Modeling mRNA Populations

R. Urquidi Camacho[a], N. Pollesch[b,1], M.A. Gilchrist[a,c,d,1,*]

[a] *Genome Science and Technology Program, University of Tennessee, Knoxville, TN 37996-XXX*
[b] *Department of Mathematics, University of Tennessee, Knoxville, TN 37996-1320*
[c] *Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996-1610*
[d] *National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville, TN 37996-3410*

Abstract

This paper presents a model to describe the dynamics of protein translation. A system of ordinary differential equations is derived to describe the number of ribosomes bound to a strand of mRNA at a given time. The number of ribosomes bound to an mRNA at a given time is referred to its ribosome load. The mRNA is classified based on its ribosome load and whether or not it's decapped for future degradation. Distribution of ribosome counts is assumed to be related to the translation initiation rate, translation completion rate, degredation marking rate, and length of the mRNA. The length of the mRNA's coding region plays the role of controlling the number of ribosome counts which, in turn, determines the number of ODEs in the system. A goal of this work is to see how the equilibrium distribution between classes as changes with coding region length. A closed form solution to the density in the i^{th} ribosomal class in a system with $i_{\max}$ states is presented for the equilibrium distribution of the decapped classes in terms of the capped classes. The equilibrium solutions in the capped classes are shown to be related to the full determinant of the tri-diagonal matrix used to describe the system, as well as all the determinants of the minors associated to it. In general, there is no closed form for the determinant of a tri-diagonal matrix, only a recurrence relation that can be used to find determinants. However, in this model a closed form exists for the full determinant as it changes with changing values of $i_{\max}$ and its formula is presented . This closed form for the determinant provides a method to efficiently find equilibrium solutions for the entire system. Additionally, a continuous approximation using PDE is derived and also used to find equilibrium solutions to the system. Both of these methods for determining equilibrium solutions are utilized in an effort to find the set of parameters that maximizes the likelihood of a given data set. A process for mapping the equilibrium model results to data is also presented and used to begin preliminary estimation of model parameters

*Corresponding author
Email address: mikeg@utk.edu (M.A. Gilchrist)

and to verify model function.

alternate abstract: Modeling Ribosomal Loading of mRNA

A model is presented to describe the dynamics of protein translation related to the ribosomal load of an mRNA. The number of ribosomes bound at a given time is referred to as ribosome load, and using this value a population of mRNA are classified. A system of ordinary differential equations (ODEs) is derived and solved for the equilibrium distribution of a population of mRNA. Distribution of ribosome counts is assumed to be related to the translation initiation rate, translation completion rate, degradation marking rate, and length of the mRNA. Methods are developed to find analytical equilibrium solutions to the system of ODEs and a system of partial differential equations (PDEs) are derived to find numerical approximations to the ODE system at equilibrium as well. Both the PDE continuous approximation and the analytical solutions to the ODE system agree offering two different methods for finding solutions at equilibrium within optimization routines. Additionally, a tool is developed and presented that is used to compare the model results to empirical microarray data measures of ribosome load.

Keywords: bioinformatics, mRNA population, protein translation, ribosome loading, ribosome count, polysome, mathematical model

---

Paper Outline

1. Motivation - **(Mike)**
   (a) Why is this process important?
   (b) What will this model enable researchers to do?
   (c) Other modeling efforts?
2. Derivation and Assumptions
   (a) Physical processes captured (Ideally, have a quick discussion of process and inline definitions of variables used to represent process, followed by a total recap in a table) - **(Nate)**
      i. System described as population model: Dichotomy of decapped and capped mRNA. State variables based on an mRNA's ribosome load.
      ii. Process of mRNA production
      iii. Process of Marking mRNA for degradation supposed
      iv. Three processes of : Initiation, translation, and completion
   (b) Definition/Discussion of system boundaries - (**MIKE**)
      i. Physical boundaries as a cell and relation to parameters
      ii. Discussion of perceived upper and lower limits to state variables and parameters
      iii. Temporal boundaries and relation to steady state
   (c) Assumptions: Such as initial assumptions of specific functional forms, i.e. marking rate constant among classes - **(Nate)**
   (d) Justify consideration of system as two subsystems, decapped and capped. - **(Nate)**
3. Model Formulation: Total model presented and then analysis of capped and decapped systems - **(Nate)**
   (a) ODE/Discrete system
      i. Present system of ODEs (Total, capped, and decapped)
      ii. Matrix Representation of ODE model (Total, capped, and decapped)
      iii. Steady state formulations
   (b) PDE/Continuous system

i. Explain motivation for deriving PDE

ii. Explain framing as 'non-linear birth and death process'

iii. Explain derivation using Taylor expansion

iv. Present PDE for capped class

v. Present non-dimensionalized system

vi. Present 2nd order ODE to be solved for non-dimensionalized PDE at Equilibrium

vii. Motivate and present equation for decapped class at equilibrium

viii. (Make decision to present results for steady state values for PDE here or in a separate section to follow)

4. Results - **(Nate)**

   (a) Present solution strategies/methods

      i. ODE/Discrete system: Matrix inversion technique

      ii. PDE/Continuous system: Numerical solver of 2nd order ODE that arises at equilibrium

      iii. Discussion of alternative solution approaches

   (b) Present actual solutions for a couple sets of parameters: Highlight agreement of ODE and PDE system

   (c) Present solutions for discrete system under further simplifications for translation and initiation

5. Opportunities for Future Research - **(Nate and Mike)**

   (a) Application of model to real data. Can highlight sources of data.

   (b) Alternate functional forms and relaxed assumptions

   (c) Further establish connection (in simplified system) to potential probability distributions

   (d) How to move forward with analytical solutions, specifically connection to solving 2nd order partial difference equation arising from tri-diagonal form of matrix, note here that boundary conditions exist that may be utilized which are not normally present.

1. Introduction

This section addresses such topics as why modeling this process important, what this model will enable researchers to do, and what other modeling efforts exist that seek to achieve the same goals.

1.1. mRNA and Translation

Intro Outline 3.1.1. Gene regulation, translation and mRNA stability 3.1.1.1. Short introduction to Gene expression, transcription, translation, and the regulation of mRNA populations both dependent and independent of translation 3.1.2. Ever increasing methods of measuring mRNA decay and Translation provide ample grounds for testing and knitting together hypothesis underlying the mechanism of translation. 3.1.2.1. Ribo-seq, microarrays, polysome profiling, proteomics and live imaging. 3.1.2.1.1. But most of these approaches are not measurements of single transcripts, but ensemble measurements of populations 3.1.3. Mathematical modeling as a tool to interpret and generate hypotheses to better understand translation 3.1.3.1. TASEP 3.1.3.2. Riboflow 3.1.3.3. Other bulk "cell-wide" approaches shah 2013 3.1.4. Our model acts as an intermediate between cell wide approaches and single transcript models such as TASEP and Riboflow. Our coarse-grained model of translation focuses on the behavior of transcript populations. This includes effects originating from transcription and mRNA decay as well as translation initiation and elongation/termination. By modeling translation at the population level, we can also use the model in the future to better understand the information held in ribo-seq and proteomics experiments.

1. Gene expression short overview
    (a) Gene expression is often stated as the central dogma in which genetic information encoded in the DNA is transcribed into mRNA which is subsequently translated into protein.
    (b) Often, a greater amount of attention is focused on explaining gene expression at the transcriptional level and prevailing changes of mRNA transcript levels.
    (c) However, multiple studies across all kingdoms of life have shown that transcript expression level is only moderately predictive of the final protein expression.
    (d) Gene expression at the post transcriptional level is controlled by mRNA transcript stability and degradation, translation and protein maturation/degradation.
    (e) The model presented in this paper encompasses gene expression regulation occuring at the translational and the mature mRNA population level.

2. Biology controlling mRNA stability and translation

   (a) Mature mRNAs in the cytosol are called the free mRNA pool, and are in one of three states.

   (b) They are actively being translated by ribosomes and will continue to initiate new rounds of translation until the transcript is degraded.

   (c) Transcripts are degraded directly from the free mRNA pool.

   (d) Transcripts are protected from degradation by RNA binding protein chaperones or are found in processesing bodies awaiting translation initiation or degradation.

   (e) Degradation of mature mRNAs is controlled by numerous processes depending on whether they are bound to ribosome, in processing bodies or in the free mRNA pool.

   (f) Free mRNAs can be decapped or deadenylated followed by exonuclease digestion.

   (g) Ribosomes can destine transcripts to degradation under multiple conditions.

   (h) The first ribosome to bind to a freshly exported transcript performs the "pioneer round of translation", which is charged with assesing the mRNA's quality.

   (i) There are 3 processes which occur in the pioneer round of translation, all of which detect different mRNA defects.

   (j) No Go Decay (NGD) detects a stalled ribosome, either due to mRNA structural features, slowly translating sequence or inteference of translation elongation.

   (k) No stop decay (NSD) detects a missing stop codon and nonsense mediated decay (NMD) detects potential mis splicing or nonsense mutations.

   (l) All three decay mechanisms, NMD, NSD and NGD lead to the eventual degradation of their bound transcripts.

   (m) While NSD and NMD are a restricted to the pioneering round of translation, NGD can also uccur during the following rounds of translation.

   (n) As transcripts are cleared by the pioneering round of translations more ribosomes can attach to the transcript, once more than one ribosome is on a transcript this ribosome mRNA complex is called a polysome.

   (o) Transcripts associated to ribosomes are generally assumed to be protected from degradation and only degraded once ribosomes are off the transcript, however both NGD, (sRNA silencing) and a process called cotranslational decay can degrade actively translated transcripts.

(p) Cotranslational decay involved the decapping of actively translating mRNA transcripts and subsequent 5' to 3' mRNA degradation which follows a 3 nucleotide periodic pattern in step with the Ribosome.

3. Current Models/Research and how our model fits in the current field

   Stuff for intro

4. mRNA degradation mechanisms (Cao and Parker 2001, Cao and Parker 2003, Wu 2013, Wu 2016, Zupanic 2016, reviewed in Ashworth 2019) and translation (Reuveni 2011, Nanikashvili 2019, Raveh 2016, Shaw 2003, Shah 2013) have been modeled separately in the past, but only rarely together (Reuveni 2011, Valleriani 2011).

5. in our model we are going to be focusing on the process of mRNA 5' decapping.

1. The basic representation of the central dogma dictates that epression of protein coding genes starts from genes encoded in DNA that are transcribed to mRNA and subsequently translated to Protein.

2. A more careful representation considers that the final protein production is dependent on both the maintenance of an actively translating mRNA population, the association of ribosomes on the population and finally the degradation of the protein itself.

3. The maintenance of mRNA populations relies on the balance of mRNA transcription rates, the translation status of transcripts and numerous mRNA decay pathways.

4. mRNA degradation relies on removing protective and translation enhancing componenets of the mRNA. These include the 5' mG cap and the 3' polyadenosine tail.

5. Additionally mRNA degradation can be promoted through endonucleolytic cleavage by RISC (and siRNAs).

6. mRNA degradation can occur in both a ribosomal associated or a ribosome free manner.

7. Ribosomal association of transcripts can lead to both protection of viable transcripts as well as quality control degradation of faulty transcripts.

8. When a viable transcript is bound by the ribosomal and translational machinery, the 5' cap is bound by translational initation factors and the 3' tail is bound by poly A binding proteins. This protects transcripts from exonucleic attack and degradation.

9. Endonucleic degradation is still possible, but reduced due to a reduced accessibility of the siRNA binding sequence on the transcript through competition with ribosomes.

10. However multiple mechanisms of mRNA decay are carried out in association with the ribosome. Nonsense mediated decay, no go decay and no stop decay all rely on ribosomes detecting faults in the transcript and susbsequently interacting with degradation machinery to remove the faulty transcript.

11. With some mechanisms of mRNA decay, decay can occur co-translationally. This is mainly seen in 5' decapping. When a translating transcript is decapped the 5' to 3' exonucleic degradation machinery trails the most upstream ribosome. As the ribosome translates the mRNA is degraded.

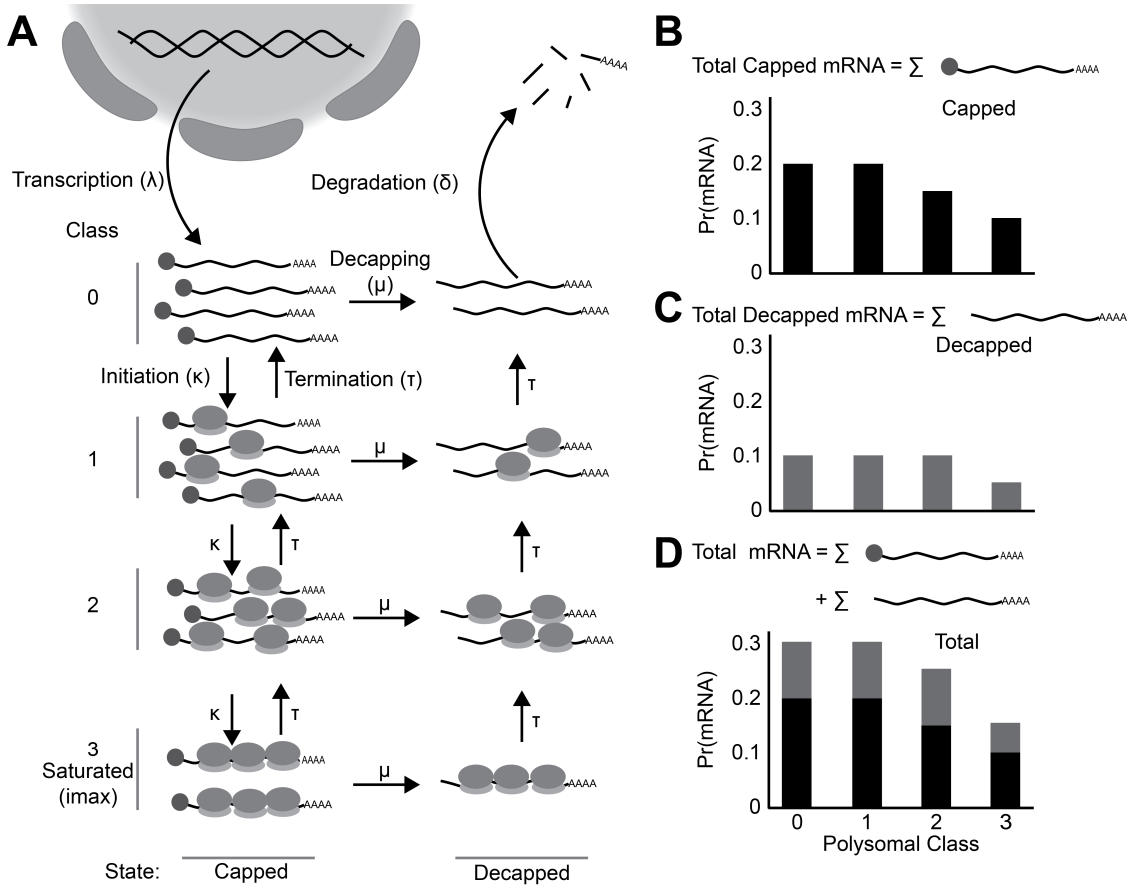
12.

many models seek to understand separate aspects of mRNA biology. Some focus on the mechanistic aspects of decay, separate from the interaction with the transaltional machinery. Others model translation directly, but not decay (TASEP and RFM). Others focus on a bulk measure of all processes, but with no particular allusion to specific types of degradation. here we present a model focused on integrating from mRNA procution to final degradation, the maintenance of mRNA populations with regards to translation and co transaltaional decay.

2. Methods

2.1. Model Overview

Figure 1: Cartoon Representation of model in biological context. A) Model overview. Transcripts enter the sytem into the capped state at class 0 (no ribosomes bound). They enter the state at rate λ through transcription. Transcripts are free to move up and down ribosomal classes at rates κ for translation initiation and τ for elongation/termination. Transcripts can also be decapped and enter the decapped state at rate μ. Finally, upon reaching class 0 in the decapped state transcripts are fully degraded at rate δ. B) Probability of finding an mRNA in each class in the capped state. C) probability of finding an mRNA in each class in the decapped state. D) Joint probabilty of finding an mRNA in each class across each state. This reflect the total protein production potential.



The model captures some of the basic processes governing mRNA populations: transcript production, degradation and the process of translation (Figure 1A). Transcripts can exist in one of two states: capped and decapped which captures the the role of the 5' cap in mRNA protection and translation initiation. Capped transcripts are translationally competent, meaning that new ribosome can be loaded onto the transcript. Individual transcripts in the cell will be found with a set number of ribosomes (none, 1, 2, etc). The number of ribosomes on a transcript determines that transcripts polysomal class. The model seeks to determine how the population of transcripts of a single gene are distributed between ribosomal classes and capped and decapped states.

Transcripts enter into the model as defined by the transcription rate λ into the capped state with no ribosomes (polysome class 0) From capped class 0 a transcript can have two fates. The transcript can be decapped, thus marked for degradation at rate μ and move into the decapped class 0. Alternatively, a ribosome can initiate translation at rate $\kappa(1 - i/i_{\max})$ and be loaded onto the transcript and move it into capped class 1. Where i is the current transcript class and $i_{\max}$ is the maximal ribosomal occupancy on the transcript. The first term, κ is the average initiation rate on an empty transcript. For ribosomes initiating on transcripts already harboring ribosomes, this initiation rate is scaled to reflect the probability of a ribosome being present at or near the start codon. This attempts to account for the ribosomal density dependent effects on initiation and is called the density dependent initiation (DDI) model. As our model does not track ribosomal positions, we assume a uniform distribution of ribosomes across a transcript.

A ribosome on transcript can then elongate and terminate at a rate of $\tau' \times i$. After a ribosome fully elongates and terminates it leaves the transcript and the transcript falls to a lower class. The term τ' is calculated by using the average elongation rate τ_0 on a particular transcript divided by the length of the transcript $\tau' = \tau_0/\text{protein length in aa}$. As the number of ribosomes on a transcript increase, the probability of a ribosome being at the end of the transcript also increases, again following a uniform distribution. To better represent this mathematically, τ' can also be written as:

$$\tau' = \frac{\tau_0}{protein\ length} = \frac{\tau_0}{9 \times i_{\max}} \tag{1}$$

Where the saturated state of a transcript is denoted as $i_{\max}$, meaning that that transcript can no longer accept any more ribosomes. The factor nine arises from the average number of codons that is occupied by a ribosome. We can now formulate $\tau' \times i$ in the same way we formulated κ.

$$\tau' \times i = \frac{\tau_0}{9} \times \frac{i}{i_{\max}} \tag{2}$$

Where $\tau_0/9$ is the scaled elongation/termination rate, and $i/i_{\max}$ is the probability that a ribosome is at the end of a transcript.

Capped transcripts move through rounds of translation initiation and elongation-termination and distribute along the different polysomal classes. From any ribosomal class in the capped state the transcript can be decapped at rate μ and move into the decapped state while maintaining the same polysomal class. Decapped transcripts can no longer initiate new rounds of translation, but allow for currently loaded ribosomes to complete translation. This process represent co-translational decay, a common method of mRNA decay in eukaryotes (Hu 2009, Pelechano 2015, Collart 2020) After all ribosome complete translation, the mRNA is in decapped class 0 and completely degraded at a rate δ. The model produces two outputs. First, the total mRNA in each state and therefore the system (Figure 1B-D). Second, The distribution of the mRNAs in each mRNA in each ribosomal class. (Figure 1 B-D). The total protein output at steady state from our model can be obtained by calculating the average ribosomal class in the system by the total mRNA in the system (Figure 1D).

2.2. Formal Model Definition

We formalize the DDI model presented in Figure 1 by converting each state in to a series of ordinary differential equations (ODEs) representing the mRNA population for each polysomal class. The functional form of the capped mRNA sub population is:

$$
\begin{aligned}
\frac{dm_0}{dt} &= \lambda + \frac{\tau_0}{9}\frac{1}{i_{\max}}m_1 - \left(\kappa_0\left(1-\frac{0}{i_{\max}}\right)+\mu\right)m_0 \\
\frac{dm_1}{dt} &= \kappa(0)m_0 + \frac{\tau_0}{9}\frac{2}{i_{\max}}m_2 - \left(\frac{\tau_0}{9}\frac{1}{i_{\max}} + \kappa_0\left(1-\frac{1}{i_{\max}}\right)+\mu\right)m_1 \\
&\vdots \\
\frac{dm_i}{dt} &= \kappa(i-1)m_{i-1} + \frac{\tau_0}{9}\frac{i+1}{i_{\max}}m_{i+1} - \left(\frac{\tau_0}{9}\frac{i}{i_{\max}} + \kappa_0\left(1-\frac{i-1}{i_{\max}}\right)+\mu\right)m_i \\
&\vdots \\
\frac{dm_{i_{\max}}}{dt} &= \kappa_0\left(1-\frac{i_{\max}-1}{i_{\max}}\right)m_{i_{\max}-1} - \left(\frac{\tau_0}{9}\frac{i_{\max}}{i_{\max}}+\mu\right)m_{i_{\max}}
\end{aligned}
\tag{3}
$$

Table 1: State variables and model parameters for ODE model of mRNA populations. Variable $i_{\max}$ is in the domain of non-negative integers; all other variables are non-negative real numbers.

Symbol	Description	Unit
State Variables		
m_i	Abundance of mRNAs with a ribosome load of i in capped state.	$mRNA$
m_i^*	Abundance of mRNAs with a ribosome load of i in decapped state.	$mRNA$
Model Parameters		
i	ribosomal load index	Ribosome
$i_{\max}$	Maximum number of ribosomes able to bind to mRNA; defines number of state variables and is a function of gene length.	Ribosome
$\kappa(i)$	Translation initiation rate for unmarked mRNAs with a ribosome load of i.	$1/s$
$\tau(i)$	Translation completion rate for the marked and unmarked mRNAs with a ribosome load of i.	$1/s$
$\mu(i)$	Marking rate for unmarked mRNAs with a ribosome load of i.	$1/s$
λ	Production rate of newly produced, ribosome free, and unmarked mRNA to the m_0 class.	$mRNA/s$
δ	Removal rate of marked mRNA with a ribosome load of 0 from the m_0^* class.	$1/s$

Similarly, the functional form of the decapped mRNA sub population is:

$$\begin{aligned}
\frac{dm_0^*}{dt} &= \mu m_0 + \frac{\tau_0}{9}\frac{1}{i_{\max}} m_1^* - \delta m_0^* \\
\frac{dm_1^*}{dt} &= \mu m_1 + \frac{\tau_0}{9}\frac{2}{i_{\max}} m_2^* - \tau(1) m_1^* \\
&\vdots \\
\frac{dm_i^*}{dt} &= \mu m_i + \frac{\tau_0}{9}\frac{i+1}{i_{\max}} m_{i+1}^* - \tau(i) m_i^* \\
&\vdots \\
\frac{dm_{i_{\max}}^*}{dt} &= \mu m_{i_{\max}}^* - \frac{\tau_0}{9}\frac{i_{\max}}{i_{\max}} m_{i_{\max}}
\end{aligned} \tag{4}$$

A less constrained version of the model does not account for the DDI effects and doesn't scale κ by $(1 - i/i_{\max})$. This is the density independent initiation (DII) version of the model. Parameters and their units are fully defined in Table 1. All parameters are assumed to fixed for a given gene, but may vary between genes.

2.2.1. Analytical steady state solutions of the capped transcript population

Analytical exploration of the model's capped system presents no closed form solution for the capped system. However, the model solution can be represented in the following form,

$$\vec{m} = \frac{\lambda}{\mu}\vec{p}_m \tag{5}$$

Where $\vec{m}$ is a vector of the steady state mRNA abundances in each polysomal class. $\vec{m}$ is calculated from by scaling the vector $\vec{p}$, which represents the distribution of the mRNA across the polysomal classes, by transcript production rate λ and the decapping rate μ scale s. The individual components of $\vec{p}$ are functions of i, $i_{\max}$, the translation initiation rate κ, the elongation rate τ_0 and μ and have no closed form solution.

2.2.2. Analytical steady state solutions of the decapped transcript population

The solution for the decapped system is dependent on the underlying distribution of the capped system and can be represented as:

$$\begin{aligned}
m_0^* &= \frac{\mu}{\delta}\sum_{j=0}^{i_{\max}} m_j \\
m_1^* &= \frac{\mu}{\tau}\sum_{j=1}^{i_{\max}} m_j \\
&\vdots \\
m_i^* &= \frac{\mu}{i\,\tau}\sum_{j=i}^{i_{\max}} m_j \\
&\vdots \\
m_{imax}^* &= \frac{\mu}{i_{\max}\,\tau}\sum_{j=imax}^{i_{\max}} m_j
\end{aligned}$$

We can simplify the model by converting the mRNA quantity m_j to the probability p_j by 5. Additionally, for any $i = j$ where S_j is cumulative probability from $i =$ class j to $i = i_{\max}$.

$$S_j = \sum_{i=j}^{i_{\max}} \vec{p_i} \tag{6}$$

Now the solution becomes,

$$
\begin{aligned}
m_0^* &= \frac{\lambda}{\delta} S_0 = \frac{\lambda}{\delta} \\
m_1^* &= \frac{\lambda}{\tau} S_1 \\
&\vdots \\
m_i^* &= \frac{\lambda}{i\,\tau} S_i \\
&\vdots \\
m_{imax}^* &= \frac{\lambda}{i_{\max}\,\tau} S_{i_{\max}}
\end{aligned}
\tag{7}
$$

2.3. Calculation of the decapped mRNA population

The total transcript population in the decapped state does not have a closed form solution. However it can be summarized as follows,

$$
m_{tot}^* = \sum_{i=0}^{i_{\max}} m_i^* = \frac{\lambda}{\delta} + \frac{\lambda}{\tau} S_1 + \ldots + \frac{\lambda}{i\tau} S_i + \ldots + \frac{\lambda}{i_{\max}\tau} S_{i_{\max}}
$$

This can be further shortened to:

$$
m_{tot}^* = \lambda \left(\frac{1}{\delta} + \frac{1}{\tau} \vec{S} \cdot \vec{l} \right) \tag{8}
$$

Where $\vec{S}$ is a vector of all the cumulative sums and $\vec{l}$ is a vector of $1, 1/2, ..., 1/i, ..., 1/i_{\max}$.

2.4. Probability distribution in the decapped state

To get the probability distribution of transcripts across the decapped state we can divide $\vec{m^*}/m_{tot}^*$ which results in,

$$
p_0^* = \frac{1}{1 + \frac{\delta}{\tau} \vec{S} \cdot \vec{l}} \tag{9}
$$

$$
p_j^* = \frac{S_j}{j(\frac{\tau}{\delta} + \vec{S} \cdot \vec{l})} \quad \text{for } j = 1, 2, ..., i, ..., i_{\max} \tag{10}
$$

2.5. Calculation of the total mRNA population and its distribution between capped and decapped states

The total mRNA (M_{tot}) in the system is defined by,

$$
M_{tot} = \frac{\lambda}{\mu} + \lambda \left(\frac{1}{\delta} + \frac{1}{\tau} \vec{S} \cdot \vec{l} \right) \tag{11}
$$

To understand how mRNA is divided between we start with the probability of finding an mRNA in the capped state.

$$p_{mtot} = \frac{1}{(1 + \frac{\mu}{\delta} + \frac{\mu}{\tau}\vec{S} \cdot \vec{l})}$$

Then you calculate the odds,

$$odds_m = \frac{1}{\mu(\frac{1}{\delta} + \frac{1}{\tau}\vec{S} \cdot \vec{l}} \tag{12}$$

2.6. Calculating expected ribosomal load and protein production

The expected ribosomal load for either the capped or decapped state is calculated by:

$$E(ribosome) = \sum_{i=0}^{i_{\max}} j \times p_i \tag{13}$$

Where $\vec{p_m}$ is the distribution in either state and i is the polysome class.

To find the global mean ribosomal load we obtain,

$$\text{Total Ribosomal Load} = p_{mtot} \times E(ribosome)_{mtot} + (1 - p_{mtot}) \times E(ribosome)_{mtot^*} \tag{14}$$

2.7. Numerical solution implementation in R

Code to solve the model was written in the R package Ribosome (`https://github.com/rurquidi/Ribosome`). To solve the capped subsystem of the model, the solve.tridiag algorithm from limSolve package (V 1.5.6) (Soetaert,K 2009). The decapped solution was obtained by using the capped solutions into 7. Utility functions, plots and statistics were created using R (v 3.6) (R core team), and data.table (v1.14.0) (Dowle 2021).

2.8. Data Sources

In order to biologically contextualize and illustrate our model's behavior, we will focus on parameter ranges derived from the literature. The range of $i_{\max}$ is determined from the distribution of protein lengths obtained from yeast (*saccharomyces cerevisiae*) and the plant *Arabidopsis thaliana*. To determine $i_{\max}$, protein lengths are divided by the average number of codons covered by a ribosome, which is 9 codons (Figure 2A and C). The range of $i_{\max}$ is 48 $\pm$ 36 for yeast and 47 $\pm$ 30 for Arabidopsis. Protein lengths were extracted from the Ensembl (version 109) and Ensembl plants (version 56) respectively (Cunningham 2022, Yates 2022, Kinsella 2011). The marking rate between the capped and uncapped system was approximated from the protein half-lives from Presnyak 2015 for yeast (Figure 2B) and Sorenson 2018 for Arabidopsis (Figure 2D). We approximated gene specific μ from the half

lives with the following:

$$\mu_i = \frac{ln(2)}{t_{1/2_i}}$$

Where $t_{1/2}$ is the half-life. The resulting range of μ is from $1.3 \times 10^-3 \pm 1.8 \times 10^-3$ for yeast and $1.7 \times 10^-4 \pm 2 \times 10^-4$ for Arabidopsis.

Translation initiation and average elongation rates (κ and τ_0) were obtained for Yeast from Duc and Song 2018. In Duc and Song 2018, the authors used 850 highly translated transcripts from the ribo-seq dataset from Weinberg 2016. They employed a TASEP model to estimate the initiation rates and correct the empirical elongation rates from the footprint distributions. We calculated an average gene specific elongation rate from the corrected elongations rates. We scale the each gene specific initiation rate by dividing it by the gene specific elongation rate This simplifies the model behavior to one generalized parameter with a unique response (Figure 2E). The scaled initiation rate ranges from $0.1s^{-1}$ to $0.001s^{-1}$.

The transcription rate, λ only acts as a scaling factor throughout the model and does not affect the distribution of the ribosomes. For solutions provided in this work λ has been set to one. However, as a point of reference, the transcriptomic results from Weinberg 2016 are included in Figure 2F. In short, reads per kilobase million from Weinberg were further converted into a log10 fold change based on the median expression level. Figure 2F shows that the absolute range of transcriptional expression ranges just under 5 orders of magnitude.

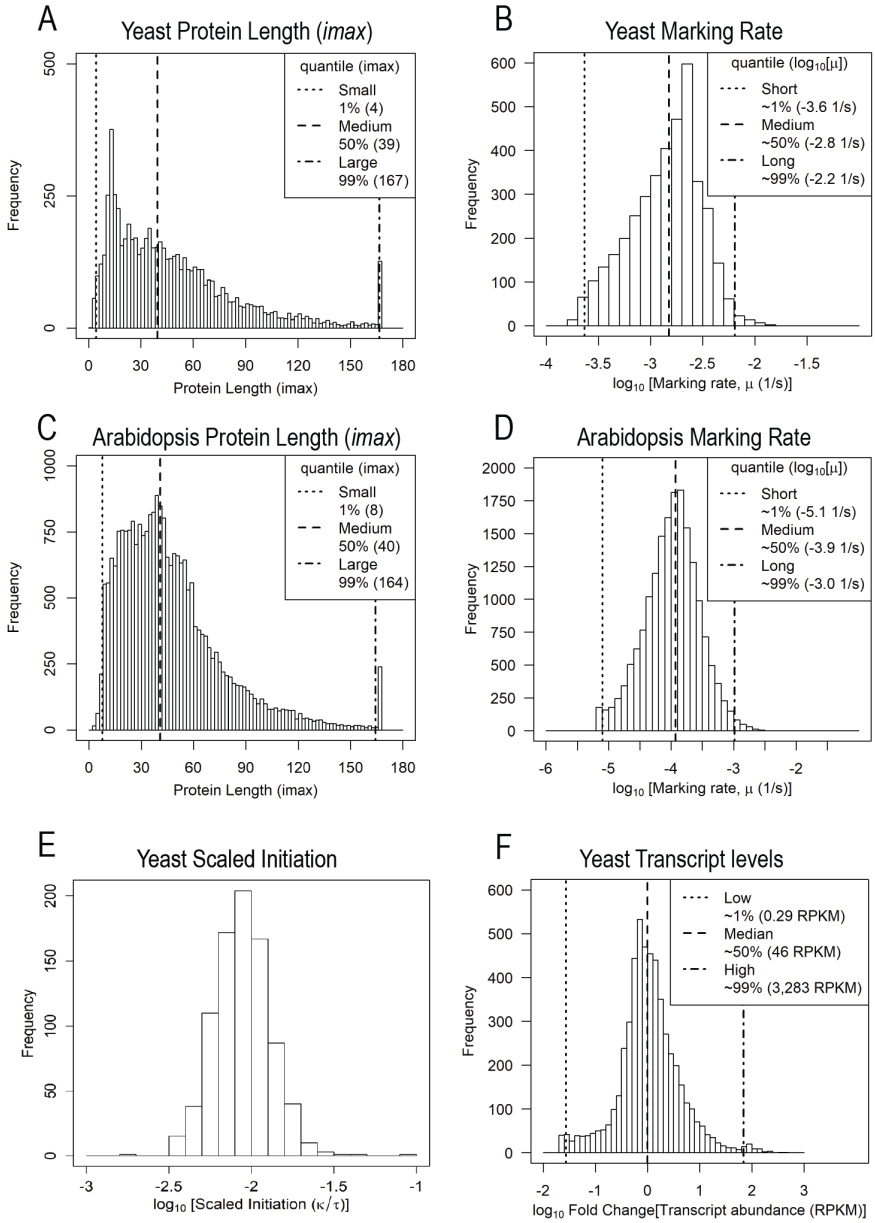
The degradation rate δ only determines the accumulation of transcripts in the m_0^* class, which for simplicity of interpreting results has been set to be $>> \tau$ and thus will not accumulate transcripts in m_0^*.

The empirical mean ribosomal load (MRL) for the 850 genes in Duc and song 2018 was calculated from the mRNA-seq read per kilbase million (mRNA RPKM) and the ribo-seq footprints (RPF RPKM) from Weinberg 2016. The following equation was used.

$$MRL_i = \frac{RPF\ RPKM_i}{mRNA\ RPKM_i \times \frac{200}{lengthmRNA_i}} \tag{15}$$

Where the gene specific scaling factor $\frac{200}{lengthmRNA_i}$ corrects for the bias in read counts due to longer transcripts producing more fragments. The value 200 arises from the average fragment size of a library prep and can be adjusted according the experimental method used.

Figure 2: Histograms of empirical values of model parameters. A) Yeast protein lengths. B) Yeast half-life C) Arabidopsis Protein Lengths. D) Arabidopsis Half-Life. E) Yeast Scaled elongation rates (Translational initiation rate/average translation elongation rate) on a per gene basis. F) Log 10 Fold Changes between all transcripts compared ot the median transcript expression in yeast.



3. Results

3.1. Model provides a unique distribution of mRNAs across ribosomal classes for each scaled initiation rate

The capped solution splits the two funtions of the marking rate μ; Its effect on transcript number and its effect on transcript distribution. And allows for their separate analysis. The mRNA population is defined by the ratio of the transcription rate λ to μ.

The probability of finding a transcript in each ribosomal class for a particular parameter set is dependent on the initiation rate κ, elongation rate τ_0 and μ. In figure 3A the mRNA distribution in the capped state is presented for four different scaled initiation rates (as shown in Figure 3B) for a median length protein with a long (52 minute) half life. To summarize the model results across a range of parameters a heatmap where each row is the distribution of mRNA at a particular scaled initiation rate is shown. As the scaled initiation rate increases the density moves to the right and spreads out in the capped system. The distribution in the capped system is bounded at class 0 and class $i_{\max}$ and roughly symmetrical away from the boundaries (Figure 3B).

The decapped system is centered around the lowest classes (Figure 4). This is due to the distribution in 9 having the following arrangement $S_0 = 1$ and $S_0 \geq S_1 \geq ... \geq S_i \geq ... \geq S_{imax}$ dependant on the distribution of $\vec{m}$ of the capped state. Exploring the result we find a few properties of our system. Transcription rate (λ) again serves only to scale the entire system. The first decapped class's population m_0^* is only dependent on the degradation rate (δ). The whole system incorporates the transcript distributions from both capped and decapped systems. This is apparent in the bimodal peaks at higher scaled initiation values, with a peak at lower ribosomal load representing the decapped system and a high ribosomal load from the capped system (Figure 5).

Figure 3: mRNA distribution in Capped state. A) Individual distribution profiles for four scaled initiation values B) Heatmap of model output across a range of scaled initiation values. Lines represent slice represented in A). Results produced with $i_{\max}$ of 39 and a long half life of 52 minutes (99th percentile). Color bar shows probability of finding mRNA in particular ribosomal class.

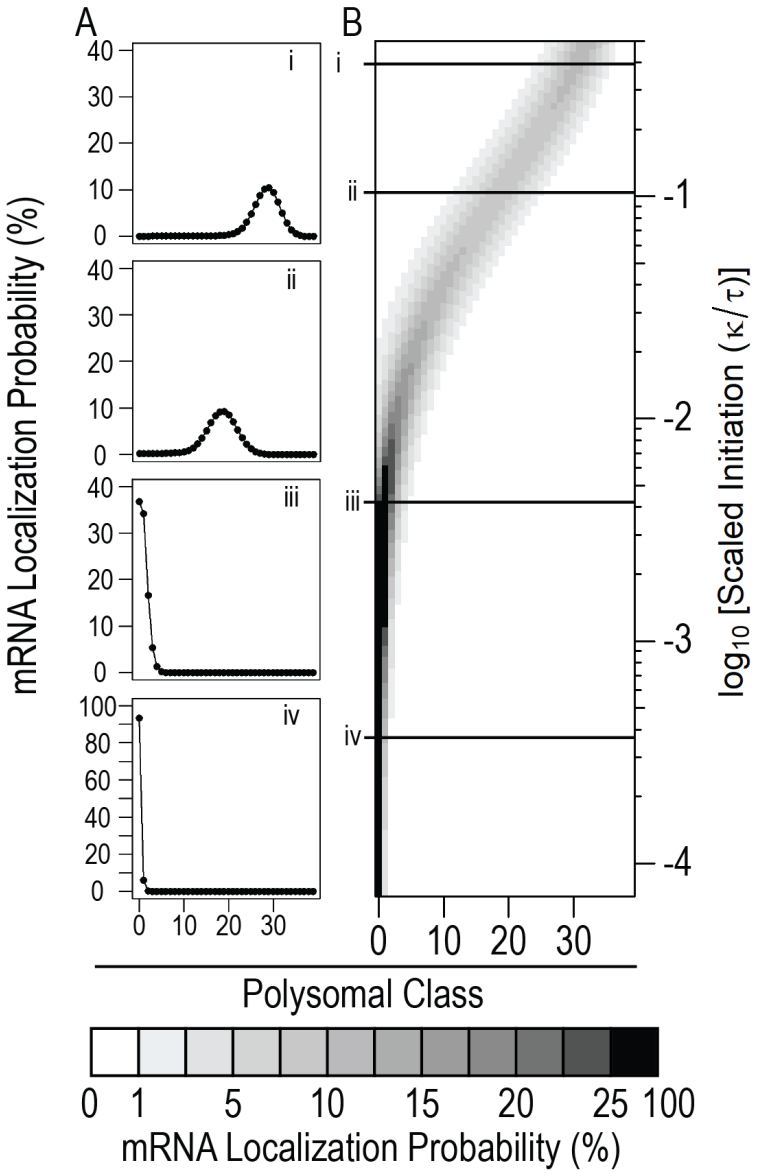
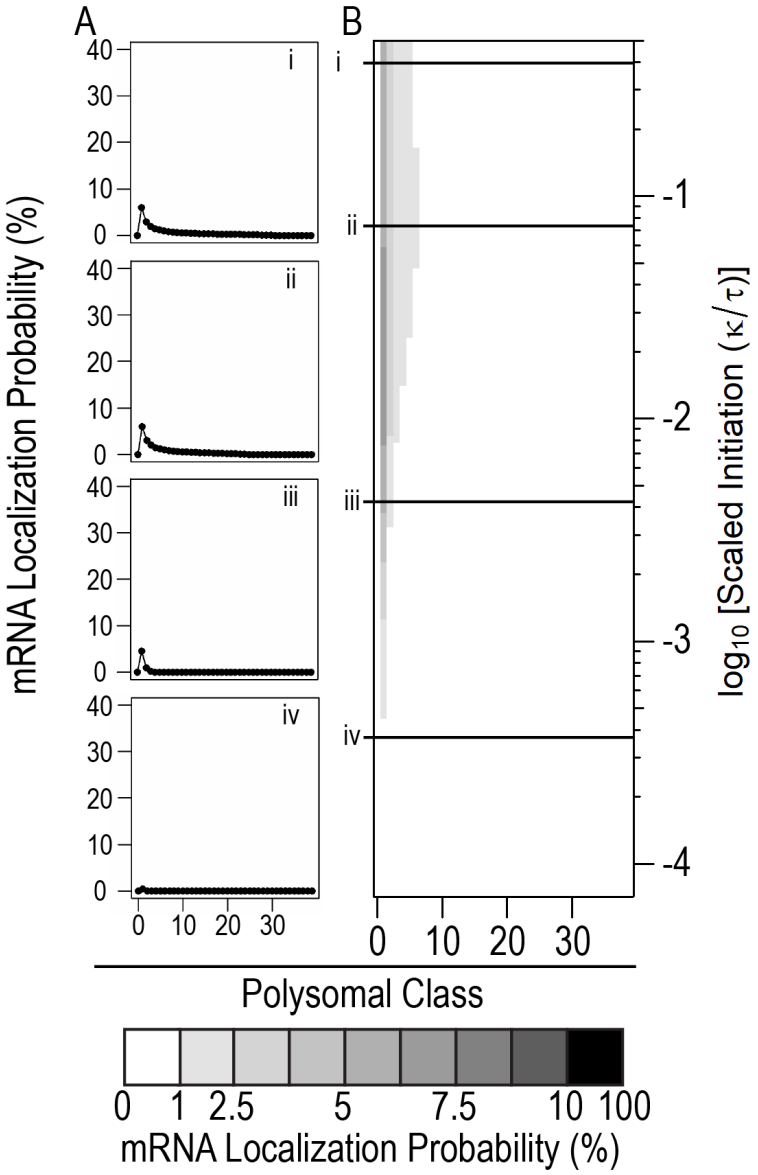


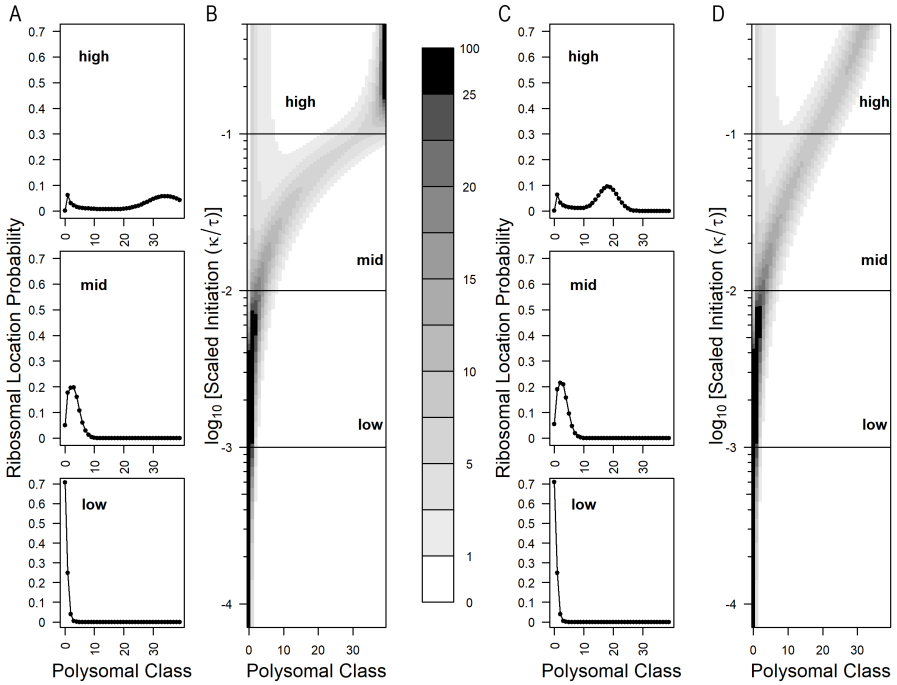
Figure 4: mRNA distribution in decapped state. A) Individual distribution profiles for four scaled initiation values B) Heatmap of model output across a range of scaled initiation values. Lines represent slice represented in A). Results produced with $i_{\max}$ of 39 and a long half life of 52 minutes (99th percentile). Color bar shows probability of finding mRNA in particular ribosomal class.



3.2. Initiation interference due to increasing ribosome density is noticeable beginning at moderate to high ribosomal loads

On a physical level, ribosomes initiating on a transcript depend on having sufficient space around the start codon. Transcripts at with higher ribosomal loads therefore have a higher probability of having an obstructed start codon. In order to explore the effect of ribosomal load on initiation we created two versions of the model. The first is dependent on the ribosomal load of a transcript, the density dependent initiation (DDI) model. The DII model is is independent on the ribosomal load of a transcript, the density independent model (DII) presented in figure 5 A-B, with individual probability profiles presented in 5A and the summary heatmap in 5B. The DDI model is presented in Figure 5C-D. Individual probability profiles are generated at low, mid and high scaled initiation rates as determined from Duc and Song 2018. Note that in the DII model the system saturates just above the high scaled initiation rate, while the DDI model doesn't. Additionally, model profiles are very similar at low to mid scaled initiation values for both models. Figures 5 B and D suggest that the density dependent effects start appearing between mid and high scaled initiation rates. Polysome profiling experiments usually ony resolve 8-10 ribosomal peaks, with the majority of the signal arising from polysomes 2-5. This generally agrees with single molecule imaging of nascent peptides where ribosomal densities are on the range of 0.5% - 30% (Morisaki 2016, Wang 2016, Wu 2016, Yan 2016). While both the DII and DDI models show similar behavior and low density at low to mid scaled initiation values, empirical evidence and physical reality would indicate that a transcript is unlikely to ever reach saturation. This supports the DDI model. From here on out all results will be solely based on the DDI model.

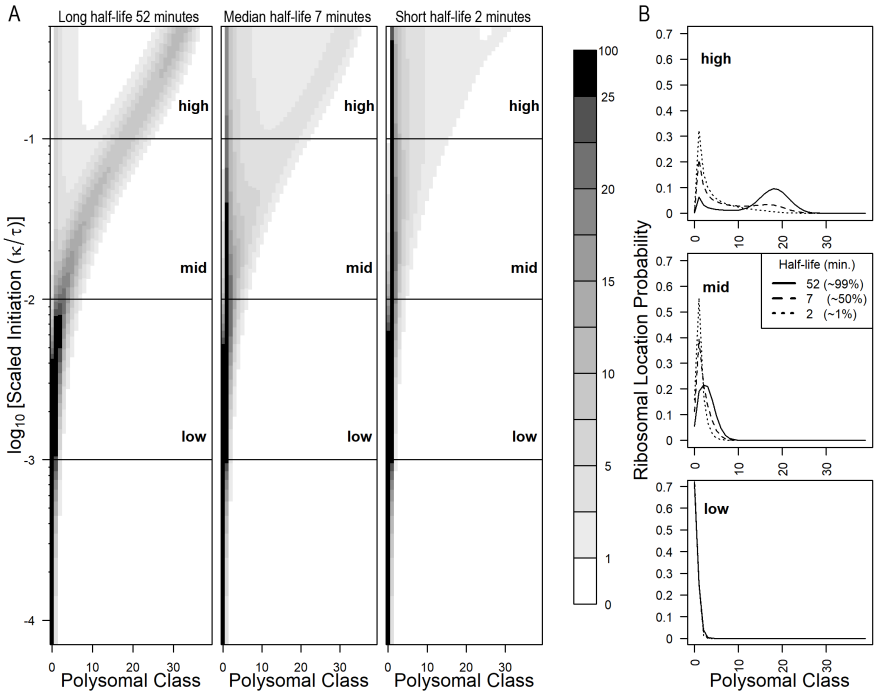
Figure 5: Comparison of density independent initiation (DII) and density dependent initiation (DDI) in full system. A) DII individual density profiles for low (0.001), mid (0.01) and high (0.1) scaled initiation values. B) DII density Heatmap for the full system. C) DDI individual density profiles for low (0.001), mid (0.01) and high (0.1) scaled initiation values. B) DDI density Heatmap for the full system. All results calculate with $i_{\max} = 39$ and long half life of 52 minutes (99th percentile).



3.3. Higher marking rates reduce capped state ribosomal loads

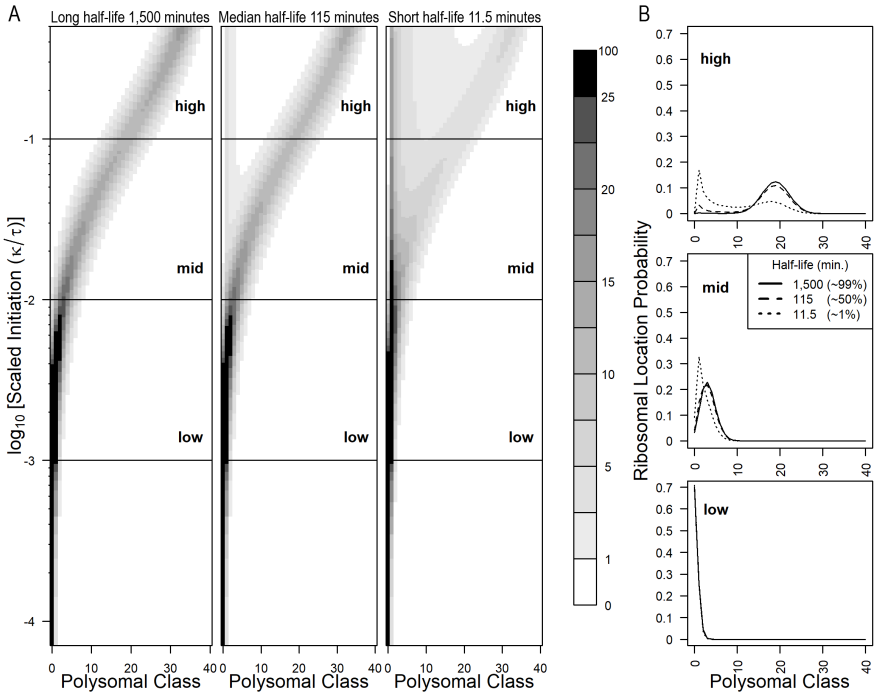
The interplay between the translational machinery, mRNA degradation machinery and mRNA properties such as codon optimality, secondary structure or modifications all have been reported to play a role in mRNA stability (Wu 2019, Medina-Munoz 2021, Bae and Collier 2022). To explore the role of mRNA stability on mRNA populations we varied the marking rate from the 1st percentile, median and 99th percentile values as determined from the half-life values in Presnyak 2015. We find that as half-life decreases the distribution of mRNAs change in two ways. First there is shift to lower ribosomal classes in the capped state (Figure 6). This is likely due to the mRNAs leaving the capped state at a higher rate and driving the equilibrium towards lower ribosomal loads. Secondly, as half-life decreases, a larger proportion of the mRNA is found in the decapped state. This is further explored later.

Figure 6: Shorter half-lives reduce ribosome load in capped system in yeast. A) Heatmaps for the full system. Left) long half life (52 minutes) Center) median half life (7 minutes) Right) short half life (2 minutes) B) individual density profiles for low (0.001), mid (0.01) and high (0.1) scaled initiation values for each half life value. All results calculate with $i_{\max} = 39$.



Multicellular eukaryotes, such as plants, face a different set of environmental challenges and tend to have slower translation initiation and elongation rates as well as slower cell division when compared to single celled organisms. This is highlighted by the current gold standard study of mRNA half-lives in the model organism *Arabidopsis thaliana*, where the half-lives measured are one two two orders of magnitude longer than those in yeast. To explore this we ran the model using the same scaled initiation rate as in yeast, the median Arabidopsis $i_{\max}$ of 41, and Arabidopsis half-lives (long half life (1,500 minutes), median half life (115 minutes), and short half life (11.5 minutes)). As expected, the longer half-lives have higher ribosomal loads and are mostly in the capped state.

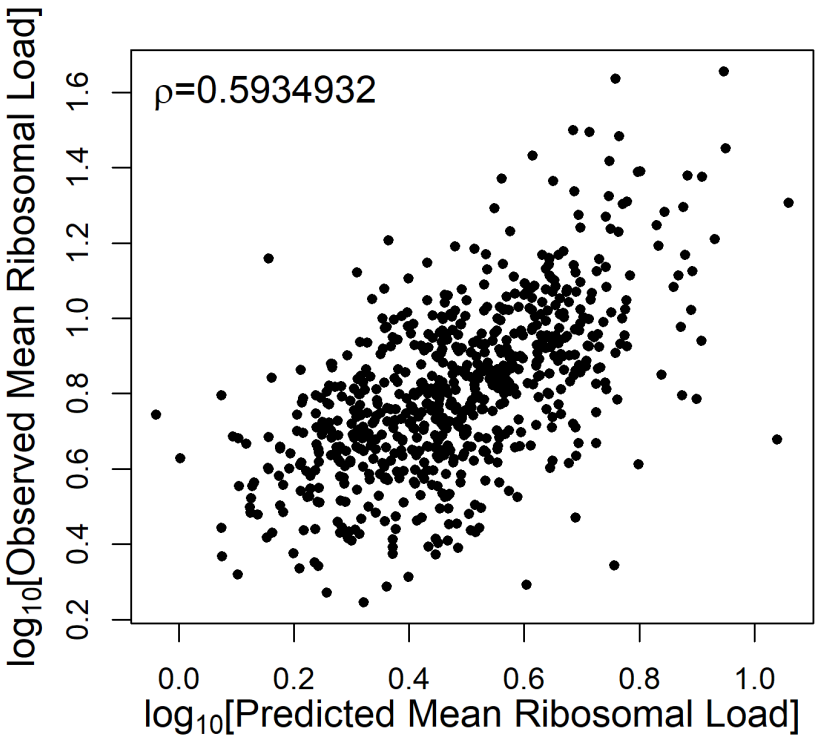
Figure 7: Longer half-lives in Arabidopsis result in a smaller effect on ribosomal load in the capped system. A) Heatmaps for the full system. Left) long half life (1,500 minutes) Center) median half life (115 minutes) Right) short half life (11.5 minutes) B) individual density profiles for low (0.001), mid (0.01) and high (0.1) scaled initiation values for each half life value. All results calculate with $i_{\max} = 41$.



3.4. Model replicates ribo-seq and single molecule measurements of translation

Gene specific MRL were calculated for the genes analyzed in Duc and Song 2018 and compared to the empirical MRL calcualted from raw data from Weinberg 2016. Model predictions of MRL showed a significant positive correlation to empirical MRLs. This result is reassuring as the model performs well despite no model fitting being performed. Model performance is further corroborated with single molecule imaging analyses. Rescaling ribosome abundances from each single molecule study to an $i_{\max}$ of 40 results in loads of 1, 2.4 and 4 ribosomes from (Morisaki 2016), 3 ribosomes (Yan 2016), 4 ribosomes (Wang 2016) and 1.6 (Wu 2016). All of which align with low to mid scaled initiation MRL predictions. Finally, mRNA distributions for the whole system agree with signal from polysome traces (Lokdarshi 2020, Dasgupta 2023).

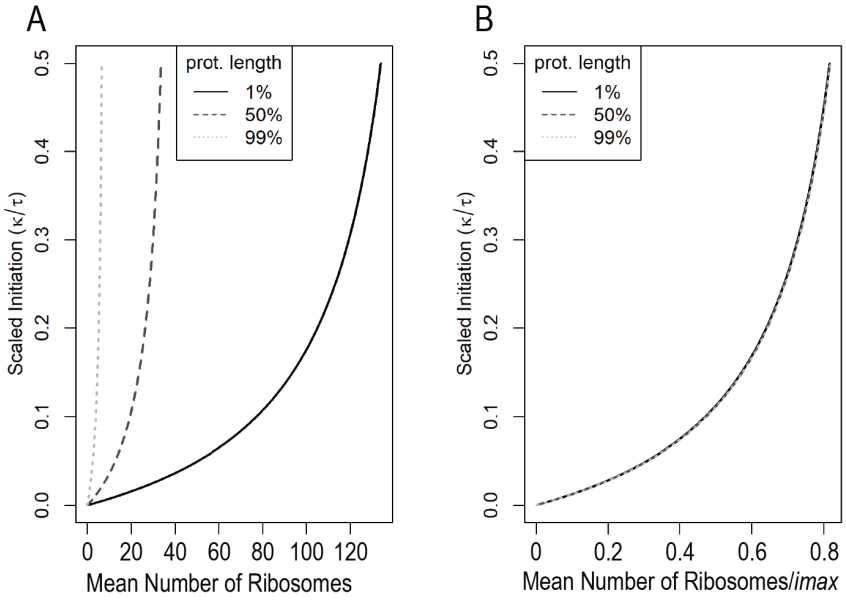
Figure 8: Predicted mean ribosomal loads coincide with observed mean ribosomal loads from Weinberg 2016. Using the 850 genes from Duc and song 2018, marking rates from Presnyak 2015 were mined. Gene specific MRL were calculated and compared to the empircal MRL. Spearman's ρ was calculated and found to be significant, pvalue $< 10^{-16}$



3.5. Under equilibrium, protein length does not affect the ribosomal density on transcripts

One particular aspect of analyzing the model at equilibrium is that while the total number of ribosomes on a particular transcript is dependent on length (Figure 10A), the density per unit length is not (Figure 10B). In other words, the flux of ribosomes through a transcript is independent of the length of the transcript. At equilibrium and under the same parameters, except for $i_{\max}$, the number of ribosomes initiating and terminating is the same regardless of $i_{\max}$. This property means that for any model solution, if the ribosomal classes are converted into densities (i.e. dividing the x axis of figure 5 by 39), you now have a general solution for all transcripts with those parameters. It also implies that the rate of protein production is independent of transcript length. 0

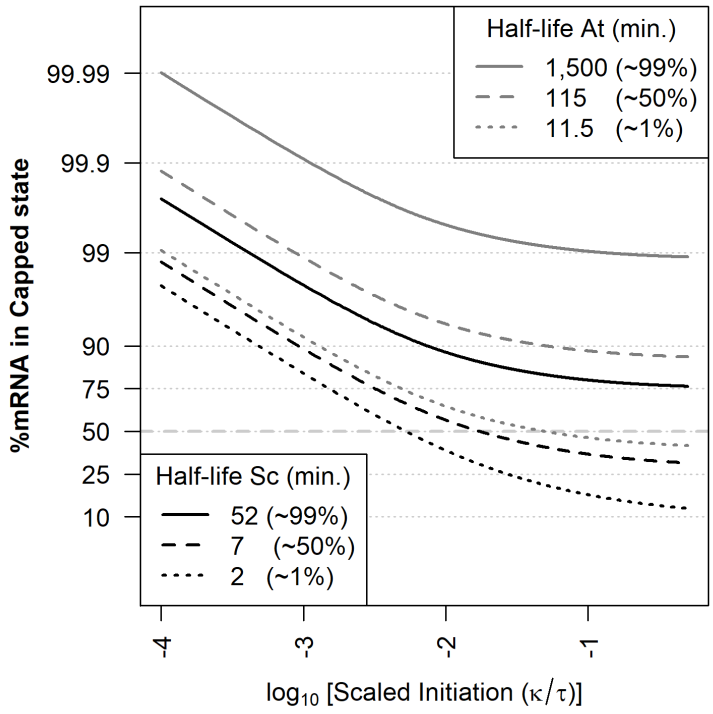
Figure 9: The ribosomal density on a transcript is independent of their length. A) Ribosomal load per transcript is higher for longer transcripts. B) When the ribosomal load is corrected for the length of the transcript, the ribsomal density collapses to the same curve for all transcripts under the same parameters. For this plot Arabidopsis parameters were used $i_{\max} = 8$ (1% percentile), 40 (1% percentile), 164 (99% percentile), 1500 minute half life, over the full scaled initiation range 0.0001- 0.5.



3.6. Marking rate and ribosomal load determine mRNA distribution between states

As shown in previous results, shorter half-lives (faster marking) leads to lower ribosomal load and a shift towards the decapped state. To explore this shift we can use the results in equation 5, which splits the mRNA population in the capped state and the distribution of reads within and equation 8 the total transcript population to derive the log odds in equation **??**. Using equation **??**, we can see under which parameter regimes mRNA is more abundant in the capped state. We produced output across all scaled initiation values and under the 1%, 50% and 99% percentiles for marking rates in both yeast and Arabidopsis (Figure 10). We note two patterns. First as the scaled initiation rate increases the amount of mRNA in the decapped class increases. Secondly, shorter mRNA half-life bias transcripts towards the decapped class as previously seen in Figures 6 and 7.

Figure 10: Log odds of finding mRNA in the capped state for a range of marking rates in yeast and Arabidopsis.



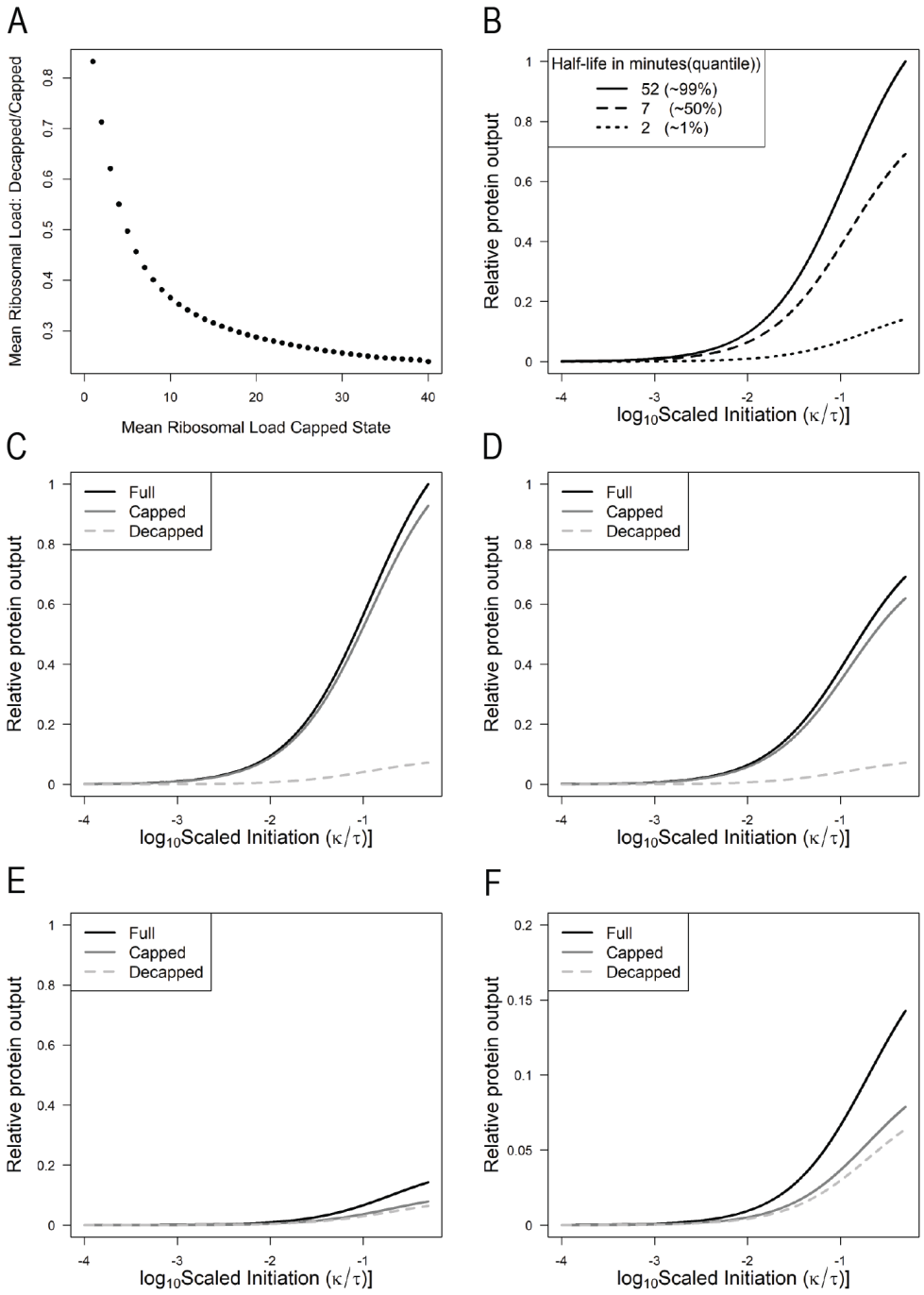
3.7. Decapped state can be a significant source of protein production

Protein production is a function of the rate at which ribosomes leave the transcript and the underlying total number of transcripts. Naively, the average protein production rate (PPR) for the capped state could be estimated by multiplying the total number of transcript by the initiation rate. However, this ignores the DDI and marking effects in the system and the actual PPR will be lower. Contributions from the decapped state are harder to estimate from the parameters as the distribution is dependent on the capped state's distribution.

Another way of comparing relative protein output (RPO) is to calculate the mean ribosomal load (MRL) for each state using 13 and multiply is by the total number of transcripts in their respective state. The full systems protein production can be found with 14. In our model the MRL of the capped system will always be equal to or greater than the decapped MRL Figure 11 A). As the scaled initiation rate increases and the MRL of the capped class increases, the distributional properties of the uncapped class mean that it will always be proportionally less to the MRL of the capped class.

The RPO of a state is dependent on both the MRL of each state and the total transcript population of each state (Figure 11B). As the scaled initiation rate increases RPO will increase. The marking rate μ plays an important role in controlling both the underlying mRNA population and the distribution of transcripts in the capped state. Therefore, and increase in marking rate reduces RPO through two mechanisms. Each of the three cases in Figure 10 B, has been broken down into the RPO contributions from the capped and decapped states (Figure 11 C-E). A surprising finding from out model is that under certain biologically relavant parameter combinations, almost half of all protein production can arise from the decapped state. This can be seen inf Figure 11 F, wherea substantial portion of all protein arises from the decapped state. The reason behind this is despite the the relative MRL of the decapped state being lower than the capped state as scaled initiation rises, the amount of mRNA in the decapped state rises faster.

Figure 11: Estimated average protein production in yeast. A) Ratio of the expected ribomal load of Decapped over Capped states for a protein of $i_{\max}$ 40 and low marking rate for distribution of the capped state that result in MRL of 0-40. B) Protein protein production across different marking rates. Total protein production is normalized to the maximal protein production across all parameters. C-F) Contribution of capped and decapped states to total protein production. C) Low marking D) Medium marking E) High marking F) zoom in of high marking result.



4. Discussion

In this study we develop and study a novel model of mRNA populations which includes the contributions of mRNA transcription, the initiation and termination of translation as well as mRNA degradation through 5' decapping and cotranslational decay.

Model Steady State Behavior

(Add relevant equation and figure references to points below)

1. Although our model is only a very simplified description of protein synthesis (translation), it includes interactions with the process of mRNA degradation which is has been underexplored (Yadav 2021).
   (a) The process of translation is dependent on the underlying population of capped, translationally competent mRNAs.
   (b) However, empirical measurements suggest that $\sim$ 12% of transcripts are undergoing cotranslational decay (Pelechano 2015).
   (c) Our model includes 5' mRNA decapping followed by cotranslational decay, permiting the analysis of the decapped mRNA state (called degratome in Ma 2020), changes in MRL for capped and decapped states and the contribution of cotranslational decay to protein production.
2. In addition to being more biologically realistic, structuring the mRNA population by its polysome classes (ribosome load) and the status of its 5' cap allows us to understand how the the rates mRNA production λ, decapping μ, elongation τ, and clearance rate δ shape the steady state distribution of a gene's mRNA population across polysome classes and capping state. Overall, we find that
   (a) Capped mRNA population
      i. Analytical and numerical solutions show that the transcription rate λ acts as a scaling factor, such that the entire mRNA population is proportional to λ. This indicates that as long as $\lambda \neq 0$, the distribution of transcripts across polysomal classes is independent of transcription.
      ii. The sum of the of capped mRNA classes $\hat{\vec{m}}$ is solely determined the ratio of λ to the 5' decapping rate μ, i.e. $\sum_{i=0}^{i_{i_{\max}}} \hat{m}_i = \lambda/\mu$.

A. This result suggests that, under biologically realistic conditions, as μ decreases, the capped population increases. Further exploration of the empirical data used for this model suggest that μ buffers λ, with most genes implying a constant λ (supplemental figure 1).

iii. Impact of κ/μ

A. When $\kappa/\mu \ll 1$, the capped mRNA distribution $\hat{\vec{m}}$ is greatest in the polysome class $i = 0$ and declines rapidly with i.

B. As κ/μ increases, the capped mRNA distribution shifts away from the lower bound of $i = 0$ appears to follow a truncated gaussian distribution.

C. At very high and generally unrealistic values of κ/μ, density dependent interference effects slow MRL from reaching $i_{\max}$. Eventually at $\kappa/\mu > 10$ saturation is reached.

(b) Decapped mRNA population

i. The steady state abundance of the decapped, ribosome free mRNA class class $\hat{m}_0^*$ is decoupled from the dynamics of the rest of the population. This decoupling has a number of important implications.

A. The steady state abundance of $\hat{m}_0^* = \lambda/\delta$ and, thus, depends only on the ratio of the mRNA transcription rate λ to the degradation rate δ. If the transcription rate λ of new, capped, but ribosome free mRNAs $\hat{m}_0$ is lower than the per capita degradation rate of decapped, ribosome free mRNAs *delta*, such that $\lambda \ll \delta$, then $\hat{m}_0^* \ll 1$ and our model predicts that there will be few mRNAs in the $\hat{m}_0^*$ class.

B. Because $\hat{m}_0^*$ has no impact on the rest of the mRNA population, this result allows us to greatly simplifying our analysis since we need not consider $\hat{m}_0^*$ nor the mRNA degradation parameter δ.

ii. Shifting our focus to the steady state abundance of the decapped, ribosome occupied mRNA classes $\hat{m}_i^*$ where $i > 0$

A. The distribtuion of $\hat{m}_i^*$ with $i > 0$ depends on the gene specific ribosome elongation rate τ_0 (where 'elongation' includes the ribosome's reading of the mRNA's stop codon) and the distribution of capped mRNA $\hat{m}_i$ with $i > 0$. This implies that

B. Item 1

C. Item 2

D. The density of $\hat{m}_i^*$ monotonically decreases with i. Thus, the decapped mRNA distribution is generally skewed towards lower polysomal classes.

E. The MRL of the decapped class is bounded, such that it never exceeds than MRL of the capped class.

(c) Combined distributions of the capped and decapped polysome class

i. Distribution is unimodal when $\sum \hat{\vec{m}} >> \sum \hat{\vec{m}}^*$ or when $\kappa/\mu << 1$.

ii. The distribution is bimodal otherwise. The bimodal peaks arised from the distributions from the capped and decapped states.

iii. Increasing marking rate and scaled initiation rate both increase the proportion of transcripts in the decapped state.

A. The model formalizes the interplay between mRNA decapping μ and translation initiation κ.

B. As expected, increasing μ results in an increase in the proportion of decapped transcripts $\hat{\vec{m}}^*$ compared to capped transcripts $\hat{\vec{m}}$.

C. However, increasing ratio of initiation to elongation rates κ/τ also results in an increase of $\hat{\vec{m}}^*$.

D. As κ/τ (scaled initiation rate) increases the MRL of the capped population $\hat{\vec{m}}$ increases, transcripts enter the decapped state at higher polysomal classes and thus take longer to reach m_0^*.

(d) A surprising prediction from our model is that genes with high marking rates (μ 5×10^{-3}) can have almost half of their protein produced in the decapped state.

i. As μ increases, a greater proportion of the MRL arises from the decapped states.

ii. This increase is mainly a function of a shift in transcripts to the decapped state $\hat{\vec{m}}^*$ due to larger μ.

iii. As the decapped MRL $\leq$ the capped MRL for any μ, or scaled initiation rate κ/τ.

iv. The high relative protein production from the decapped state suggests that high marking rate transcripts can produce more protein than expected from their capped MRL alone.

4.1. Model Validation

In addition to studying the general behavior of our model, we validate this behavior using empirically based parameter values from the literature. In general, we find that our model's predictions of mRNA distributions are highly consistent with a wide range of empirical data. For example,

1. Previously observed polysome gradients (Lokdarshi 2020, Dasgupta 2023),
2. Ribo-seq (Weinberg 2016, Figure 8)
3. single molecule measurements of translation (Morisaki 2016, Yan 2016, Wang 2016, Wu 2016, Section 3.4).
4. Predicted capped to decapped mRNA ratios also align with empirical measurements (Pelechano 2015)
5. MRL is largely independent of the protein length (single molecule imaging of translation (Wu 2016), Figure 9).
6. Fraction of mRNA in the decapped class under XXX conditions are consistent with population wide estimates (Pelechano 2015, Figure 10).

4.2. Model limitations, extensions and future work

1. Our model's assumptions about the process of mRNA decapping, the continued competence of ribosomes present prior to decapping, and degradation of mRNA solely from the decapped and ribosome free class m_0^* closely resembles the biological process of co-translational mRNA decay.
   (a) The existence of co-translational mRNA decay is well established (Sorenson 2018).
   (b) However, other mechanisms exist with different outcomes for translation.
      i. 3' decay would result in no ribosomes terminating and send all transcripts into the m_0^* class .
      ii. Endonucleolytic decay due to NGD or NMD would potentially allow ribosomes downstream to terminate but not those upstream (Urquidi-Camacho 2020, Merchante 2017).
      iii. Thus, depending on the ditribution of mRNAs in the capped class, and the site of the endonucleolytic decay a transcript in m_i would end up in m_j, where $j < i$.

2. A better understanding of how different factors affect a gene's mRNA stability and, in turn, protein expression, relevant to a wide range of applied molecular biology (e.g. the design of efficient heterologous genes expression and mRNA vaccines. (Cheng 2023 viruses, Boo and Kim 2020). Current debate is focused on the contributions of the protective effects of ribosome association vs. ribosome stalling to mRNA transcript stability.

   (a) Our model currently cannot distinguish between the protective effects of translation or codon effects.

   (b) In the current implementation of the model we did not directly explore the protective effects of ribosomal loading which could be modeled by weighting the marking rate μ by $(1 - i/imax)$, or having separate marking rates for m_0 and the other polysomal classes.

   (c) Our model does not consider codon specific effects such as pausing sites, difficult to fold regions of a protein or codon optimality (Wu and Bazzini 2023).

   (d) Splitting each polysome class into two regions would approximate a ribosome flow model of only two regions, bounded by the pausing site.

   (e) The protective effects of translation could increase per ribosome, but eventually at high loads could trigger ribosome associated decay pathways through ribosomal collisions. This would require analysis on an individual transcript basis.

5. Supplemental figures

Figure 12: Marking rate determines the gene abundance for most transcripts. The parameter λ was calculated by multiplying the RNA-seq RPKM ($\sum_0^{i_{\max}} \tilde{m}_i$) values of Weinberg 2016 with the marking rates μ obtained from Presnyak 2015 as is described in 5. As μ increses, λ increases linearly for many genes, suggesting μ accounts for the accumulation of transcripts. Some genes significanly deviate from this pattern, indicating additional production from λ. Spearmans ρ is presented with a pvalue $< 10^{-16}$.