, RNAs and asRNAs) (shown in grey). P-values were calculated based on a T-test (2-tailed, unequal variance).

Figure 7. A combinatorial model accurately predicts translation of leaderless mRNAs.

A.) Line graph showing the predicted $TIE_{leaderless}$ scores on the X-axis and the number of RNAs on the Y axis. The solid blue line represents natural leaderless mRNAs. The black dotted line represents the RNAs that are not leaderless RNAs. The grey line represents all RNAs. RNAs with short leaders are shown in Fig S1. B.) ROC curve (shown in solid blue, with “random” shown as a dotted line) with true positive rate on Y-axis and false positive rate on X-axis. The area under curve (AUC) was calculated to be 0.99. C.) TIE reporter levels compared to $TIE_{leaderless}$ scores. For the leaderless TIE reporters tested (Table S1) the YFP reporter level (Y-axis) is plotted compared to the $TIE_{leaderless}$ (X-axis). The trendline is the result of a least-squares

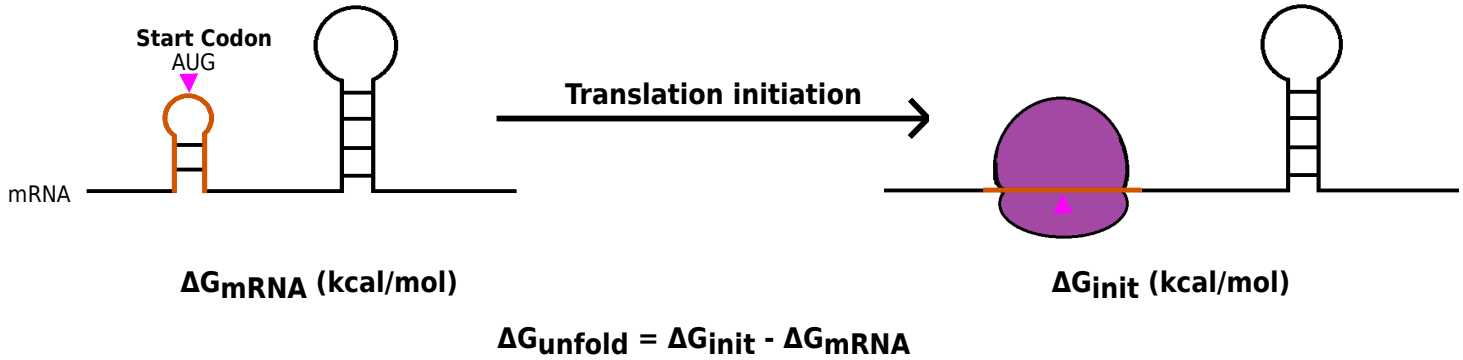
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

fit yielding a slope of 2050 with R^2=0.87. Error bars represent the standard deviation of at least three biological replicates. D.) Translation efficiency (TE) of leaderless mRNAs (Y-axis) is plotted compared to $TIE_{leaderless}$ (X-axis). The trendline is the result of a least-squares fit yielding a slope of 2.4 and R^2=0.44. E.) Model design showing ribosome binding to the AUG trinucleotide (pink triangle) at the 5′ end when it is highly accessible as shown in the left. The ribosome binding is prevented when the region becomes more structured and the accessibility decreases.

Fig 1

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A.



$$\Delta G_{unfold} = \Delta G_{init} - \Delta G_{mRNA}$$

B.

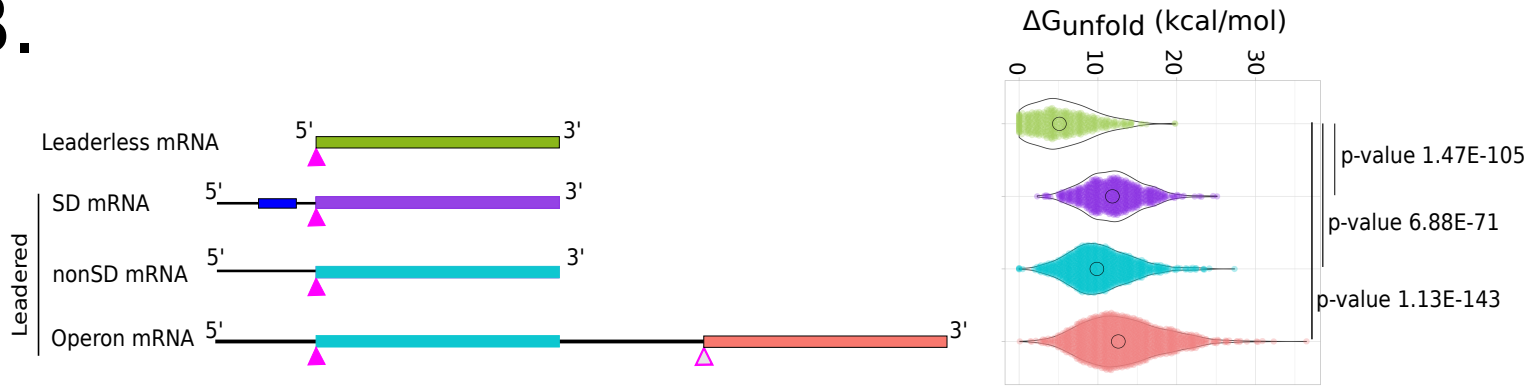
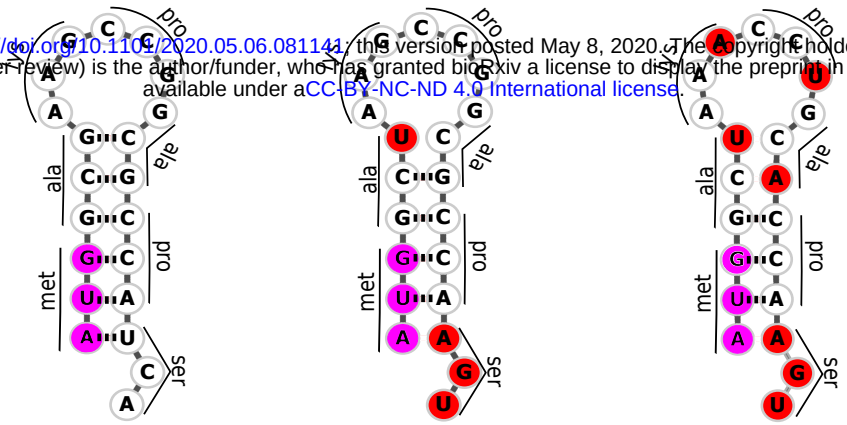


Fig 2

A.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



ΔG_{unfold} (kcal/mol) =	8.71	7.93	4.94

B.

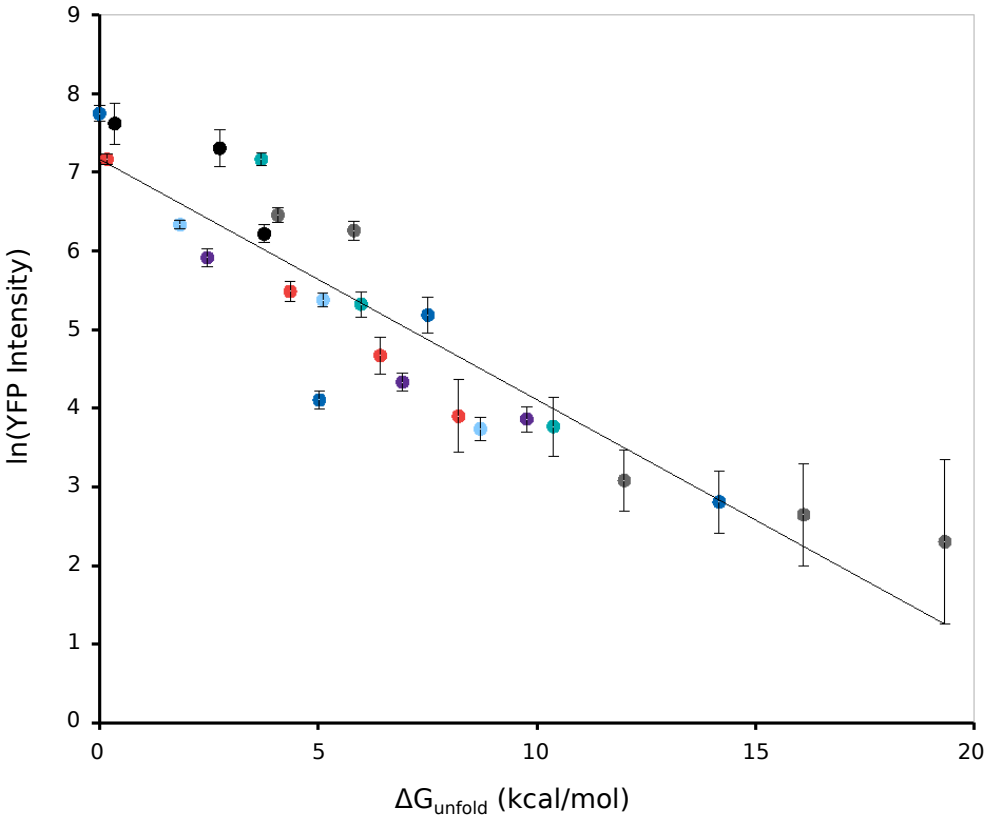


Fig 3

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

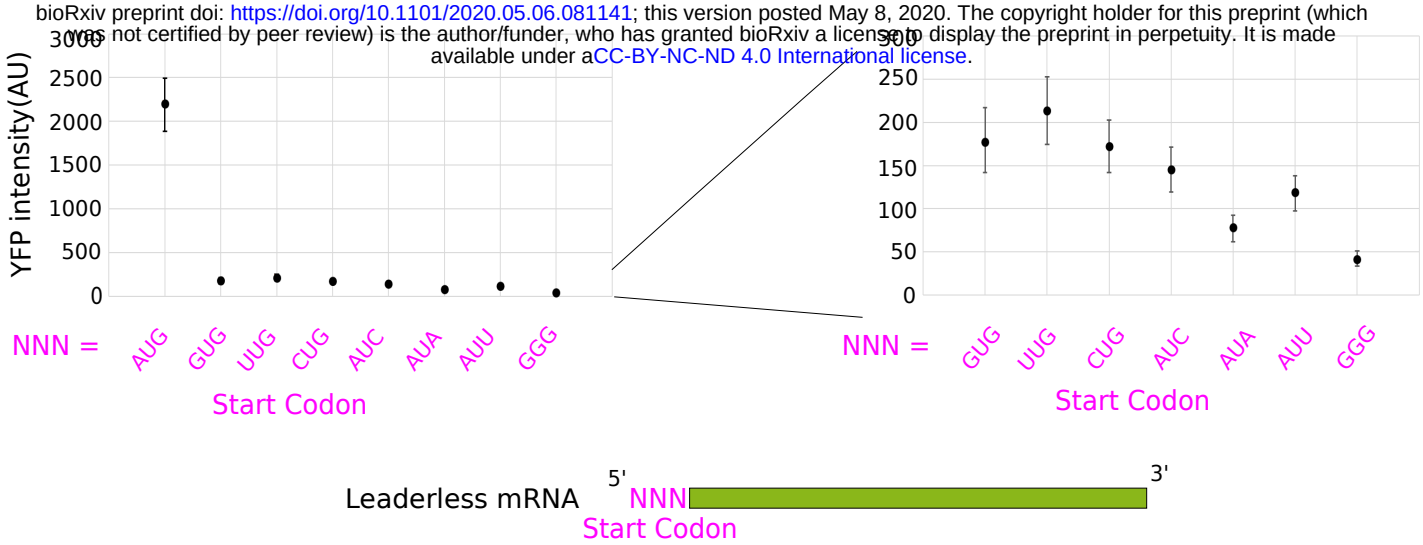


Fig 4

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

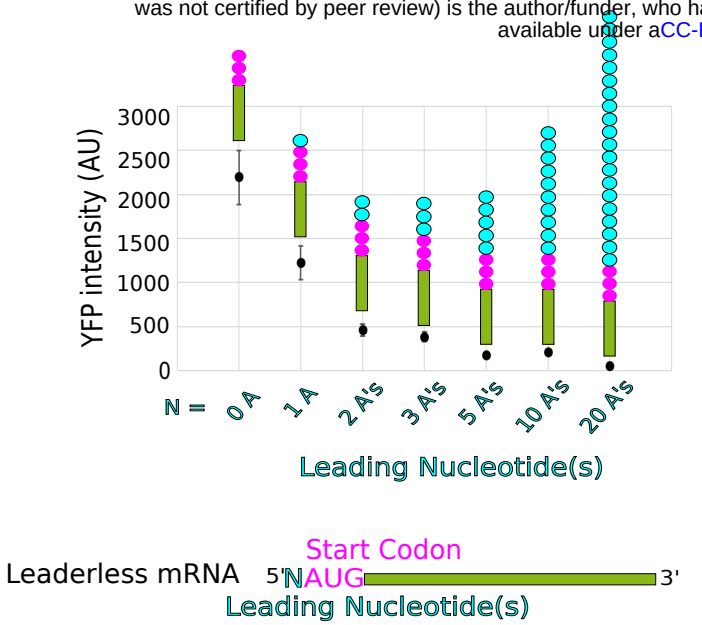
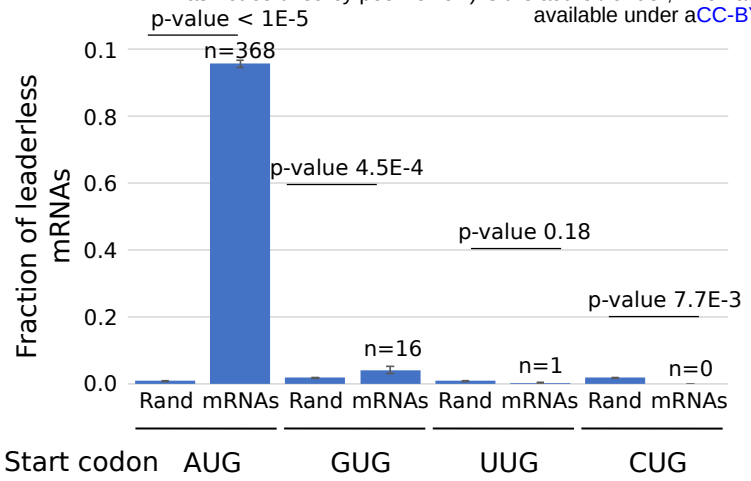


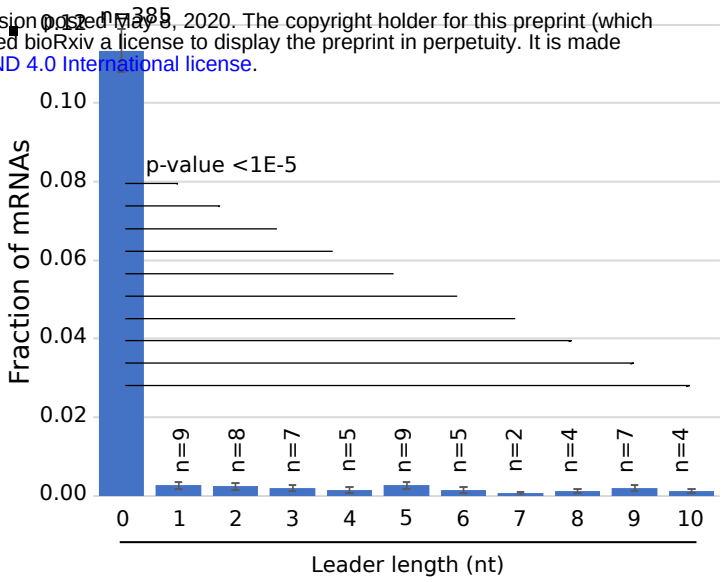
Fig 5

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

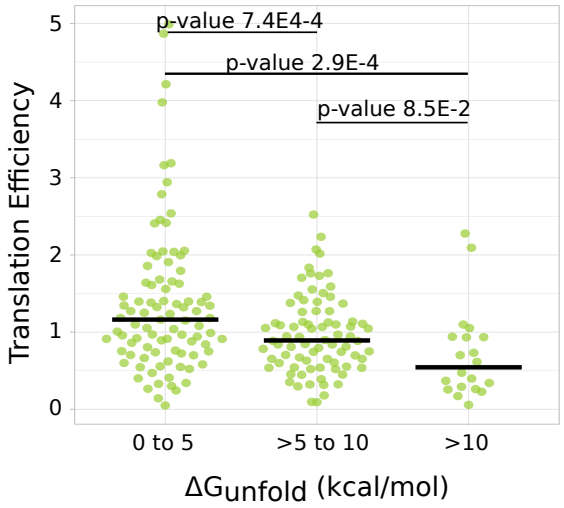
A.



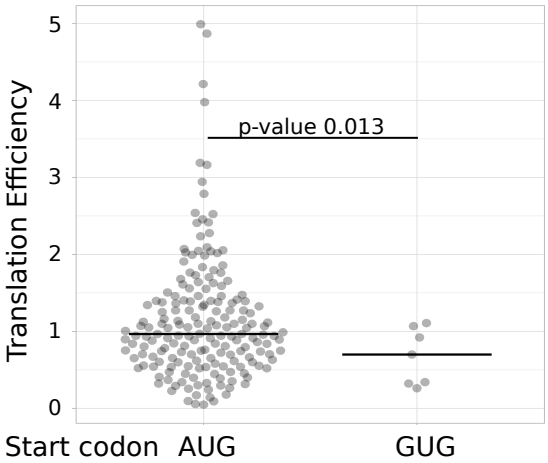
B.



C.



D.



E.

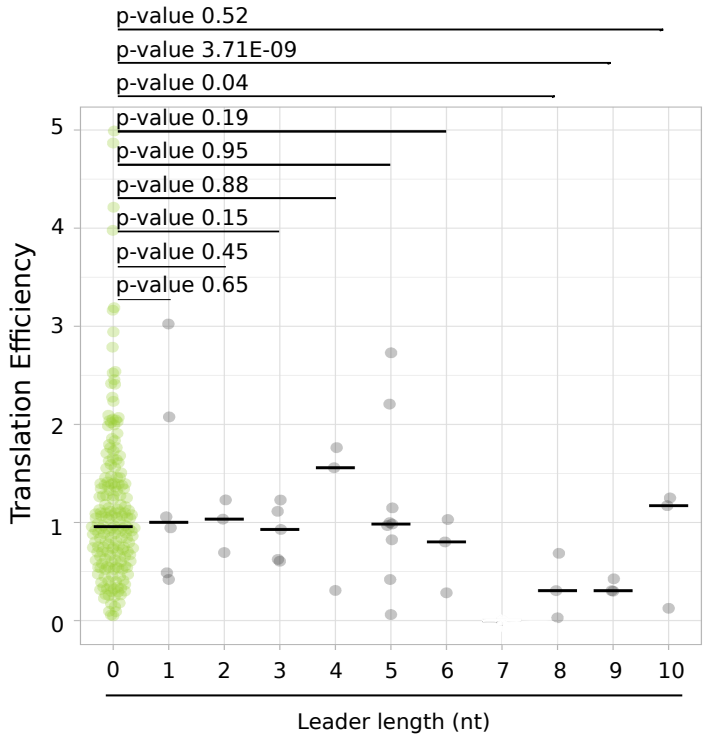


Fig 6

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

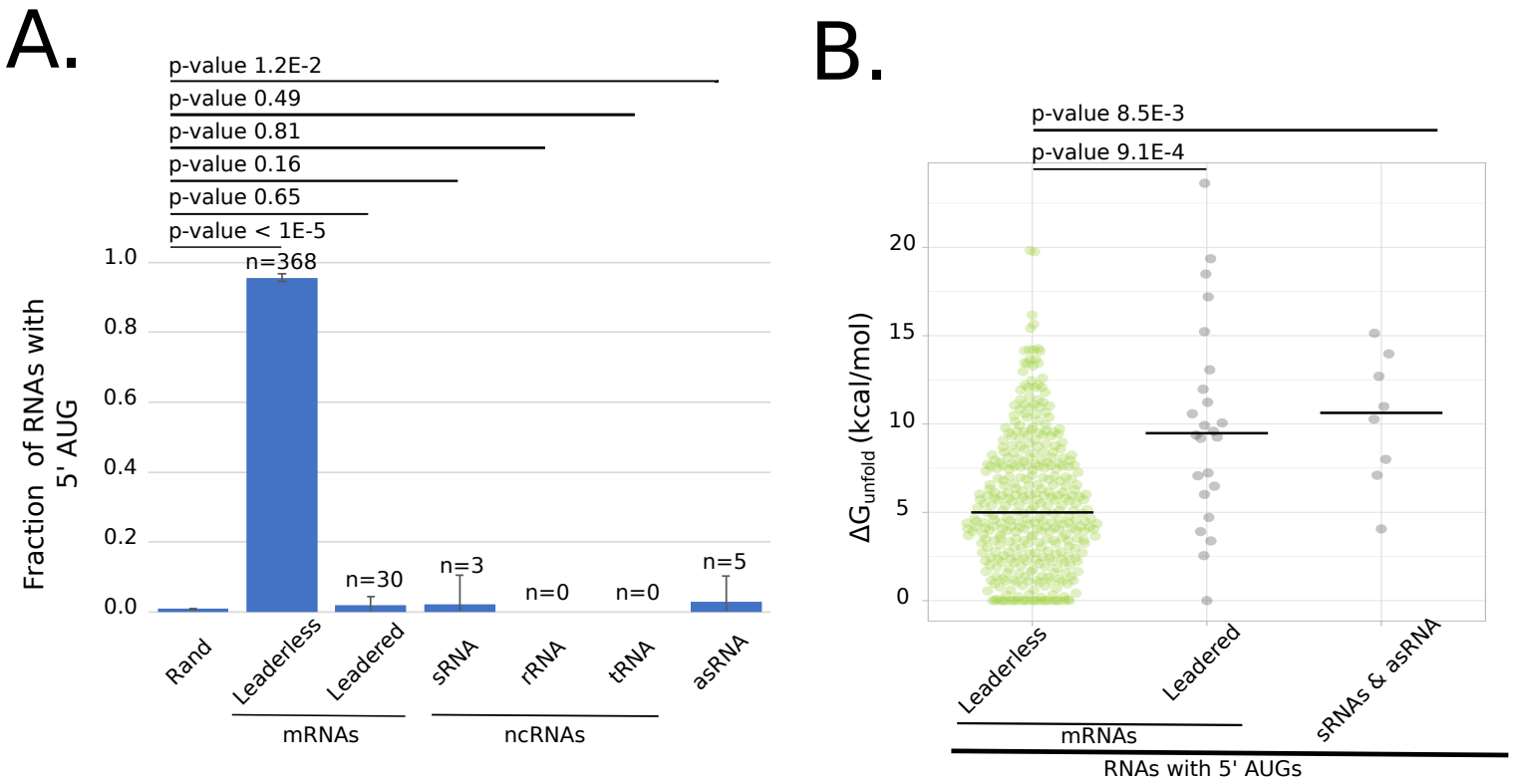
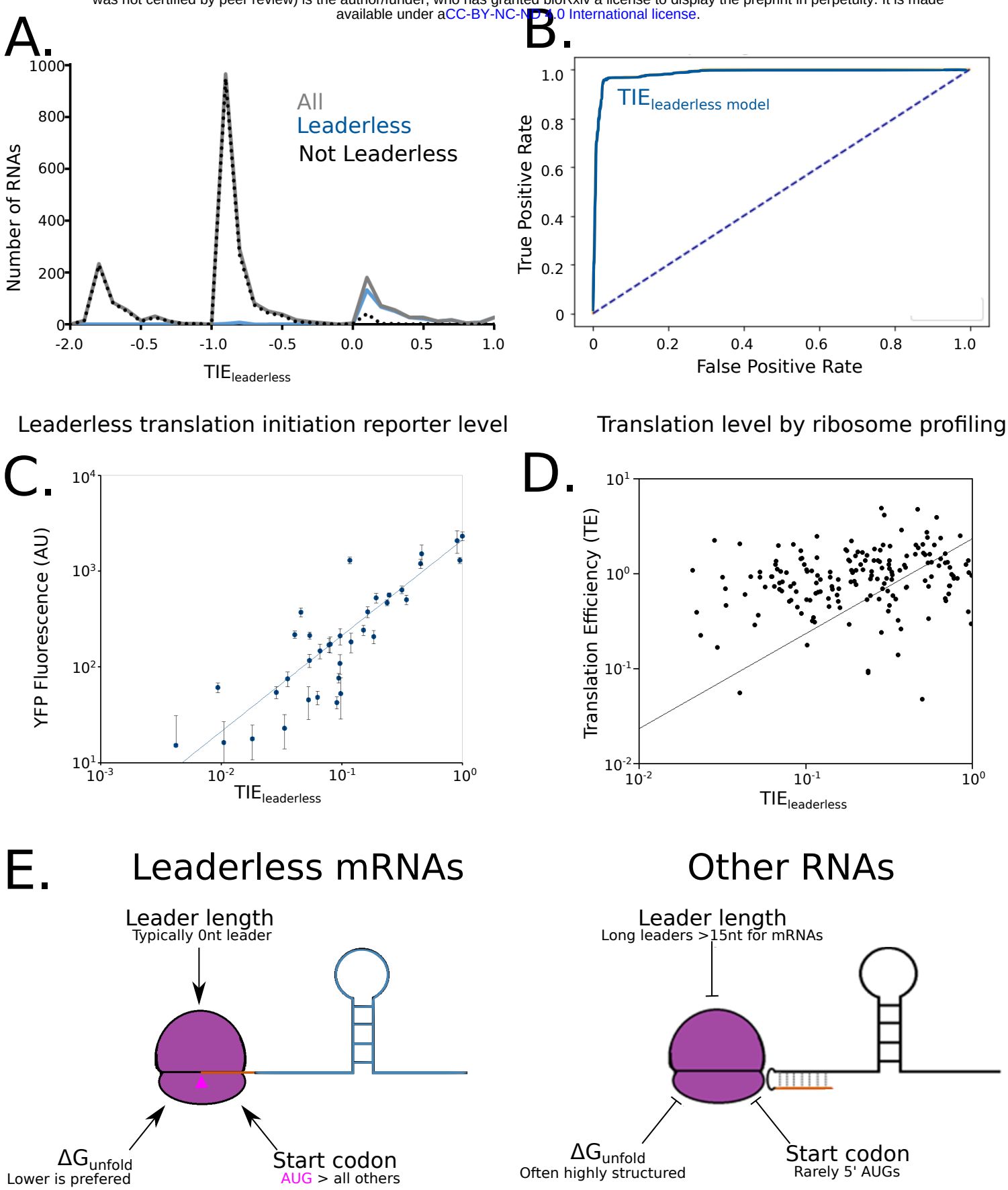


Fig 7

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.06.081141; this version posted May 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Leaderless mRNA classification