# Resolving the mRNA Life Cycle across Cell Compartments

Inaugural-Dissertation

## Jason Marcel Müller

A Thesis presented for the Degree of

Doctor of Natural Sciences

Institute for Medical Statistics and Computational Biology

University of Cologne

Germany

December 2023

Whenever we proceed from
the known into the unknown
we may hope to understand,
but we may have to learn at
the same time a new
meaning of the word
"understanding".

*Werner Karl Heisenberg*

*Physics and Philosophy:*

*The Revolution in Modern*

*Science (1958)*

# Abstract

In a eukaryotic cell, the life cycle of an mRNA consists of several processing steps, including its synthesis, export from the nucleus into the cytosol, and, after fulfilling its function as a protein template, cytosolic degradation. Eventually, the transcript can also be degraded in the nucleus upon aberrant processing. Yet, the dynamics of these processes that determine how mRNA flows through the cell are still not fully understood. Understanding these dynamics is crucial as impairments in mRNA metabolism have been associated with severe human diseases.

Recent advances in biotechnology allow for the investigation of eukaryotic mRNA metabolism by combining metabolic labeling sequencing techniques with subcellular fractionation protocols. In such experiments, newly synthesized mRNA is labeled with nucleoside analogs that allow for differentiation of pre-existing transcripts. However, the analysis of this data provides many computational challenges.

In this dissertation, I develop a model to quantify essential parameters of eukaryotic mRNA metabolism, from which nuclear and cytosolic RNA half-lives can be derived. The model uses metabolic labeling sequencing data from HeLa-S3 cells that were separated into nuclear and cytosolic compartments by subcellular fractionation prior to sequencing. The presented method is demonstrated to overcome critical biases of metabolic labeling sequencing data. At the same time, it outperforms the corresponding methodology of the widely used tool GRAND-SLAM for analyzing such data.

The model reveals that nuclear export is slow, while cytosolic degradation is fast. Consequently, an average mRNA spends most of its life in the nucleus, not the cytosol. I show that mRNA metabolism is consistent along the cell cycle, and its quantities can be used to pinpoint potential regulators of metabolic subprocesses such as nuclear export.

In summary, the presented model resolves the life cycle of an eukaryotic mRNA across cell compartments with high accuracy and, therefore, will be a valuable tool for RNA research.

# Zusammenfassung

In einer eukaryotischen Zelle besteht der Lebenszyklus einer mRNA aus mehreren Schritten, einschließlich ihrer Synthese, ihres Exports aus dem Zellkern in das Zytosol und, nachdem sie ihre Funktion als Proteinvorlage erfüllt hat, ihres Abbaus im Zytosol. Die Dynamiken dieser Prozesse, die bestimmen, wie die mRNA durch die Zelle fließt, müssen jedoch noch erforscht werden. Das Verständnis dieser Dynamiken ist von entscheidender Bedeutung, da Beeinträchtigungen des mRNA-Metabolismus mit schweren menschlichen Krankheiten assoziiert sind.

Fortschritte in der Biotechnologie ermöglichen die Untersuchung des eukaryotischen mRNA-Metabolismus durch die Kombination von Sequenzierungstechniken mit metabolischen Markern und subzellulären Fraktionierungsprotokollen. In solchen Experimenten werden neu synthetisierte mRNAs mit Nukleosidanaloga markiert, um sie von bereits vorhandenen Transkripten unterscheiden zu können. Die Analyse dieser Daten ist jedoch mit zahlreichen bioinformatischen Herausforderungen verbunden.

In dieser Dissertation entwickle ich ein Modell zur Quantifizierung wesentlicher Parameter des eukaryotischen mRNA-Stoffwechsels, woraus sich Halbwertszeiten von nukleärer und zytosolischer RNA ableiten lassen. Das Modell verwendet metabolische Markierungs-Sequenzierungsdaten von HeLa-S3-Zellen, die vor der Sequenzierung durch subzelluläre Fraktionierung in nukleäre und zytosolische Kompartimente aufgeteilt wurden. Es wird gezeigt, dass die vorgestellte Methode kritische Probleme metabolischer Markierungs-Sequenzierungsdaten überwindet. Gleichzeitig übertrifft sie die entsprechende Methodik des weit verbreiteten Programms GRAND-SLAM für die Analyse solcher Daten.

Das Modell zeigt, dass der Export von RNA aus dem Zellkern langsam und der Abbau im Zytosol schnell erfolgt. Folglich verbringt eine mRNA die meiste Zeit ihres Lebens im Zellkern und nicht im Zytosol. Ich zeige, dass der mRNA-Metabolismus über den Zellzyklus weitgehend konstant ist. Darüber hinaus können seine Schätzungen verwendet werden, um potenzielle Regulatoren von Subprozessen wie dem Export von nukleärer RNA zu identifizieren.

Zusammenfassend lässt sich sagen, dass das vorgestellte Modell den Lebenszyklus einer eukaryotischen mRNA über Zellkompartimente hinweg mit hoher Genauigkeit auflöst und daher ein wertvolles Programm für die RNA-Forschung sein wird.

# List of Abbreviations

**ARE** AU-rich element

**CTD** C-terminal domain

**DNA** Deoxyribonucleic acid

**EM** Expectation-maximization

**eRNAs** Enhancer RNAs

**FISH** Fluorescence *in situ* hybridization

**h** Hour

**IAA** Iodoacetamide

**ISH** *In situ* hybridization

**IQR** Inter-quartile range

**lncRNAs** Long non-coding RNAs

**MCMC** Markov chain Monte Carlo

**min** Minutes

**miRNAs** Micro RNAs

**mM** Millimolar

**mRNA** Messenger RNA

**mRNP** Messenger ribonucleoprotein

**m6A** N6-Methyladenosine

**NEBD** Nuclear envelope breakdown

**NPC** Nuclear pore complex

**nt** Nucleotide

**PCR** Polymerase chain reaction

**PIC** Pre-initiation complex

**Pol** Polymerase

**RBP** RNA-binding protein

**RNA** Ribonucleic acid

**rRNA** Ribosomal RNA

**RT** Reverse transcription

**SLAM-seq** Thiol(SH)-linked alkylation for the metabolic sequencing of RNA

**SNPs** Single-nucleotide polymorphisms

**snRNA** Small nuclear RNA

**s** Second

**TREX** Transcription-export

**tRNAs** Transfer RNAs

**UTR** Untranslated region

**WCE** Whole-cell extract

**4sU** 4-thiouridine

$\mu$**M** Micromolar

$\mu$**m** Micrometre

# Contents

# 1 Introduction

## 1.1 Ribonucleic Acid: an Informant, Messenger, and Catalyst of Life

Ribonucleic acid (RNA) plays a crucial role in every living organism. According to Francis Crick's central dogma of biology, RNA acts as a messenger that transmits the genetic information of deoxyribonucleic acid (DNA) for the production of proteins [1]. For this reason, an RNA that codes for a protein is called messenger RNA (mRNA). Today, it is recognized that the function of RNA goes beyond this role as a messenger, as it has two essential properties of both DNA and proteins. Like DNA, RNA carries genetic information, and like proteins, it can catalyze biochemical reactions [2]–[4]. These properties have sparked discussions about RNA being the pioneer molecule of life on Earth before the evolution of DNA and proteins [5]–[8]. This hypothesis was termed "The RNA World" by Walter Gilbert in 1986 [9].

Since Francis Crick publicized his hypothesis about the central dogma of biology, various RNA types have been discovered. Such types include, but are not limited to, long non-coding RNAs (lncRNAs), enhancer RNAs (eRNAs), micro RNAs (miRNAs), and transfer RNAs (tRNAs). Each of these transcript classes possesses its characteristics and regulatory functions. For instance, lncRNAs are non-protein-coding transcripts typically longer than 200 nucleotides (nt), are usually poly-adenylated, and are involved in essential gene regulation processes such as X-chromosome inactivation [10]–[12]. In contrast, eRNAs are transcribed from enhancers, which are cis-acting elements that promote the transcription of a particular gene. These transcripts are usually not poly-adenylated and do not code for proteins but can regulate, for instance, chromatin accessibility [12], [13]. These examples show that the regulatory nature of RNAs is diverse. Yet, the exact role of many RNA classes remains to be elucidated. Among all these classes, mRNAs have been the most comprehensively researched as they are fundamental for protein synthesis [14].

# 1.2  The Five Steps of the mRNA Life Cycle

In a eukaryotic cell, mRNA undergoes five critical steps in its life cycle. The first step involves its synthesis in the nucleus, which is coordinated by RNA polymerase (Pol) II. This process of transcription starts with an initiation phase, where a set of general transcription factors assemble with RNA Pol II at the promoter region of a respective gene, forming the pre-initiation complex (PIC) [15]–[17]. This initiation phase is succeeded by an elongation phase, whereby the polymerase moves along and unwinds the double-stranded DNA in the 3'- to 5'-direction to transcribe a nascent RNA transcript nucleotide by nucleotide in the 5'- to 3'-direction. During this elongation phase, RNA Pol II does not synthesize continuously but pauses at several sites of a gene sequence [18]. After the transcript has been fully synthesized, transcription terminates, releasing RNA Pol II from the DNA. This process yields an immature pre-mRNA molecule, which requires further processing.

During the second stage of the mRNA life cycle, pre-mRNA undergoes additional processing in the nucleus to produce a mature transcript suitable for protein synthesis. The transcript's 5'-end is capped while the 3'-end is cleaved and poly-adenylated to prepare it for export from the nucleus into the cytosol and increase stability [19], [20]. The spliceosome facilitates alternative splicing of the pre-mRNA, removing introns from the RNA sequence [21]. RNA modifications like N6-Methyladenosine (m6A) introduced by writer proteins and RNA editing by ADAR enzymes can modify mRNA in such a way that these modifications affect a transcript's stability, translation efficiency, and even the protein sequence [22]–[27].

If correctly processed, an mRNA proceeds to the third stage of its life cycle, where it undergoes nuclear export. This process is regulated by various RNA-binding proteins (RBPs), which associate with the transcript, forming a messenger ribonucleoprotein complex (mRNP) [28]. The protein composition of these mRNPs decides the transcript's fate [28], [29]. Several pathways control the nuclear export of mRNA types, with the NXF1/TAP-pathway and CRM1/XPO1-pathway forming the main two export routes in human cells (Figure 1) [30].
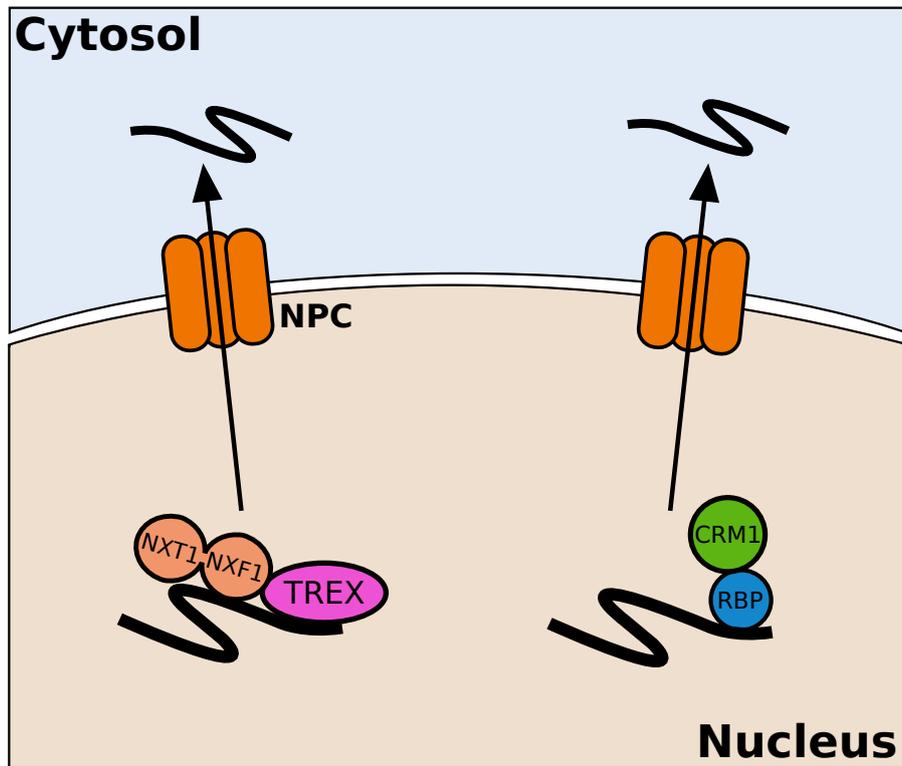
**Figure 1: Main Nuclear Export Routes of mRNA in Eukaryotes.** NXF1 forms a heterodimer with NXT1, which is subsequently recruited by the Transcription-Export (TREX) complex. By direct binding, NXF1 promotes the nuclear export of most mRNAs through the nuclear pore complex (NPC). CRM1, in contrast, cannot bind directly to RNA but has been shown to interact with several RBPs to translocate mRNA through the NPC [30].

In the NXF1/TAP-pathway, NXF1 forms a heterodimer with NXT1, which is recruited by the Transcription-Export (TREX) complex during splicing (Figure 1). Through its direct binding to RNA, NXF1 promotes the nuclear export of most mRNAs through the nuclear pore complex (NPC). In contrast, the export of small transcripts like snRNAs, rRNAs, but also a subset of mRNAs is mediated by CRM1. This pathway is independent of the NXF1/TAP-pathway since it does not involve the TREX complex or the NXF1-NXT1 heterodimer. On the contrary, CRM1 cannot bind directly to RNA but has been shown to interact with various RBPs to translocate mRNA through the NPC along a Ran-GTP gradient [30].

However, if the mRNA is not correctly processed in the nucleus, it is targeted for nuclear degradation. In this fourth stage of the mRNA life cycle, these transcripts could lead to the synthesis of aberrant proteins, so they must not be exported from the nucleus. The

main driver of nuclear degradation is the exosome, a multiprotein complex that cleaves mRNA from its 3'- to 5'-end [31].

Cytosolic degradation is the fifth and last step in the mRNA life cycle. After the transcript has fulfilled its function as a protein template for translation, it gets degraded in the cytosol. Usually, cytoplasmic mRNA decay starts with the shortening of the poly-A tail, triggering downstream pathways. The most prominent ones are 5' to 3' degradation mediated by XRN1 or 3' to 5' degradation mediated by the exosome [32]–[34].

Although the life cycle of mRNA can be subdivided into the previously mentioned steps, there are overlaps between these processes. For instance, pre-mRNA processing is coupled with transcription via RNA Pol II [19], [35], [36]. Its C-terminal domain (CTD) interacts with factors that mediate 5'-capping, poly-adenylation, and splicing [37]–[40]. Furthermore, the TREX complex is recruited during transcription to facilitate nuclear export [41]. These examples suggest a high interconnectedness of the metabolic processes, increasing the complexity of mRNA metabolism. However, there is a high need to unravel the underlying factors and mechanisms involved as impairments in mRNA metabolism have been associated with severe human diseases like motor neuron disease [42], [43].

## 1.3 Quantification of mRNA Metabolic Processes

Various methods for quantifying mRNA metabolism have been developed over the last decades. For instance, radioactive labeling and RNA extraction were combined to monitor RNA turnover from cell lysates in the early 1960s [44]. However, this approach had limited sensitivity and specificity because they could not distinguish between gene-specific transcripts and their subcellular localization. The development of *in situ* hybridization (ISH) and its successor fluorescence *in situ* hybridization (FISH) allowed the targeted monitoring of RNA expression and localization [45]–[48]. In the mid-1980s, the introduction of polymerase chain reaction (PCR) by Kary Mullis [49] further revolutionized the quantification of mRNA metabolism: reverse transcription (RT-)PCR has been used to study mRNA synthesis, splicing, and degradation rates [50]–[54]. Although these tools provide valuable insights into the processes of RNA metabolism, they also have limitations and

drawbacks. On the one hand, working with radioactive materials is hazardous for the experimentalist. On the other hand, FISH and PCR methods cannot be easily scaled up to study an entire genome or transcriptome. In addition, these methods are often combined with protocols that use inhibitors such as actinomycin D and thiolutin to block transcription [51], [54], [55]. Consequently, these settings can severely affect cellular processes and lead to biased results [56].

A less invasive method of studying RNA dynamics is the metabolic labeling of RNA with non-radioactive nucleoside analogs such as 4-thiouridine (4sU). This analog is incorporated into newly synthesized transcripts during transcription [57], [58]. The resulting newly synthesized fraction can be purified from the total RNA fraction using biotinylation and streptavidin-based pulldown, allowing simultaneous monitoring of RNA synthesis and degradation rates using microarrays [58]–[61]. With the advances in high-throughput sequencing, microarray-based analyses have been replaced by RNA-seq, providing even higher specificity and sensitivity in the quantification of RNA metabolism [62]–[66]. However, the accuracy of these measurements highly depends on the biochemical separation efficiency of newly synthesized and pre-existing RNA, which was shown to be limited [67].

In 2017, Herzog *et al.* published a method called thiol(SH)-linked alkylation for the metabolic sequencing of RNA (SLAM-seq) that bypasses the limited efficiency of biochemical separation [68]. According to the SLAM-seq protocol, cells are treated with 4sU for a certain period before the RNA is isolated and treated with iodoacetamide (IAA). This compound performs a thiol-alkylation by attaching a carboxyamidomethyl group to 4sU. When preparing the cDNA library, reverse transcriptase interprets this modified base as a cytidine, not a uridine. As a result, the sequence alignments of the newly synthesized RNAs contain cytosines instead of thymines. The relative amount of newly synthesized transcripts can be quantified from the sequencing data alone and does not require physical separation from the pre-existing fraction. Therefore, SLAM-seq provides even more accurate measurements than biotinylation-based methods.

Bioinformatics methods and mathematical models are required to analyze the ever-increasing data obtained by metabolic labeling methods. One prominent model that has been widely applied to estimate RNA decay rates is the exponential decay model [58], [60], [64], which is an ordinary differential equation of the form:

$$\frac{dR(t)}{dt} = -\lambda R \tag{1}$$

Here, $R$ describes the quantity of a particular transcript and $\lambda$ a positive degradation rate constant. The exponential decay model requires the biological system to be in steady state, so metabolic rates do not change over time. Under this assumption, this differential equation has the solution

$$R(t) = R_0 \cdot e^{-\lambda t} \tag{2}$$

where $R(t)$ describes the quantity of a transcript $R$ at time point $t$ and $R_0$ the initial quantity at time point 0 (the quantity in steady state). An intuitive way to interpret an RNA metabolic rate is its inverse, the half-life. The RNA half-life describes the time it takes for the amount of a transcript to decrease to half its steady-state value. The RNA half-life can be calculated as:

$$t_{1/2} = \frac{ln(2)}{\lambda} \tag{3}$$

In addition, steady-state models have been used to quantify RNA synthesis rates. They use the principle that the expression level of a transcript is determined by a balance between RNA synthesis and decay at equilibrium [58], [60], [64]. However, the steady-state assumption does not hold if the biological system is subjected to intrinsic or external stress. In such a case, metabolic rates are not constant but can change over time. For this reason, some studies have proposed differential equations in which the metabolic rates themselves are time-dependent [53], [62], [69]. However, as model complexity increases, so does the risk of overfitting the model. For this reason, steady-state models are still commonly used.

Sequencing data generated by methods such as SLAM-seq also has some limitations. T>C conversions in a read induced by 4sU labeling would be detected as a mismatch during sequence alignment to a genome. Yet, 4sU-labeling-induced T>C conversions could be mimicked by single-nucleotide polymorphisms (SNPs), sequencing errors, or RNA editing [26], [27]. In addition, the rate of 4sU incorporation into newly synthesized transcripts

is not 100%, but in the range of 2% [68]. Due to this low labeling efficiency, some reads from new transcripts will not exhibit a single T>C nucleotide conversion and will appear as pre-existing. Therefore, a naive count of labeled reads would lead to an underestimation of the relative abundance of newly synthesized transcripts. The labeling efficiency cannot be substantially increased by prolonging labeling periods or the concentration of 4sU [58], [68], as high concentrations of 4sU can trigger cellular stress [70], which could bias transcript dynamics. Further, stress responses violate the steady-state assumption of the exponential decay model, so it cannot be used for downstream analysis. Therefore, these issues must be addressed during data processing and modeling.

To manage the mapping problem of SLAM-seq data, Herzog *et al.* (2017) used a custom pre-processing pipeline called SlamDunk in their original publication [68], [71]. This tool can perform T>C conversion-aware alignments that account for annotated SNPs. One year later, Jürges *et al.* (2018) published GRAND-SLAM, a tool that corrects SLAM-seq data for the low 4sU incorporation rates to derive RNA degradation kinetics [72]. To this end, they proposed a Bayesian framework to quantify newly synthesized and pre-existing RNA proportions on a genome-wide scale. These proportions can be used to calculate $\frac{New}{Total}$ RNA ratios. Working with these relative quantities is more robust than working with absolute quantities since no external normalization with control samples is required. GRAND-SLAM then fits an exponential decay model to these $\frac{New}{Total}$ RNA ratios over time and, subsequently, computes RNA half-lives.

Although GRAND-SLAM has been used in several studies with SLAM-seq experiments [73]–[78], this tool has some potential drawbacks. GRAND-SLAM applies a binomial mixture model that accounts for T>C conversions occurring in unlabeled RNA due to sequencing errors to distinguish between newly synthesized and pre-existing transcripts. However, it does not account for the possibility that T>C conversions might also be missed in labeled RNA if the sequencing machine erroneously calls a base other than cytosine (C>X error, where X is either A, T, or G). Furthermore, their binomial mixture model assumes that the 4sU labeling efficiency is known. Therefore, they estimate this labeling efficiency using an expectation-maximization (EM) algorithm [79]. Instead of using all sequencing reads for this step, GRAND-SLAM selects only a subset of reads with sufficient T>C conversion to be distinguishable from T>C sequencing error. These pre-selected reads could

be inflated by T>C conversions arising from RNA editing and SNPs. If fewer than 10,000 reads can be selected for estimating the labeling efficiency, the authors describe in their publication that GRAND-SLAM stops the computation [72]. Another limitation of GRAND-SLAM is that it was initially designed for sequencing data from whole-cell extracts. Although RNA synthesis, degradation, and splicing rates can be modeled from such data, quantifying nuclear export rates and attributing RNA degradation to the nucleus and cytosol requires data from subcellular fractionation experiments [73], [76], [80].

Recently, we have published a probabilistic modeling framework that overcomes many biases of metabolic labeling sequencing data [80]. This pipeline corrects the data for potential RNA editing sites and SNPs more conservatively than SlamDunk and GRAND-SLAM. Further, it models the 4sU labeling efficiency using an EM algorithm while accounting for both the T>C and C>X sequencing errors. Additionally, this EM algorithm returns $\frac{New}{Total}$ RNA ratio estimates for each measured gene. We calculated these ratios from subcellular fractionation SLAM-seq data of the nucleus and cytosol from HeLa-S3 cells to fit an ordinary differential equation system of the following form:

$$\frac{dN(t)}{dt} = \mu - (\nu + \tau)N \tag{4}$$

$$\frac{dC(t)}{dt} = \tau N - \lambda C \tag{5}$$

Here, $N(t)$ and $C(t)$ denote the time courses of nuclear and cytosolic RNA abundances, respectively. These quantities depend on 4 parameters: the synthesis rate $\mu$, the nuclear degradation rate $\nu$, the nuclear export rate $\tau$, and the cytosolic degradation rate $\lambda$ (Figure 2).
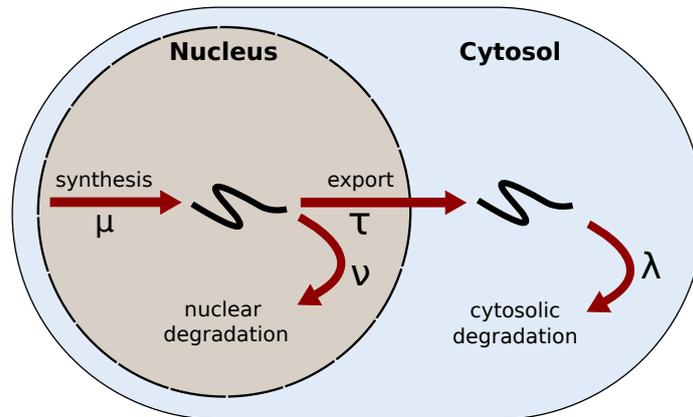
**Figure 2: Metabolic Rates along the mRNA Life Cycle.** Shown is the typical life path of a transcript. In the nucleus, an mRNA molecule is synthesized with a specific synthesis rate $\mu$. Then, an mRNA can be either exported from the nucleus with rate $\tau$ or degraded in the nucleus with rate $\nu$. After the transcript has fulfilled its purpose as a template for protein synthesis, it gets degraded in the cytosol with rate $\lambda$.

Our two-compartment model quantifies all these 4 metabolic processes simultaneously. We found that nuclear degradation of polyadenylated mRNA is negligible, and its nuclear export is slow, whereas cytosolic mRNA degradation is comparatively fast. With a nuclear RNA half-life of roughly 300 minutes and a cytosolic RNA half-life of approximately 60 minutes, an average mRNA spends most of its life in the nucleus rather than the cytosol. We have also shown that applying a simple exponential decay model to individual subcellular fractions does not capture the interdependence between RNA metabolic processes as our two-compartment model does. When quantifying RNA metabolism, it is crucial to consider the compartmentalization of the cell. Otherwise, the stability of the transcript is mainly attributed to mechanisms of cytosolic RNA degradation rather than RNA export from the nucleus.

As RNA metabolic labeling and subcellular fractionation experiments have only recently been combined [73], [76], [80], our modeling framework and results might be subject to certain biases. We determined rates of mRNA metabolism in bulk sequencing data of HeLa-S3 cells. Since these cells were not synchronized regarding the cell cycle, the metabolic RNA rates reflect average measures across all cell cycle stages. However, since the nuclear envelope breaks down and, therefore, the former nuclear and cytosolic fractions could mix during Mitosis, the mRNA export rates could be distorted by this process. Also, the modeling framework should be tested with another SLAM-seq time series.

In this thesis, I developed a methodology that expands on our previously published modeling framework [80]. To this end, I put particular focus on the EM algorithm used to estimate the 4sU labeling efficiency and, subsequently, the $\frac{New}{Total}$ RNA ratios. I implemented a new EM algorithm that estimates a global labeling efficiency from all measured transcripts. Initially, our system estimated transcript-specific labeling efficiencies and then calculated the median of the resulting distribution to determine the final $\frac{New}{Total}$ RNA ratios. Simulations showed that this new EM algorithm estimates a wide range of 4sU incorporation rates with high precision. It also outperforms the EM algorithm used by GRAND-SLAM. In addition, I applied the presented model to new SLAM-seq time series data from HeLa-S3 cells explicitly generated for this study. These data have the unique feature that the cells were synchronized regarding the cell cycle. The time series were generated from cells in G1 and Mitosis. This data allowed for investigating RNA metabolic rates for differences between bulk and cell cycle-specific measurements. I demonstrated that applying the presented model returns highly reliable rates of RNA metabolism during the cell cycle.

# 2 Data and Methods

## 2.1 Data

### 2.1.1 SLAM-seq Time-Series from synchronized HeLa-S3 Cells

Two SLAM-seq time-series datasets were generated by Elisabeth Altendorfer and Susanne Freier from the Mayer lab at the MPI for Molecular Genetics (Berlin, Germany). For this purpose, HeLa-S3 were synchronized at early S-phase using a 2.5 mM double thymidine block. Cells were released into fresh and prewarmed media supplemented with 24 µM 2-deoxycytidine (Sigma) for 6 hours (late G2) prior to labeling with 500 µM 4sU for 0, 15, 45, 75, 120 and 180 minutes. At each time point, cells were fractionated to obtain nuclear and cytosolic cell fractions. Similarly, a time series in the G1 cell cycle phase (12-hour release time after double thymidine block) was measured at 0, 15, 45, 75, 145, 180, 240, and 300 minutes. Prior to phenol extraction of RNA, 1,9 µl 1:100 diluted ERCC Exfold RNA spike-ins per 1 million cells were added to each sample. 4sU-RNA was alkylated based on Herzog *et al.* (2017) [68] and poly-A enriched SLAM-seq libraries were prepared using 3'-QuantSeq REV protocol (Lexogen) according to manufacturers instructions [81]. The final libraries were sequenced using an Illumina NovaSeq6000 with 150 nt read length in paired-end mode. Table 1 lists the crucial experimental parameters.

**Table 1:** Experimental Parameters of the SLAM-seq cell cycle data.

| | |
|---|---|
| **Cell line** | HeLa-S3 |
| **Time-series for Mitosis** | 0, 15, 45, 75, 120, 180 minutes |
| **Time-series for G1** | 0, 15, 45, 75, 145, 180, 240, 300 minutes |
| **Subcellular Fractions** | Nuclear and cytosolic fraction |
| **Sequencing Protocol** | 3'UTR Quant-seq REV with poly-A enrichment (Lexogen) |
| **Read Length** | 150 nucleotides per read |
| **Sequencing Mode** | Paired-end (NovaSeq6000) |
| **Spike-ins** | ERCC ExFold RNA Spike-in Mix (Thermo Fisher Scientific) |

### 2.1.2   SLAM-seq Time-Series from unsynchronized HeLa-S3 Cells

The second dataset used in this work is the SLAM-seq time-series data we recently published on bioRxiv [80]. In brief, HeLa-S3 were incubated with 500 µM 4sU for 0, 15, 30, 45, 60, 120 and 180 minutes immediately followed by sub-cellular fractionation into nuclear and cytosolic compartments as proposed by Nojima *et al.* (2016) [82]. Library preparation was performed according to the 3' QuantSeq REV protocol (Lexogen) [81]. Each library was sequenced in single-end mode with 100 nt read length on an Illumina HiSeq 2000 machine. Sequencing reads were mapped to human genome hg19 using SlamDunk [71] and assigned to 3'UTRs using annotation downloaded from UCSC Table Browser [83]. Before counting T>C conversions, the data was corrected for putative SNPs and RNA editing sites (see Data and Methods 2.2.7). After conversion count, $\frac{New}{Total}$ RNA ratios were estimated using an EM algorithm. Subsequently, the two-compartment model was fit to these ratios to obtain metabolic rate estimates.

Since hg19 and GRCh38 differ in annotation, RNA half-life estimates from these unsynchronized and the synchronized HeLa-S3 cells (Data and Methods 2.1.1) were compared on the gene level. To that end, the 3'UTRs from both annotation files were mapped to their corresponding gene name.

### 2.1.3   RNA Half-life Estimates from Whole-Cell Extracts

For comparison of RNA metabolic rates obtained by the two-compartment model with estimates from the literature, RNA half-lives published by Schueler *et al.* (2014) [63] were used. This dataset contains estimates from MCF7 and HEK293 cells. The cells were incubated with 4sU, followed by separation of the newly synthesized RNA from the total RNA fraction using biotinylation. No subcellular fractionation of the respective cells was performed, but whole-cell extracts (WCE) were used. The authors describe that RNA half-lives were calculated according to Schwanhäusser *et al.* (2011) [66]. These estimates were compared with the RNA half-lives obtained from the SLAM-seq cell cycle data (Data and Methods 2.1.1) based on the respective gene names.

### 2.1.4   RNA-binding Protein Profiles from eCLIP Experiments

eCLIP [84] data comprising binding profiles of 120 RNA-binding proteins obtained from K562 cells were downloaded from ENCODE [85]. The profiles are stored in BED files containing the genomic coordinates of cross-linked binding sites. Only BED files for GRCh38 were considered to match the genome assembly used for this work. Furthermore, only binding sites with ($-log10(p$-value$) \leq 5$ and $log2($fold-enrichment$) \geq 3$) were kept for the analysis to control for false positives. GENCODE annotation [86] was used to map the binding sites to gene identifiers.

A 3'UTR was defined as bound by a certain RBP if this 3'UTR can be assigned to a single gene, which in turn contains an exonic binding site for this RBP. Subsequently, this procedure yielded lists of both RBP-bound and RBP-unbound transcripts. A Wilcoxon rank-sum test was applied to test for significant differences in RNA half-lives between these two groups. This procedure was done for each subcellular fraction separately. For robustness, only RNA half-life estimates that fulfilled stringent reliability criteria were considered (see Data and Methods 2.2.12).

### 2.1.5   Annotation of AU-rich Element Transcripts

AU-rich elements (ARE) in the 3'UTRs of specific transcripts affect RNA stability [87]. To compare RNA half-lives between ARE and non-ARE transcripts, annotation of known ARE-containing RNAs was downloaded from the ARED-Plus database [88]. The respective gene names of these ARE RNAs were used to assign RNA half-lives from the SLAM-seq cell cycle data. A Wilcoxon rank-sum test assessed significant differences between ARE and non-ARE transcripts. Only RNA half-life estimates that fulfilled stringent reliability criteria were considered (see Data and Methods 2.2.12).

### 2.1.6   Annotation of Protein Complexes

The annotation of protein complexes was downloaded from the CORUM database [89]. The sums of the cell cycle-specific nuclear and cytosolic RNA half-life estimates (see Data and Methods 2.1.1) were assigned to the corresponding protein subunits in each complex using the gene names. Only RNA half-life estimates that fulfilled stringent reliability cri-

teria were considered (see Data and methods 2.2.12). The median RNA half-life of the protein complexes was calculated. Only protein complexes with at least 5 subunits were considered to obtain a more stable median estimate of the RNA half-life. The log2 fold changes in RNA half-life between the individual subunits and the whole complex were then calculated for each protein complex.

## 2.2   **Methods**

### 2.2.1   **An EM Algorithm to determine the Proportions of newly synthesized Transcripts**

SLAM-seq enables the differentiation between newly synthesized and pre-existing transcripts. From these two fractions, $\frac{New}{Total}$ RNA ratios can be calculated for each measured transcript. A time series of these $\frac{New}{Total}$ RNA ratios can be used to derive parameters of RNA metabolism, such as synthesis and degradation rates. However, calculating these ratios is a difficult task as the 4sU incorporation rate (respectively, 'labeling efficiency') was shown to be in the range of 2% [68] but is highly dependent on 4sU labeling time and concentration [58], [68]. This low labeling efficiency leads to the problem that many newly synthesized transcripts do not show any T>C conversions, for which they will be misclassified as pre-existing. As a consequence, the actual proportion of newly synthesized transcripts is underestimated, which in turn leads to biased estimates of RNA metabolism. SNPs, RNA editing sites, and T>C (false-positive) sequencing errors in high-throughput sequencing data that mimic T>C conversions induced by 4sU labeling exacerbate this problem. Furthermore, a 4sU-labeling induced T>C conversion can also be missed due to C>X (false-negative) sequencing errors. Despite all these possible sources of error, the challenge is to derive an accurate estimator for the labeling efficiency, which determines the proportion of newly synthesized transcripts.

Analogous to Jürges *et al.* (2018) [72], the proportion of newly synthesized transcripts is modeled with a binomial mixture model of the following form:

$$P(o_j^r; U_j^r, \rho^r, \epsilon_+, a) = (1 - \rho^r) \cdot Bin(o_j^r; U_j^r, \epsilon_+) + \rho^r \cdot Bin(o_j^r; U_j^r, a) \qquad (6)$$

Here, $o_j^r$ denotes the observed T>C conversions and $U_j^r$ the number of potential conversion sites in a read $j$ mapped to a certain 3'UTR $r$. Further, $\rho^r$ is the $\frac{New}{Total}$ RNA ratio of this region $r$, and $\epsilon_+$ the T>C false-positive sequencing error. In contrast to Jürges *et al.* (2018) [72], $a$ does not directly reflect the labeling efficiency (in the following denoted as $\ell$) but a corrected labeling efficiency which accounts for the T>C false-positive sequencing error $\epsilon_+$ and the C>X false-negative sequencing error $\epsilon_-$. This conversion rate $a$ is calculated as:

$$a = 1 - (\ell\epsilon_- + (1 - \ell)(1 - \epsilon_+)) = \ell(1 - \epsilon_- - \epsilon_+) + \epsilon_+ \tag{7}$$

Equation (7) of $a$ is derived from the tree diagramm shown in Figure 3.
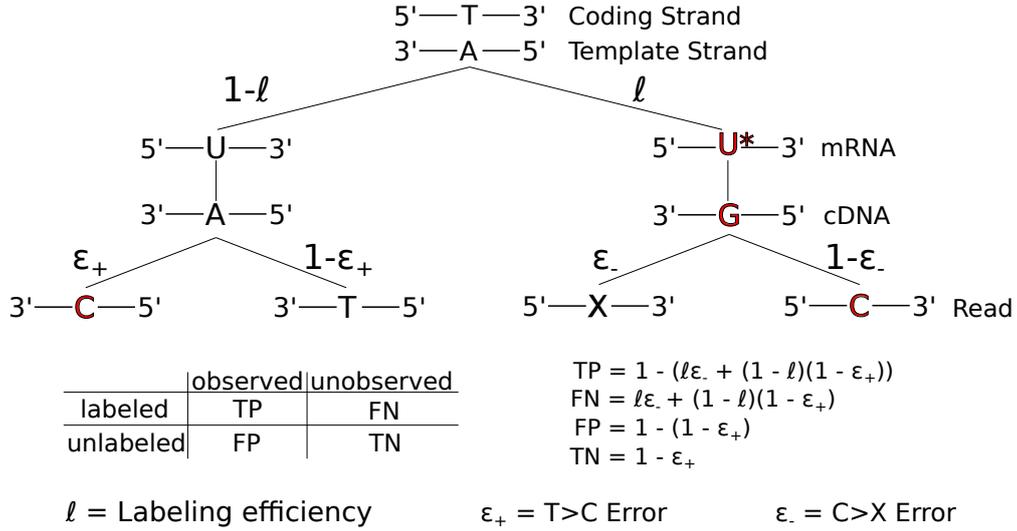


**Figure 3: Tree Diagramm of a T>C Conversion Event.** Represents the probability paths of observing a T>C conversion in 4sU labeling experiments. Here, $\ell$ denotes the 4sU labeling efficiency, whereas $\epsilon_+$ and $\epsilon_-$ denote the false-positive and false-negative sequencing errors, respectively. T>C conversion event paths are highlighted by red letters.

Each branch in the tree diagram in Figure 3 depicts the probability paths of observing T>C conversions in 4sU labeling experiments. Suitable estimates for $\epsilon_+$ and $\epsilon_-$ can be directly derived from control experiments of a SLAM-seq time-series, given that SNP and RNA editing sites were masked beforehand (see Data and Methods 2.2.7-2.2.8). Then, the labeling efficiency $\ell$ can be estimated with the following EM algorithm.

Let $R$ denote the set of all 3'UTRs that are expressed across a SLAM-seq time-series experiment. For $r \in R$, let $J^r$ be the number of reads mapped to region $r$. Given a read $j = 1, ..., J^r$, let $U_j^r$ be its number of potential conversion sites and let $o_j^r \in 0, 1, ..., U_j^r$ the number of positions at which we observe T>C conversions. Let $h_j^r$ be a hidden variable that denotes whether the read $j$ is derived from a pre-existing RNA ($h_j^r = 0$), or a newly synthesized RNA ($h_j^r = 1$). Let $\rho^r \in [0, 1]$ be the proportion of newly synthesized RNAs in the RNA population belonging to region $r$. Let $\ell \in [0, 1]$ be the labeling efficiency, i.e., the probability by which a 4sU is incorporated into a newly synthesized RNA. Let $\epsilon_+ \in [0, 1]$

be the probability of a T>C sequencing error (false-positive) and $\epsilon_- \in [0, 1]$ be the C>X sequencing error (false-negative) probabilities. Denote the unknown parameters by $\Theta = (\ell, \rho^r; r \in R)$. Then,

$$P(o_j^r, h_j^r; \Theta) = P(h_j^r; \rho^r) \cdot P(o_j^r \mid h_j^r; U_j^r, \ell, \epsilon_+, \epsilon_-) \tag{8}$$

In the above expression, $P(h_j^r; \rho^r) = Bernoulli(h_j^r; p = \rho^r)$ and

$$P(o_j^r \mid h_j^r; U_j^r, \ell, \epsilon_+, \epsilon_-) = \begin{cases} Bin(o_j^r; n = U_j^r, p = \epsilon_+) & \text{if } h_j^r = 0 \\ Bin(o_j^r; n = U_j^r, p = a) & \text{if } h_j^r = 1 \end{cases} \tag{9}$$

where $a = 1 - (\ell\epsilon_- + (1 - \ell)(1 - \epsilon_+))$ as denoted in Equation (7). Taking logs of Equation (8) results in:

$$\log P(o_j^r, h_j^r; \Theta) = \begin{cases} \log(1 - \rho^r) + \log \binom{U_j^r}{o_j^r} + o_j^r \log \epsilon_+ + (U_j^r - o_j^r) \log(1 - \epsilon_+) & \text{if } h_j^r = 0 \\ \log \rho^r + \log \binom{U_j^r}{o_j^r} + o_j^r \log a + (U_j^r - o_j^r) \log(1 - a) & \text{if } h_j^r = 1 \end{cases} \tag{10}$$

**E-step.**

Let $H = (h_j^r; j = 1, ..., J^r)$ and $H_{-1} = (h_j^r; j = 2, ..., J^r)$. Given a parameter set $\Theta' = (\rho^{r\prime}, \ell')$, the function $Q(\Theta, \Theta')$ has to be optimized with respect to $\Theta = (\rho^r, \ell)$:

$$Q(\Theta; \Theta') := \mathbb{E}_{P(H|O;\Theta')} \log P(O, H; \Theta)$$

$$= \sum_H P(H \mid O; \Theta') \log P(O, H; \Theta)$$

$$= \sum_{r \in \mathcal{R}} \sum_{H^r_{-1}} \sum_{h^r_1 \in \{0,1\}} \left[ P(H^r_{-1} \mid O^r_{-1}; \Theta') \cdot P(h^r_1 \mid o^r_1; \Theta') \right]$$

$$\cdot \left[ \log P(O^r_{-1}, H^r_{-1}; \Theta) + \log P(o^r_1, h^r_1; \Theta) \right]$$

$$= \sum_{r \in \mathcal{R}} \sum_{h^r_1 \in \{0,1\}} P(h^r_1 \mid o^r_1; \Theta') \log P(o^r_1, h^r_1; \Theta)$$

$$+ \sum_{r \in \mathcal{R}} \sum_{H^r_{-1} \in \{0,1\}^{J-1}} P(H^r_{-1} \mid O^r_{-1}; \Theta') \log P(O^r_{-1}, H^r_{-1}; \Theta)$$

$$\overset{\text{induction}}{=} \sum_{r \in \mathcal{R}} \sum_{j=1}^{J^r} \sum_{h^r_j \in \{0,1\}} P(h^r_j \mid o^r_j; \Theta') \log P(o^r_j, h^r_j; \Theta) \tag{11}$$

Let

$$c^r_{j,h^r_j} := P(h^r_j \mid o^r_j; \Theta') = \frac{P(o^r_j \mid h^r_j; \ell') P(h^r_j; \rho'^r)}{\sum_{h^r_j \in \{0,1\}} P(o^r_j \mid h^r_j; \ell') P(h^r_j; \rho'^r)} \quad , j = 1, ..., J^r \tag{12}$$

where

$$C^r_0 = \sum_{j=1}^{J^r} c^r_{j,0}$$
$$C^r_1 = \sum_{j=1}^{J^r} c^r_{j,1} \tag{13}$$

Further, if $U^r_j \geq 30$ (for determination of this threshold see Data and Methods 2.2.2 and Results 3.1.1):

$$A = \sum_{r \in \mathcal{R}} \sum_j c^r_{j,1} o^r_j$$
$$B = \sum_{r \in \mathcal{R}} \sum_j c^r_{j,1} (U^r_j - o^r_j) \tag{14}$$

Then, Equation (11) simplifies to:

$$Q(\Theta; \Theta') = \sum_{r \in \mathcal{R}} \sum_j c_{j,0}^r \left[ \log(1 - \rho^r) + \log \binom{U_j^r}{o_j^r} + o_j^r \log \epsilon_+ + (U_j^r - o_j^r) \log(1 - \epsilon_+) \right]$$

$$+ \sum_{r \in \mathcal{R}} \sum_j c_{j,1}^r \left[ \log \rho^r + \log \binom{U_j^r}{o_j^r} + o_j^r \log a + (U_j^r - o_j^r) \log(1 - a) \right]$$

$$= \sum_{r \in \mathcal{R}} C_0^r \log(1 - \rho^r) + \sum_{r \in \mathcal{R}} C_1^r \log \rho^r + A \log a + B \log(1 - a) + \text{const}$$

$$(15)$$

**M-step.**

Taking the partial derivative of $Q$ in Equation (11) with respect to $\rho^r$ and equating this expression to zero yields:

$$0 = \frac{\partial Q(\Theta; \Theta')}{\partial \rho^r} = -\frac{C_0^r}{1 - \rho^r} + \frac{C_1^r}{\rho^r}$$

$$\rho^r = \frac{C_1^r}{C_0^r + C_1^r} = \frac{C_1^r}{J^r}$$

$$(16)$$

Recall that $a = \ell(1 - \epsilon_- - \epsilon_+) + \epsilon_+$ (Equation (7)), and take the partial derivative of $Q$ in Equation (11) with respect to $\ell$. Equating the resulting expression to zero yields:

$$0 = \frac{\partial Q(\Theta; \Theta')}{\partial \ell} = \frac{\partial Q(\Theta; \Theta')}{\partial a} \cdot \frac{\partial a}{\partial \ell}$$

$$= \left( \frac{A}{a} - \frac{B}{1 - a} \right) \cdot (1 - \epsilon_- - \epsilon_+)$$

$$(17)$$

Solving Equation (17) for $a$ (assuming $1 - \epsilon_- - \epsilon_+ \neq 0$) and then solving for $\ell$ yields:

$$a = \frac{A}{A + B}$$

$$\ell = \frac{a - \epsilon_+}{1 - \epsilon_- - \epsilon_+} = \frac{\left( \frac{A}{A+B} - \epsilon_+ \right)}{1 - \epsilon_- - \epsilon_+}$$

$$(18)$$

To further stabilize the labeling efficiency estimation procedure, only regions that fulfill a certain coverage threshold can be used to consider a 3'UTR $r$ for the update step in Equation (14). This threshold was set to $J^r \geq 200$ reads per region for the SLAM-seq cell cycle

data 2.1.1, due to the high coverage per 3'UTR (Figure 6). The EM was run until conver-gence to obtain a final estimate of the labeling efficiency $\ell$. Subsequently, $\ell$ was plugged in Equation (7) to calculate the corrected labeling efficiency $a$. In turn, $a$ was plugged into the binomial mixture model given by Equation (6). With this binomial mixture model, the $\frac{New}{Total}$ RNA ratio $\rho^r$ was determined using maximum likelihood (L-BFGS-B method [90]) for each 3'UTR $r$.

### 2.2.2  Simulation Details for Determining a Read Pre-selection Threshold based on potential Labeling Sites

Given a certain labeling efficiency $\ell \in \{1\%, 2\%, 3\%, 4\%, 5\%\}$ and a certain number of 4sU incorporation sites (also referred to as "potential labeling sites") $U \in \{17, 30\}$ in a read, the number of expected T>C conversions $o \in 0, 1, ...U$ is modeled by a binomial distribu-tion:

$$Bin(k = o, n = U, p = \ell) \tag{19}$$

Based on the resulting distributions given by Equation (19), it was found that the labeling efficiency could be estimated more reliably if the reads contain at least a certain number of potential labeling sites (see Figures 11-12; Results 3.1.1). The following simulation was performed to determine this threshold.

1,000 3'UTRs were simulated with a certain coverage $T \in \{100, 200\}$. Furthermore, a $\frac{New}{Total}$ RNA ratio $\rho \in \{1\%, 5\%, 10\%, 20\%, 50\%, 90\%\}$ was defined to simulate the frac-tion of newly synthesized reads $N$ by sampling from a Binomial distribution of form $Bin(n = T; p = \rho)$. Correspondingly, the fraction of pre-existing reads was defined as $O = T - N$ minus the number of newly synthesized reads. Next, a potential number of labeling sites $U \in \{10, 20, 30, 40, 50\}$ was defined. Then, for each read of newly synthe-sized fraction $N$, the number of T>C conversions was sampled from a binomial distribu-tion $Bin(n = U; p = a)$. Here, $a = 1 - (\ell\epsilon_- + (1 - \ell)(1 - \epsilon_+))$ (Equation (7)) is a corrected labeling efficiency which accounts for the actual labeling efficiency $\ell \in \{1\%, 2\%, 3\%\}$, the T>C sequencing error $\epsilon_+ = 0.1\%$ and the C>X sequencing error $\epsilon_- = 0.1\%$. Correspond-ingly, putative T>C conversions due to $\epsilon_+$ were simulated for each read of the pre-existing

fraction $O$ by a binomial distribution $Bin(n = U; p = \epsilon_+)$. Subsequently, the number of T>C conversions was used by the presented EM algorithm (Data and Methods 2.2.1) to estimate the labeling efficiency $\ell$ and the corresponding $\frac{New}{Total}$ RNA ratio $\rho$.

### 2.2.3   Simulation to assess the Impact of Deviations in Labeling Efficiency on $\frac{New}{Total}$ RNA Ratios

Deviations in the labeling efficiency estimate from the actual value could lead to biased estimates of the $\frac{New}{Total}$ RNA ratios. The impact of such deviations was assessed with the following simulation.

1,000 3'UTRs were simulated from actual transcripts annotated for GRCh38. To that end, a BED file containing the genomic coordinates of GRCh38 transcripts was downloaded from the UCSC Table Browser [83]. For each 3'UTR $r$, a coverage ranging between 0-1,000 reads $T_r$ was sampled from a beta-binomial distribution $BetaBin(n = 1000, \alpha = 10, \beta = 90)$ with $\alpha = 10$ and $\beta = 90$ to obtain a mean of 100 reads (Figure 4).
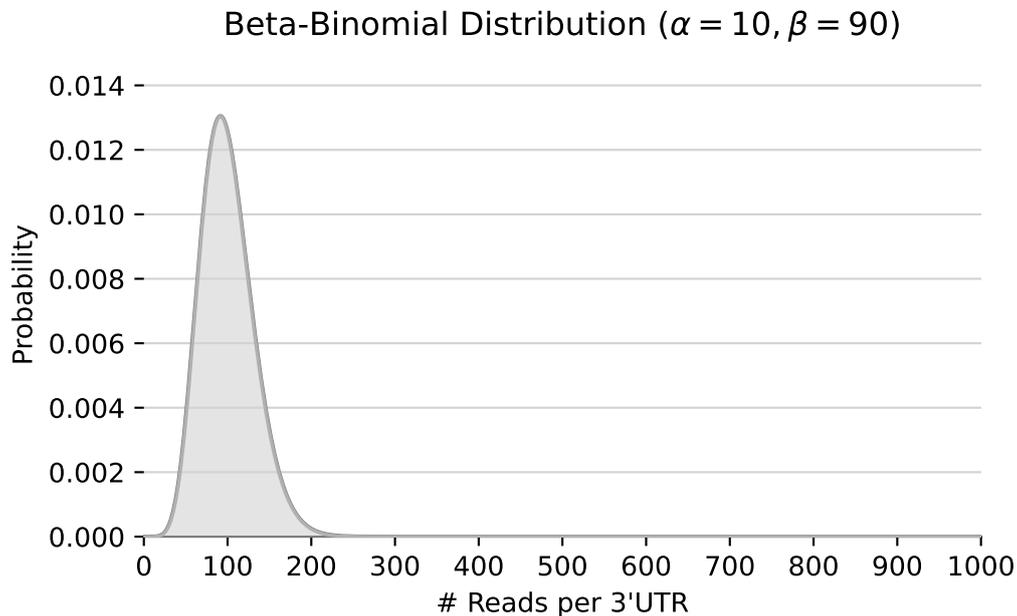


**Figure 4: Coverage Distribution used for Simulation of Annotated GRCh38 Transcripts.** The distribution was modeled by $BetaBin(n = 1000, \alpha = 10, \beta = 90)$. For visualization purposes, this beta-binomial distribution appears as a continuous distribution. However, it is a discrete distribution.

Having determined a specific coverage $T_r$ from the beta-binomial distribution (Figure 4), this number of reads was sampled from the annotated 3'UTR regions. The length of each read was set to 100 nt. Since the length of each region $r$ was 250 nt, while the length of a read was only 100 nt, each read was sampled randomly from within the 3'UTR. This procedure introduced some variability in the number of potential labeling sites between reads from the same region $r$. Next, a $\frac{New}{Total}$ RNA ratio $\rho \in \{1\%, 5\%, 25\%, 50\%, 75\%, 95\%, 99\%\}$ was used as the event probability for the binomial distribution to sample the fraction of newly synthesized $N_r$ from a binomial distribution according to $Bin(n = T_r, p = \rho)$. Correspondingly, the pre-existing fraction $O_r$ was defined as $O_r = T_r - N_r$. Then, for each read $j$ in both fractions $N_r$ and $O_r$, the number of potential labeling sites $U_j$ was determined. For reads of the new fraction $N_r$, T>C conversions were simulated using a binomial distribution $Bin(n = U_j, p = a)$ with a fixed labeling efficiency $\ell$ of 2%, where $a = 1 - (\ell \epsilon_- + (1 - \ell)(1 - \epsilon_+))$ is a corrected labeling efficiency which accounts for the T>C sequencing error $\epsilon_+ = 0.1\%$ and the C>X sequencing error $\epsilon_- = 0.1\%$ (see Equation (7)). In contrast, T>C conversions in reads from the pre-existing fraction $O_r$ were sampled from $Bin(n = U_j, p = \epsilon_+)$. Instead of using the EM algorithm to estimate this 'true' labeling efficiency $\ell$ of 2%, the labeling efficiency was manually varied between 1-3% with a step size of 0.01%. For each of these labeling efficiencies, the corrected labeling efficiency $a$ was calculated according to Equation (7). $a$ was then used to compute the $\frac{New}{Total}$ RNA ratio estimates $\hat{\rho}_r$ using the binomial mixture model as defined by Equation (6). Subsequently, the final $\frac{New}{Total}$ RNA ratio estimate $\hat{\rho}$ was obtained by calculating the median of the 3'UTR-specific estimates $\hat{\rho}_r$.

### 2.2.4   Simulation for Performance Assessment of the new EM Algorithm

The simulation described in Data and Methods 2.2.2 determined a suitable number of potential labeling sites per read. This number can be used as a threshold to pre-select reads, which are used to estimate the labeling efficiency. However, this threshold was identified using only a fixed number of potential labeling sites per read. In actual sequencing data, the number of potential labeling sites varies between reads as it depends on the genomic sequence of a respective transcript and the read length. For validation of the pre-selection threshold and comparison of the presented EM (Data and Methods 2.2.1)

with the corresponding EM algorithm of GRAND-SLAM [72], a similar simulation as described in Data and Methods 2.2.3 was performed. Again, sequencing data was simulated from actual genomic sequences of GRCh38. The details are described in the following.

15,000 3'UTRs from annotated transcripts of GRCh38 were simulated. Genomic coordinates of GRCh38 transcripts were extracted from a BED file downloaded from the UCSC Table Browser [83]. For each transcript, the last 250 nucleotides were defined as the 3'UTR. For each 3'UTR $r$, a coverage ranging between 0-1,000 reads $T_r$ was sampled from a beta-binomial distribution $BetaBin(n = 1000, \alpha = 10, \beta = 90)$ with $\alpha = 10$ and $\beta = 90$ to obtain a mean of 100 reads (Figure 4). According to the coverage $T_r$, this number of reads was sampled from the 3'UTR regions, each with a length of 100 nt. The reads were sampled randomly from within the corresponding 3'UTR, so the number of potential labeling sites per read could vary. Then, a $\frac{New}{Total}$ RNA ratio $\rho_r$ between 1-99% was randomly sampled from a uniform distribution and assigned to each 3'UTR. $\rho_r$ was then used to sample the newly synthesized fraction of reads $N_r$ from a binomial distribution according to $Bin(n = T_r, p = \rho_r)$. Correspondingly, the pre-existing fraction $O_r$ was defined as $O_r = T_r - N_r$. Then, for each read $j$ in both fractions $N_r$ and $O_r$, the number of potential labeling sites $U_j$ was determined. For reads of the new fraction $N_r$, T>C conversions were simulated using a binomial distribution $Bin(n = U_j, p = a)$, where $a = 1 - (\ell\epsilon_- + (1 - \ell)(1 - \epsilon_+))$ is a corrected labeling efficiency which accounts for the actual labeling efficiency $\ell \in \{0.1\%, 0.5\%, 1\%, 2\%, 3\%, 4\%, 5\%\}$, the T>C sequencing error $\epsilon_+ = 0.1\%$ and the C>X sequencing error $\epsilon_- = 0.1\%$ (Equation (7)). In contrast, T>C conversions of pre-existing reads were sampled from $Bin(n = U_j, p = \epsilon_+)$. The objective of the presented EM algorithm (Data and Methods 2.2.1) was to estimate $\ell$ from this data.

The performance of the presented EM algorithm was compared with the corresponding EM algorithm of GRAND-SLAM [72] using the same simulation framework. GRAND-SLAM is not open-source, so the EM algorithm was implemented as described in the corresponding publication. For ease of readability, this implementation is simply referred to as GRAND-SLAM in the following. As no start value was described for the labeling efficiency, it was naively set to $\ell = 100\%$ in the first iteration of GRAND-SLAM's EM algorithm. Conveniently, the number of potential labeling sites and the observed T>C conversions are sufficient statistics for both methods.

### 2.2.5   Quality Control and Mapping of Sequencing Data

Individual FASTQ files of the SLAM-seq cell cycle dataset (Data and Methods 2.1.1) were quality-checked with fastqc [91]. Sequencing adapters were trimmed using cutadapt [92]. Trimmed reads were mapped to human genome GRCh38 with NextGenMap [93] in paired-end mode. The number of mapped reads ranged from roughly 60-120 million reads (Figure 5).
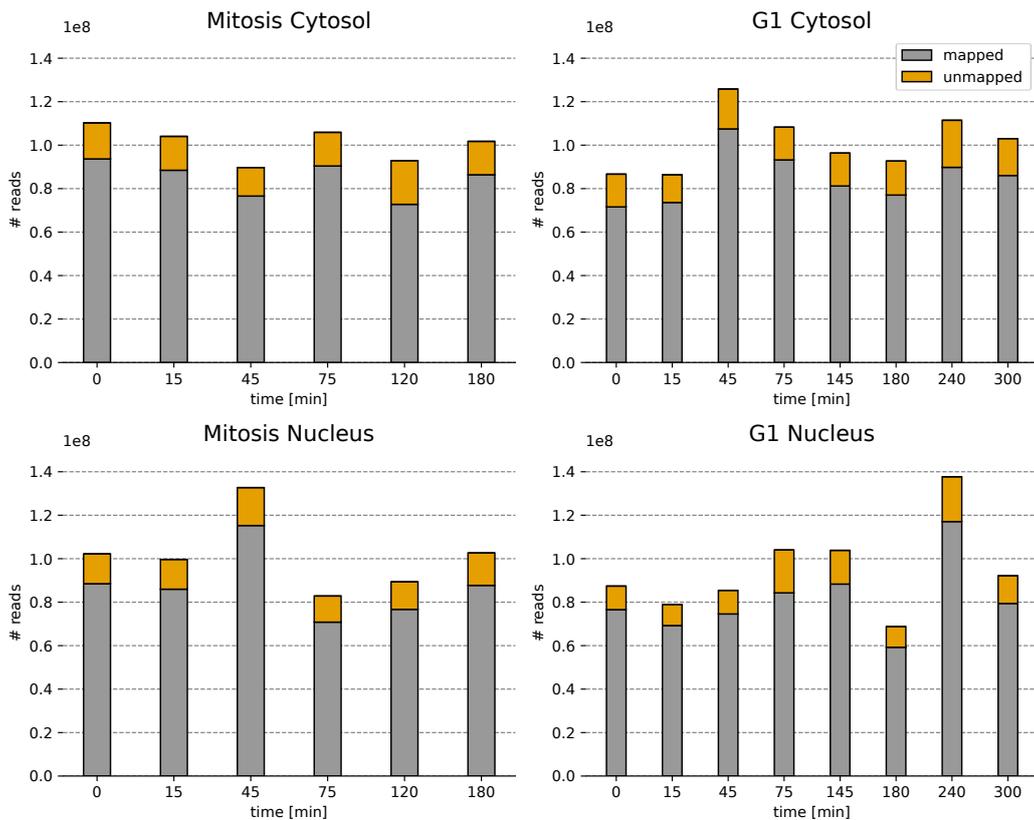


**Figure 5: Mapping Statistics of the SLAM-seq Cell Cycle Data.** Trimmed reads were mapped to human genome GRCh38 with NextGenMap [93] in paired-end mode. The number of mapped (gray) and unmapped (orange) reads are shown for each time point of the nuclear and cytosolic fractions from the cell cycle phases G1 and Mitosis.

### 2.2.6   Read Filtering and Assignment to 3'UTRs

Several quality filter criteria were applied to the mapped reads of the SLAM-seq cell cycle data (Data and Methods 2.1.1). The first read of each aligned read pair was kept for further analysis. The reason for this procedure was the observation that mates of the

same read pair often overlapped. Therefore, T>C conversions detected in reads aligned to the sense strand (A>G conversions on the antisense strand, respectively) would be erroneously counted twice, as the read pair comes only from one single transcript. Similar to the filtering procedure of SlamDunk [71], reads were filtered out if their mapping quality was lower than 2 and the sequencing identity was lower than 95%. The reads which passed these quality filters were stored in BAM files for downstream processing.

The 3'UTR annotations for GRCh38 were obtained from the UCSC Table Browser [83]. Overlapping 3'UTRs on the same strand were merged into one region. A read was assigned to a 3'UTR if its alignment overlapped with the region with at least one nucleotide position. The strandedness of the 3'QuantSeq REV protocol (Lexogen) [81] was considered during the read assignment. That means reads in the forward direction were assigned to annotated regions on the antisense (-) strand, whereas reads in the reverse direction were assigned to regions on the sense (+) strand. The reads could be assigned to 15,779 3'UTRs in G1 and 16,517 3'UTRs in Mitosis expressed across the respective time series (Figure 6).
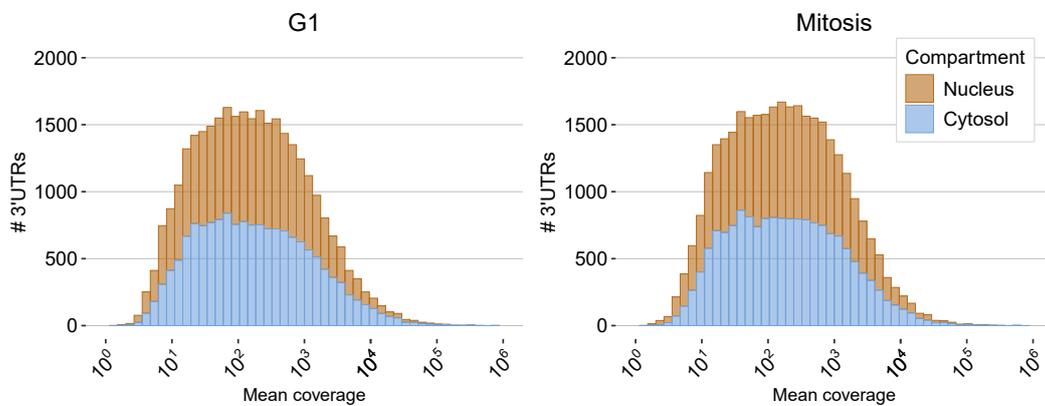


**Figure 6: Mean Coverage per 3'UTR of the Cell Cycle data.** The histograms show the mean coverages over the respective time series of G1 (left) and Mitosis (right), which were calculated for both the nuclear (orange) and the cytosolic (blue) fractions.

### 2.2.7   Correction for SNPs and RNA Editing Sites

Genomic SNPs can mimic T>C conversions and bias 4sU labeling-induced conversion counts. As these counts are essential for accurately estimating $\frac{New}{Total}$ RNA ratios, it is necessary to mask these SNP positions. Therefore, a VCF file containing SNP coordinates in

the human genome GRCh38 was downloaded from NCBI dbSNP [94]. Subsequently, these positions were excluded from further analyses.

Nonetheless, cell line-specific SNPs that could accumulate during cultivation and RNA editing sites could be present in the data. Therefore, position-wise mismatch rates were calculated in the respective control samples (0 min time points) to account for these genomic positions. If the mismatch rate of a genomic position was found to be higher than 5%, it was considered a potential SNP or RNA editing site. Such positions were collected from all individual control samples in one list, which was used to mask the respective genomic positions in each SLAM-seq cell cycle dataset sample. Subsequently, T>C and (respectively, A>G conversions on the antisense strands) were counted in each sample and monitored over time (Figure 7).
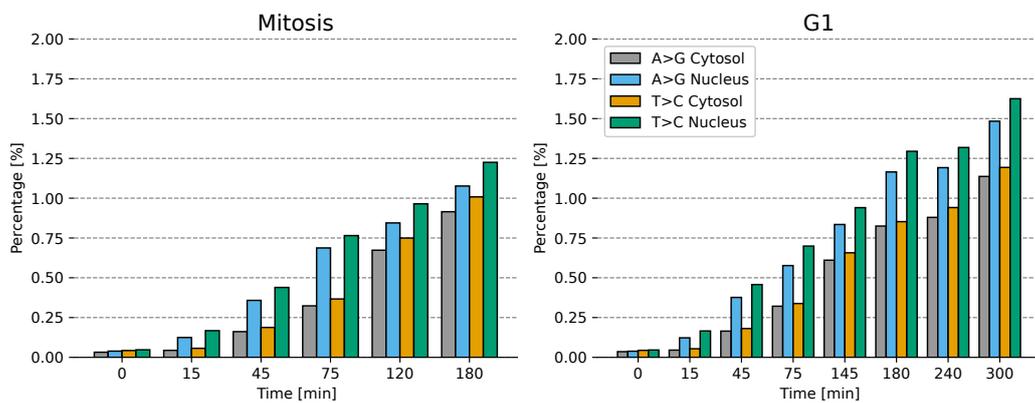


Figure 7: **Conversion Statistics of the SLAM-seq Cell Cycle Data.** Barplots depict the increase of T>C (A>G conversions on antisense strand, respectively) over time for each compartment and cell cycle phase. Conversions were counted after the correction of sequencing data for SNPs and RNA editing sites.

### 2.2.8   Estimation of Sequencing Errors

Sequencing errors were determined after correction for SNPs and RNA editing sites (Data and Methods 2.2.7) from control samples (0 min time points). Two distinct types of sequencing errors can be detected in 4sU-based sequencing experiments. The first type is the false-negative sequencing error $\epsilon_-$, describing C>X (sense strand) and G>X (antisense strand) conversions, which mask 4sU labeling induced T>C (sense strand) and A>G (antisense strand) conversions, respectively. In contrast, the false-positive sequencing error $\epsilon_+$

describes T>C conversions (respectively A>G conversions on the antisense strand), which occur in sequencing data even without 4sU labeling.

$\epsilon_+$ was directly inferred as the T>C mismatch rate (A>G mismatch rate, respectively). In contrast, $\epsilon_-$ was calculated as the average of all C>X mismatch rates (G>X mismatch rates, respectively). Finally, the strand-specific sequencing errors were averaged to obtain one error rate $\epsilon_+$ and $\epsilon_-$ for each respective time series. Table 2 depicts the sequencing errors measured in the SLAM-seq Cell Cycle Data (Data and Methods 2.1.1).

**Table 2:** Sequencing Errors in SLAM-seq Cell Cycle Data.

| Control | T>C Error $\epsilon_+$ | C>X Error $\epsilon_-$ |
|---|---|---|
| Mitosis Cytosol | 0.033% | 0.723% |
| Mitosis Nucleus | 0.036% | 0.708% |
| G1 Cytosol | 0.036% | 0.736% |
| G1 Nucleus | 0.034% | 0.692% |

### 2.2.9 A Two-compartment Model of RNA Metabolism

Using the $\frac{New}{Total}$ RNA ratios, the life cycle of a mature (polyadenylated) transcript can be modeled by four metabolic parameters [80]. Initially, the RNA is synthesized in the nucleus at a rate $\mu$. Subsequently, the mature transcript is either exported to the cytosolic compartment with rate $\tau$ or degraded in the nucleus at a nuclear degradation rate $\nu$. Transcripts that were exported from the nucleus are degraded in the cytosol with rate $\lambda$ irrespective of their function in this compartment. The two-compartment model assumes that $\mu, \tau, \nu$ and $\lambda$ are constant over time, so the RNA metabolism is in steady state. In the following, the details of the two-compartment model are described.

Given a 3'UTR, $N = N(t)$ denotes the time course of nuclear RNA abundances averaged over all cells in a bulk sample. Correspondingly, $C = C(t)$ represents the time course of the cytosolic RNA abundances. Here, $N$ and $C$ can be treated as continuous variables as they are high numbers. Subsequently, their time course can be modeled by an ordinary differential equation system involving $\mu, \tau, \nu$ and $\lambda$:

$$\frac{dN}{dt} = \mu - (\nu + \tau)N \tag{20}$$

$$\frac{dC}{dt} = \tau N - \lambda C \tag{21}$$

For any choice of initial conditions, the system has a closed-form solution at $t = 0$ but converges to its steady state in every case. This steady state is given by:

$$N_\infty = \frac{\mu}{\nu + \tau} \tag{22}$$

and

$$C_\infty = \frac{\mu}{\nu + \tau} \cdot \frac{\tau}{\lambda} \tag{23}$$

With these steady-state levels, the time courses of the pre-existing RNA fraction can be modeled by:

$$N_{old}(t) = N_\infty \cdot e^{-(\nu+\tau)t} \tag{24}$$

and

$$C_{old}(t) = C_\infty \cdot \left( \frac{\lambda e^{-(\nu+\tau)t} - (\nu + \tau)e^{-\lambda t}}{\lambda - (\nu + \tau)} \right) \tag{25}$$

Correspondingly, the time courses of the newly synthesized RNA fraction can be obtained by subtracting the pre-existing RNA levels from the steady states levels:

$$N_{new}(t) = N_\infty \cdot \left( 1 - e^{-(\nu+\tau)t} \right) \tag{26}$$

and

$$C_{new}(t) = C_\infty \cdot \left( 1 - \frac{\lambda e^{-(\nu+\tau)t} - (\nu + \tau)e^{-\lambda t}}{\lambda - (\nu + \tau)} \right) \tag{27}$$

SLAM-seq makes it possible to distinguish between newly synthesized and existing RNA without separating these two fractions physically but by sequencing alone. This proce-

dure provides accurate estimates of the relative abundance of RNA $\frac{N^{new}(t)}{N_\infty}$ in the nucleus, respectively $\frac{C^{new}(t)}{C_\infty}$ in the cytosol. In steady state, these ratios can be calculated with Equations (22) and (26) for the nucleus and with Equations (23) and (27) for the cytosol:

$$\frac{N^{new}(t)}{N_\infty} = 1 - e^{-(\nu+\tau)t} \tag{28}$$

$$\frac{C^{new}(t)}{C_\infty} = 1 - \frac{\lambda e^{-(\nu+\tau)t} - (\nu+\tau)e^{-\lambda t}}{\lambda - (\nu+\tau)} \tag{29}$$

As defined by Equation (28), $\nu + \tau$ is treated as one quantity called nuclear vanishing rate. Further, the cytosolic ratio in Equation (29) only depends on this nuclear vanishing rate and the cytosolic degradation rate $\lambda$. Therefore, these parameters $\nu + \tau$ and $\lambda$ can be estimated from the measured nuclear and cytosolic $\frac{New}{Total}$ RNA ratios by fitting Equations (28) and (29) to these quantities (see Data and Methods 2.2.11).

## 2.2.10    Variance-stabilizing Transformation of $\frac{New}{Total}$ RNA Ratios

As in our previous work [80], the $\frac{New}{Total}$ RNA ratios $\rho$ were variance-stabilized by applying an arcsin transformation [95]. Assuming that the ratios $\rho$ with a corresponding total read count $J$ follow a binomial distribution given by $Bin(n = J, p = \rho)$, the distribution of the arcsin-transformed $\frac{New}{Total}$ RNA ratios $arcsin\left(\sqrt{\rho}\right)$ approximates to a normal distribution with variance $\sigma^2 = \frac{1}{4 \cdot J}$. The variance $\sigma^2$ depends only on the number of total reads $J$ so that it can be accounted for during parameter fitting.

## 2.2.11    Fitting the Two-compartment Model to derive RNA Dynamics

Using the arcsin-transformed nuclear and cytosolic $\frac{New}{Total}$ RNA ratios (Data and Methods 2.2.10), Equation (28) and (29) of the two-compartment model (Data and Methods 2.2.9) are fitted with the following procedure.

First, fix a 3'UTR. Let $q(t)$ be the arcsin-transformed $\frac{New}{Total}$ RNA ratio of the 3'UTR in measurement in either the nuclear or cytosolic compartment at time point $t$. Let $\hat{q}(t, \Theta)$ be the true, target $\frac{New}{Total}$ RNA ratio defined by the parameter set $\Theta = (\tau + \nu, \lambda)$. The

probability of observing the arcsin-transformed $\frac{New}{Total}$ RNA ratio $q(t)$ given the parameter set $\Theta$ can be formulated as:

$$
\begin{aligned}
P(q(t) \mid \Theta) &= \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\left(\frac{q(t)-\hat{q}(t,\Theta)}{\sigma}\right)^2} \\
&= \frac{\sqrt{4J}}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot 4J(q(t)-\hat{q}(t,\Theta))^2}
\end{aligned}
\tag{30}
$$

Subsequently, taking the logarithm results in:

$$
logP(q(t) \mid \Theta) = log\left(\frac{\sqrt{4J}}{\sqrt{2\pi}}\right) - 2J(q(t) - \hat{q}(t,\Theta))^2
\tag{31}
$$

Then, a suitable parameter set $\Theta$ is obtained by maximizing the sum of the log-likelihoods over all time points $T$ of a SLAM-seq time series $\sum_t^T logP(q(t) \mid \Theta)$ with respect to $\Theta$. In this case, $log\left(\frac{\sqrt{4J}}{\sqrt{2\pi}}\right)$ can be omitted from Equation (31) as it does not depend on $\Theta$.

The Markov chain Monte Carlo method (MCMC) Metropolis-Hastings [96] was used as an algorithm to sample the parameter set $\Theta = (\tau + \nu, \lambda)$. Specifically, $\tau + \nu$ and $\lambda$ were sampled jointly in each two-dimensional MCMC run. To reduce burn-in time and to prevent the Markov chain from getting stuck in regions of low probability in the parameter space, a Differential Evolution algorithm (global optimization) [97] was used to find a suitable seed for the Markov chain. Next, the Markov chain sampled 50,000 parameter combinations of $\tau + \nu$ and $\lambda$ (Figure 8).
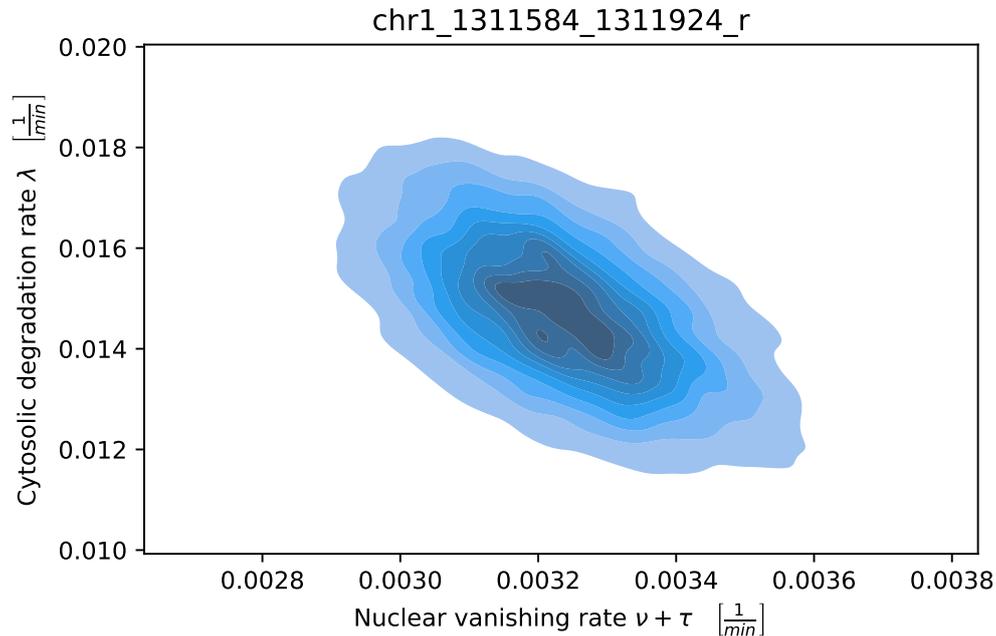
**Figure 8: Examplary MCMC Density Plot of a Parameter Fit.** The MCMC-chain sampled 50,000 parameter combinations of the joint parameter set $\Theta = (\nu + \tau, \lambda)$ for an exemplary 3'UTR with identifier "chr1_1311584_1311924_r". Darker areas indicate higher density regions than lighter areas. The elliptic form reflects broad searches through the probability space of the parameter combinations.

Median values for $\tau + \nu$ and $\lambda$ were calculated from these samples. Conveniently, credibility intervals were calculated as the 95%-percentiles of the MCMC samples. The median values were used as a seed for a Nelder-Mead algorithm (local optimization) [98] to obtain the final parameter set $\Theta$.

### 2.2.12 Applying Reliability Criteria to Rate Estimates of RNA Metabolism

As in our previous work [80], several reliability criteria were applied to the rate estimates obtained by the two-compartment model (Data and Methods 2.2.11). To that end, a parameter estimate was classified as reliable if: 1) Its corresponding 3'UTR was found to be covered with at least 30 reads on average across the time series. 2) The difference between the parameter estimate and the upper and lower credibility bound by the parameter estimate is $\leq 0.3$. 3) The expression level of the 3'UTR region is constant across the time series so that the steady-state assumption of the model holds. To that end, the ex-

pression data was normalized by calculating the average read count of the inter-quartile range (25-75%) of the read distribution of a respective time point. Then, the read count of each respective 3'UTR was divided by this average to obtain relative expression levels. A linear regression model was fitted to the expressions of each 3'UTR over time. If the slope of the regression line did not increase by more than $0.0025 \frac{1}{min}$ or decrease by less than $-0.0025 \frac{1}{min}$, the region was considered reliable. 4) The coefficient of determination $R^2$ of the model fit is $\geq 0.4$. The reliability criteria were applied to the estimates of each compartment and cell cycle phase. Additionally, cytosolic parameters were only considered reliable if the corresponding nuclear estimate was also reliable. This criterion is reasoned by the fact that the RNA content in the cytosol depends on the export from the nucleus.

### 2.2.13   Estimating $\frac{\text{Cyt}}{\text{Nuc}}$ RNA Abundances using ERCC Spike-ins

$\frac{Cyt}{Nuc}$ RNA abundance ratios were determined using ERCC Exfold RNA Spike-In Mixes (Thermo Fisher Scientific). First, a robust average read count was calculated for each sample as described in Müller *et al.* (2023) [80]. To that end, the bottom and top quartile of 3'UTRs were excluded concerning the total read count. Subsequently, the read counts of the remaining 50% of 3'UTRs were averaged. Only 3'UTRs covered with at least 1 read were considered for this procedure. This normalization strategy prevented the average expression value from being biased by regions with highly variable transcript expression levels. Then, each robust average expression value was divided by the total count of the ERCC spike-ins within a respective sample. This division yielded spike-in normalized read count averages. Lastly, $\frac{Cyt}{Nuc}$ RNA abundance ratios were calculated by dividing these count averages of the cytosolic fraction by the corresponding nuclear measurements for each time point.

### 2.2.14 Simulation Details for the Assessment of cytosolic RNA Half-life Variability

In this work and our previous study [80], the variability of cytosolic RNA half-lives was notably higher than the corresponding estimates for the nucleus. This discrepancy might result from technical variability, which is investigated with this simulation.

Consider a SLAM-seq time series $t = \{15, 30, 45, 75, 120, 180\}$ in minutes and a certain coverage $T \in \{30, 100, 250, 1000\}$. Further, assume some parameter values for the synthesis rate $\mu = 1\frac{1}{min}$ and the nuclear degradation rate $\nu = 1 \cdot 10^{-10}\frac{1}{min}$. Reasonable estimates for the export rate can be obtained by taking the prior information that the average RNA spends roughly 300 min in the nucleus. Then, we can obtain the rate by $\tau = \frac{ln(2)}{300min} = 0.0023\frac{1}{min}$. Analogously, we can obtain a cytosolic degradation rate for a cytosolic half-life of 10 min by calculating $\lambda = \frac{ln(2)}{10min} = 0.069\frac{1}{min}$. $\lambda$ values were calculated this way for 10-100 min with a step size of 10min. For each $\lambda$ and each coverage $T$, we can obtain a total fraction of nuclear RNA $N_\infty$ using Equation (22). Subsequently, the old nuclear fractions can be modeled by $N_{old} = N_\infty \cdot e^{-(\nu+\tau)t}$ for each time point $t$ (Equation (24)). Correspondingly, the new fractions are given by $N_{new} = N_\infty - N_{old}$ (Equation (26)). Similarly, the total fraction of RNA in the cytosol is given by $C_\infty = N_\infty \cdot \frac{\tau}{\lambda}$ (Equation (23)), so that $C_{old} = C_\infty \cdot \left( \frac{(\nu+\tau) \cdot e^{-\lambda t} - \lambda \cdot e^{-(\nu+\tau)t}}{\nu+\tau-\lambda} \right)$ (Equation (25)) and $C_{new} = C_\infty - C_{old}$ (Equation (27)). Nuclear and cytosolic $\frac{New}{Total}$ RNA ratios were obtained by calculating $\frac{N_{new}}{N_\infty}$ and $\frac{C_{new}}{C_\infty}$, respectively. Then, these ratios were used to estimate back the parameter set $\Theta = (\tau + \nu, \lambda)$ by fitting the two-compartment model as described in Data and Methods 2.2.11. 100 bootstrap runs were performed for each parameter combination.

### 2.2.15 Motif Search with HOMER

The reliable estimates (Data and Methods 2.2.12) of the nuclear and cytosolic RNA half-lives were categorized into 4 different groups based on quartiles. These groups describe 1) fast (0-25%-percentile) 2) moderately fast (25-75%-percentile) 3) moderately slow (50-75%-percentile) and 4) slow (75-100%-percentile) transcripts. The analysis was performed at the gene level. For this purpose, each 3'UTR was mapped to one gene. A 3'UTR was excluded from further analyses if it mapped to multiple gene identifiers. Subse-

quently, all exonic sequences of the respective genes were retrieved from the ENSEMBL annotation [99], and *de novo* RNA sequence motifs were called with HOMER [100] using default settings. The remaining groups not scanned for motifs were used as a background sequence dataset for each group's motif search. Walter Sandt performed this analysis.

### 2.2.16 Used Software

Quality metrics of sequencing data were assessed with fastqc (v0.11.9) [91], and sequencing adapters trimmed with cutadapt (v1.18) [92] before sequence alignment with NextGenMap (v0.5.5) [93].

Simulation frameworks were implemented in Python (v3.9.6) [101] using additional non-base packages: pandas (v1.4.3) [102], [103] for data table handling, scipy (v1.8.1) [104] and numpy (v1.23.1) [105] for numerical operations and statistics, and BioPython (v1.79) [106] for handling genomic data.

The two-compartment model was implemented in Python (v3.9.6) [101] and R (v4.0.1) [107]. Several additional non-base python packages were used: pysam (v0.19.1; visit `https://github.com/pysam-developers/pysam`) [108] for handling sequence alignment BAM files, pandas (v1.4.3) [102], [103] for data handling, scipy (v1.8.1) [104] and numpy (v1.23.1) for numerical operations and statistics, and pandas (v1.4.3) [102], [103] for data table handling. Furthermore, several non-base R packages were used: mcmc (v0.9-7) [109] for MCMC sampling, deoptim (v2.2-6) [97] for global optimization, and doParallel (v1.0.17) [110] for multicore processing.

For data table handling and visualization with Python, the packages pandas (v1.4.3) [102], [103], matplotlib (v3.5.2) [111] and seaborn (v0.13.0) [112] were used. Analogously, visualization and data table handling in R was performed with non-base packages reshape2 (v1.4.4) [113] ggplot2 (v3.3.5) [114], ggpubr (v0.4.0) [115] and gridExtra (v2.3) [116]. Schematic figures were created with Inkscape (v1.0.1; visit `https://inkscape.org/?switchlang=en`).

# 3 Results

This Results section is divided into two parts. In the first section, a probabilistic modeling framework for quantification of RNA metabolism was developed, and its performance was thoroughly assessed. As a 'ground truth' dataset of RNA metabolic rates does not exist, the model performance was tested on various simulations. Subsequently, the method was compared to GRAND-SLAM [72], a widely applied tool to estimate RNA half-lives from metabolic labeling sequencing data.

In the second part of this section, the probabilistic modeling framework was applied to two SLAM-seq time series of HeLa-S3 synchronized to be in Mitosis and G1 cell cycle phase (Data and Methods 2.1.1; Figure 9).
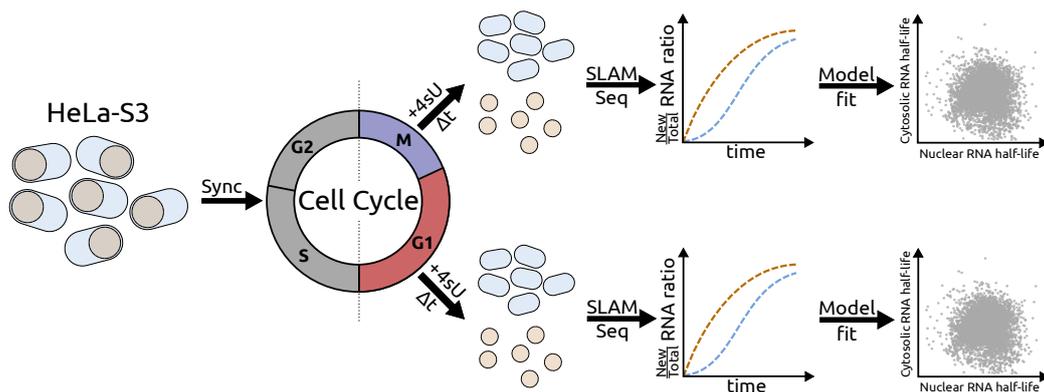


**Figure 9: Overview of the SLAM-seq Cell Cycle Experiments.** HeLa-S3 cells were synchronized regarding the cell cycle. Two SLAM-seq time series were generated, one in Mitosis and one in the G1 cell cycle phase. Cells were fractionated to obtain nuclear and cytosolic RNA samples. Subsequently, the model was applied to estimate $\frac{New}{Total}$ RNA ratios for each compartment and cell cycle phase. Then, cytosolic RNA half-lives were derived from these ratios by fitting the two-compartment model.

After data pre-processing and model fit (Figure 9), the resulting cell cycle-specific RNA half-lives were used to elucidate putative differences between the RNA metabolism during the cell cycle. Furthermore, they were compared to estimates from the literature. Lastly, the RNA half-lives were used to search for putative regulators of RNA metabolism, such as RBPs and RNA sequence elements.

# 3.1  **Model Development**

### 3.1.1    **A novel EM Algorithm estimates the Labeling Efficiency with high Precision**

In our first implementation of the modeling framework [80], the EM algorithm estimated labeling efficiencies for each measured 3'UTR and time point. The final labeling efficiency estimate was then obtained by calculating the median of the resulting labeling efficiency distribution. Subsequently, this 'global' labeling efficiency estimate was used to estimate $\frac{New}{Total}$ RNA ratios for each measured 3'UTR. While this approach worked well with the corresponding SLAM-seq time series data from unsynchronized HeLa-S3 cells, its application led to inconsistent bimodal labeling efficiency distributions between time points in the SLAM-seq cell cycle data presented here (Supplemental Figure S1). Consequently, the median-based approach for finding a 'global' labeling efficiency estimate resulted in calculating $\frac{New}{Total}$ RNA ratios to which the two-compartment model could be poorly fitted.

One possible explanation for the inconsistent labeling efficiency estimates is the time point-specific expression of certain transcripts during the cell cycle. The subsequent measurement of the corresponding 3'UTRs could skew the labeling efficiency distribution, leading to biased median estimates. To buffer this effect, only 3'UTRs, which are constantly expressed along a time series, can be accounted for when estimating the labeling efficiency. Further, instead of evaluating one labeling efficiency per detected 3'UTR, it can be assessed globally by considering the conversion counts of all regions simultaneously. This procedure is based on the assumption that the labeling efficiency is constant for each transcript. Although there might be some variability regarding the 4sU incorporation rate by Pol II during transcription, it can be assumed this polymerase does not incorporate 4sU selectively for each transcript, and, therefore, the variability is minor.

Another cause for the inconsistent estimation of the labeling efficiency could be a low number of T>C conversions in the data. Indeed, a manual inspection of the T>C conversion counts per time point indicated that many of the corresponding reads show 0 T>C conversions. Although these small numbers of T>C conversions could reflect low transcriptional activity, they might also be an artifact of the low 4sU incorporation rates. In the latter's

case, the EM would erroneously underestimate the labeling efficiency, leading to an over-estimation of the proportion of pre-existing transcripts.

Assuming that the 4sU incorporation follows a binomial distribution, reads from newly synthesized transcripts with low numbers of potential labeling sites should also have a low chance of carrying T>C conversions. The average number of potential labeling sites was inferred from the cell cycle data to assess how many expected T>C conversions can be observed within a sequencing read (Figure 10).
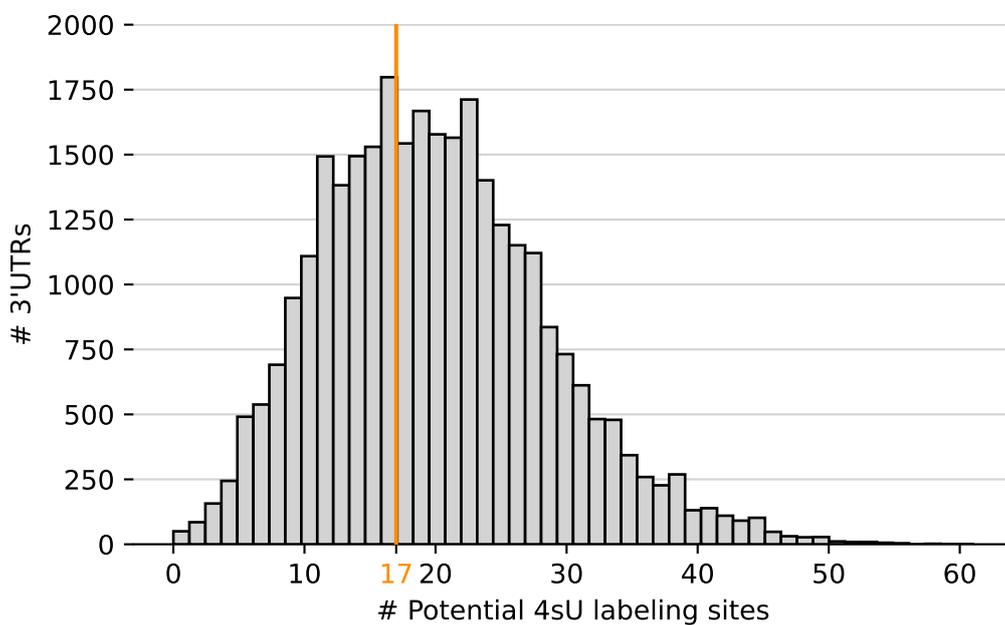


**Figure 10: Histogram of Potential Labeling Sites per 3'UTR.** Shown are the mean number of potential labeling sites in the respective reads of each 3'UTR (15 min time point sample of the nuclear fraction measured in Mitosis). The orange line depicts the mode of the distribution, indicating that most 3'UTRs contain 17 labeling sites per read.

Notably, the reads of most 3'UTRs contain 17 potential labeling sites (Figure 10). Based on this number, the expected number of T>C conversions was modeled by binomial distributions using different labeling efficiencies as event probabilities (Data and Methods 2.2.2; Figure 11).
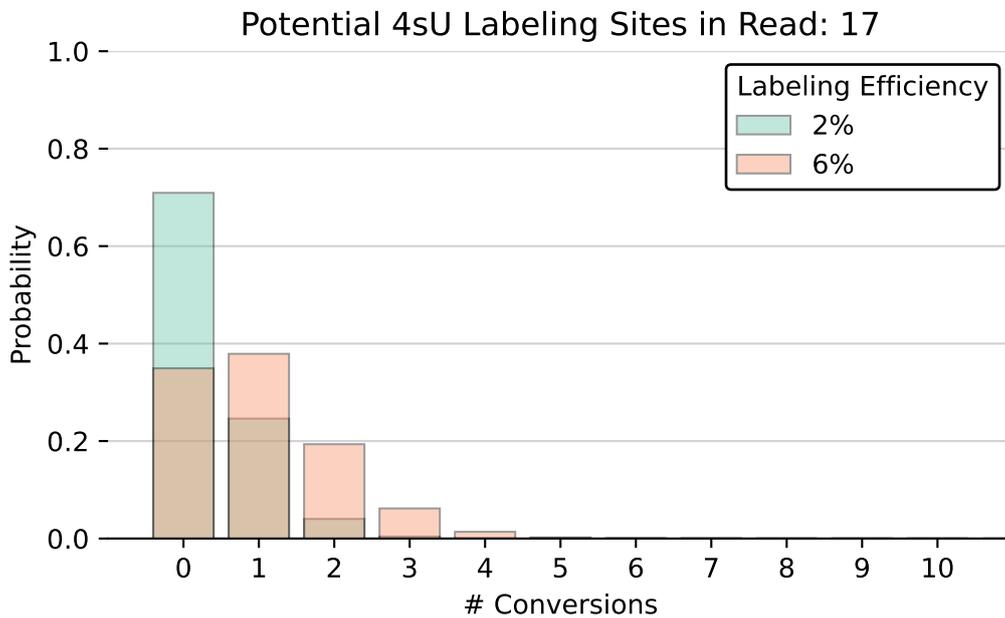
**Figure 11: Barplot of Expected T>C Conversions given 17 Potential Labeling Sites.** The colors represent the choices of the labeling efficiencies (4sU incorporation rate), which were used as event probabilities $p$ to generate conversion probabilities from binomial distributions. No T>C or C>X sequencing errors were accounted for in the binomial model to simulate 'optimal' 4sU incorporation conditions.

Assuming a labeling efficiency of 2% (which was reported to be in a realistic range [68]), the probability of observing 0 T>C conversions in a read given 17 labeling sites is roughly 70% (Figure 11), whereas the probability of observing 1 conversion is slightly higher than 20%. Even with 6% labeling efficiency, the probability of observing 0 T>C conversions is higher than 30%, whereas the probability of observing 1 conversion is slightly higher but still lower than 40%. Notably, the binomial model does not account for false-positive (T>C error) and false-negative (C>X error) rates but optimal conditions where a T>C conversion event is only dependent on the 4sU incorporation rate itself. Therefore, the probability of observing 0 T>C conversions might be higher in actual sequencing data.

One potential way to control for this bias is to infer the labeling efficiency from reads whose number of potential labeling sites exceeds a certain threshold. Figure 12 shows the expected number of T>C conversions in a read given 30 instead of 17 potential labeling sites (Data and Methods 2.2.2).

## Potential 4sU Labeling Sites in Read: 30
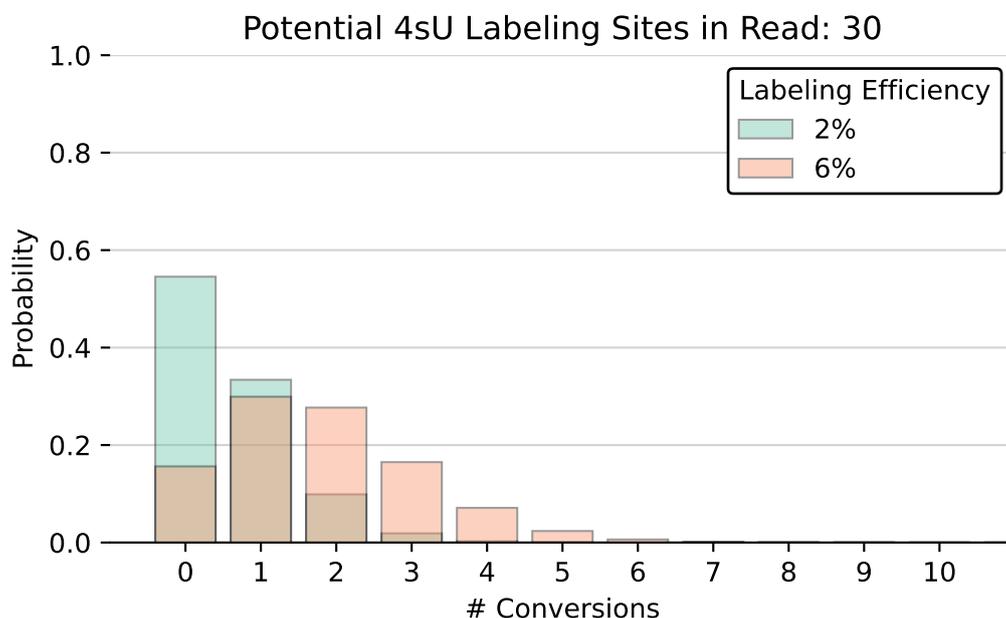


**Figure 12: Barplot of Expected T>C Conversions given 30 Potential Labeling Sites.** The colors represent the choices of the labeling efficiencies (4sU incorporation rate), which were used as event probabilities $p$ to generate conversion probabilities from binomial distributions. No T>C or C>X sequencing errors were accounted for in the binomial model to simulate 'optimal' 4sU incorporation conditions.

Assuming a labeling efficiency of 2%, the chance of observing 0 T>C conversions dropped from roughly 70% to less than 60% when a read contains 30 instead of 17 potential conversion sites (compare Figure 11-12). Also, the chance of observing at least 1 T>C conversion increases to more than 30%. If the labeling efficiency is 6%, the probability of observing 0 T>C conversions drops below 20%. Overall, these results suggest that a read pre-selection based on the number of potential labeling sites could increase the chance of capturing 4sU labeling-induced T>C conversions, improving the accuracy of the labeling efficiency estimation by the EM.

A suitable threshold of potential labeling sites per read was assessed via simulation (Data and Methods 2.2.2). Here, the labeling efficiency was fixed to 2%. 1,000 3'UTRs were generated with either 100 or 200 reads as coverage. The proportion of newly synthesized transcripts was simulated to range from 1-90%. T>C conversions were generated by a binomial model, accounting for putative sequencing errors. Given different numbers

of potential labeling sites per read, the EM was challenged to estimate the 'true' labeling efficiency of 2% (Figure 13).
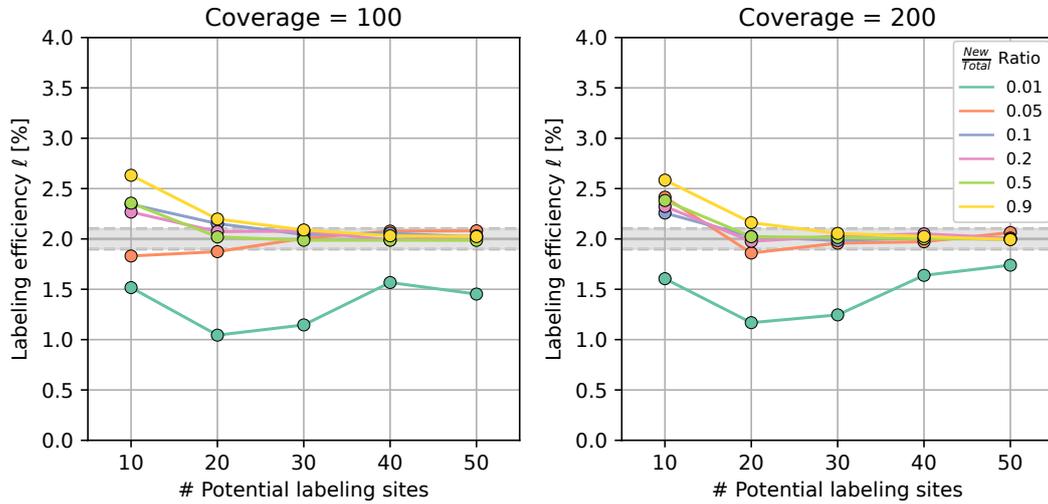


**Figure 13: Accuracy of the Labeling Efficiency Estimation with Increasing Labeling Sites per Read** In this simulation, 1,000 3'UTRs with a respective $\frac{New}{Total}$ RNA ratio ranging from 1-90% were generated (see colors). Each region was covered with 100 (left plot) or 200 reads (right plot), where the reads contained a certain number of potential labeling sites (x-axis). T>C conversions per read were sampled from a binomial model using a labeling efficiency of 2% as event probability. T>C ($\epsilon_+$) and C>X ($\epsilon_-$) sequencing errors of 0.1% were accounted for by the model. The labeling efficiency estimates computed by the EM were plotted on the y-axis, where the gray area indicates the 5% error region.

As the line plot in Figure 13 shows, at least 30 labeling sites per read are required to obtain labeling efficiency values that deviate with less than 5% relative error from the literature value of the labeling efficiency of 2% [68], given that the $\frac{New}{Total}$ RNA ratio ranges from 5-90%. If the ratio is 1%, the labeling efficiency could not be estimated with less than 5% error, irrespective of the used threshold of potential labeling sites. If a read contains only 10 potential labeling sites, the deviation in the labeling efficiency is higher than 5% for all simulated $\frac{New}{Total}$ RNA ratios. Generally, a higher coverage of 200 reads increases the accuracy of the labeling efficiency estimates. Here, the deviation in the labeling efficiency is already less than 5% for a broader range of $\frac{New}{Total}$ RNA ratios given 20 labeling sites per read (compare Figure 13).

As the main objective of the model is to determine the proportion of newly synthe-sized transcripts, the respective labeling efficiency estimates were also used to infer the $\frac{New}{Total}$ RNA ratios (Data and Methods 2.2.1; Figure 14).
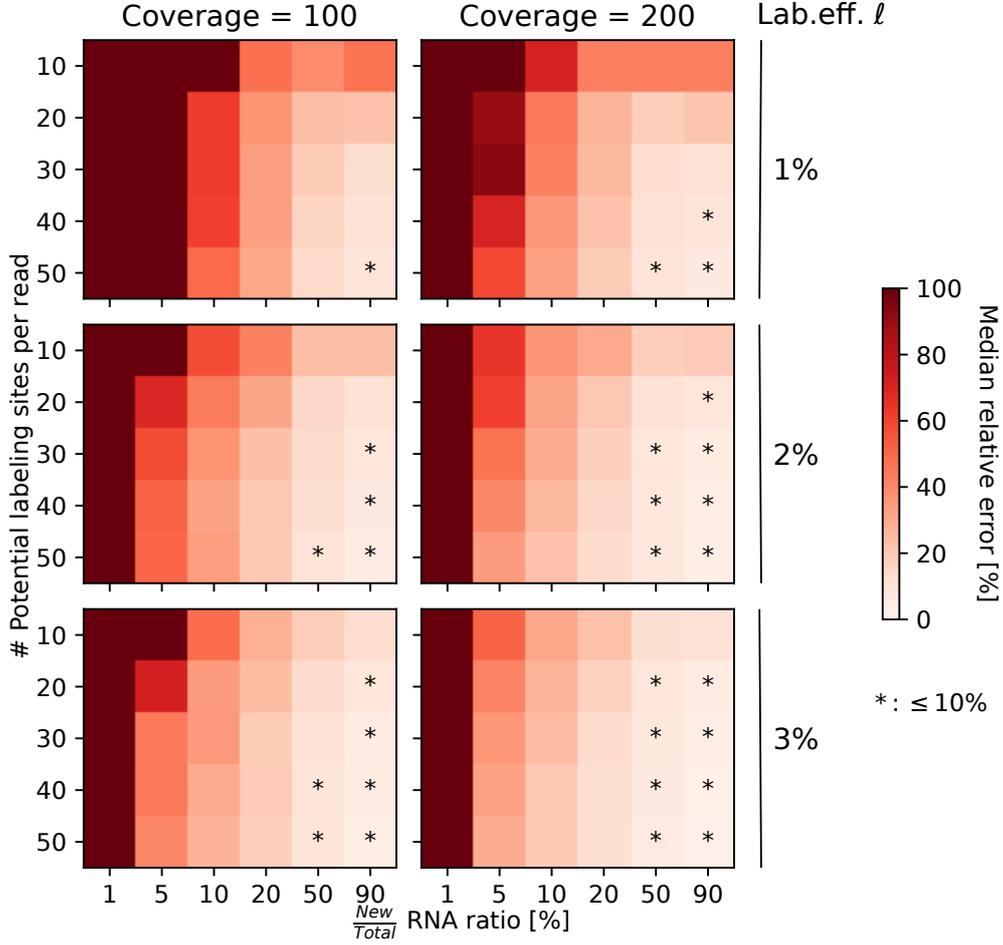
**Figure 14: Heatmap of the Relative Error of $\frac{New}{Total}$ RNA Ratios recovered from Reads with Different Numbers of Potential Labeling Sites.** 1,000 3'UTRs with 100 reads (left panel) or 200 reads (right panel) per region were generated for each simulation. Given a certain $\frac{New}{Total}$ RNA ratio (x-axis), T>C conversions were sampled from a binomial distribution using a pre-defined number of potential labeling sites per read (y-axis). Here, the labeling efficiencies $\ell$ of 1% (top panel), 2% (middle panel), and 3% (bottom panel) were used for the fraction of newly synthesized reads. The binomial models also accounted for T>C ($\epsilon_+$) and C>X ($\epsilon_-$) sequencing errors of 0.1% each. The color grading represents the median of the relative error of the $\frac{New}{Total}$ RNA ratio across all 1,000 3'UTRs. Asterisks depict errors equal to or lower than 10%.

The relative error of the $\frac{New}{Total}$ RNA ratio decreases with higher numbers of potential labeling sites per read, higher labeling efficiencies, higher coverage, and higher proportions of the $\frac{New}{Total}$ RNA ratio itself (Figure 14). In general, the lower-end ratios ranging from 1-10% have higher error rates. This behavior is expected since the proportion of newly synthesized transcripts decreases at lower ratios, reducing the number of reads with T>C con-

versions that can be differentiated from sequencing error rates. Therefore, the model is missing the statistical power to robustly infer the $\frac{New}{Total}$ RNA ratios. Additionally, the simulation was performed by fixating one $\frac{New}{Total}$ RNA ratio, and not a mixture of different ratios. In such a mixture setting, reads from 3'UTRs with higher ratios should stabilize the labeling efficiency estimation and, subsequently, the full range of $\frac{New}{Total}$ ratio estimates between 0-100%.

Based on the above results, the threshold of potential labeling sites per read was set to $\geq 30$. In theory, higher thresholds would even improve the labeling efficiency estimation process. However, the number of reads that pass these thresholds would decrease, which could influence the stability of the estimation procedure.

### 3.1.2   Performance Assessment of the EM Algorithm

Compared to the implementation of the EM algorithm from our previous work [80], the new version now incorporates two crucial changes. First, instead of estimating an individual labeling efficiency value for each genomic region, the latest version estimates one global labeling efficiency value from all genomic regions simultaneously. Second, it performs a pre-selection of reads that contain a certain number of potential labeling sites. In the previous simulation, 30 sites per read were determined to be a suitable threshold (see Results 3.1.1). To that end, all parameters, such as the number of labeling sites per read, the coverage, and the $\frac{New}{Total}$ RNA ratio, were kept constant. However, in actual sequencing data, all these parameters can differ between transcripts. Therefore, the EM algorithm was tested on simulated data with reads generated from actual human genomic sequences (Data and Methods 2.2.3-2.2.4). The necessity to obtain an accurate estimator of the 4sU incorporation rate is demonstrated, and the pre-selection step is further evaluated using this data. Lastly, the EM is tested against the corresponding implementation of GRAND-SLAM [72], a widely used tool to quantify RNA metabolic rates.

First, the impact of deviations in the labeling efficiency estimate on the $\frac{New}{Total}$ RNA ratio estimation was investigated. For this purpose, 1,000 3'UTRs were retrieved from actual sequences of GRCh38 (Data and Methods 2.2.3). A wide range of target $\frac{New}{Total}$ RNA ratios ranging from 1-99% was defined for each region. A 'true' labeling efficiency of 2% was set, which was used to model the expected number of T>C conversions at each ratio.

Then, the EM was applied to estimate the target $\frac{New}{Total}$ RNA ratios. However, the labeling efficiency was manually varied between 1-3%, while the T>C conversion counts generated with the 'true' labeling efficiency of 2% were kept constant. Subsequently, the deviation of the $\frac{New}{Total}$ RNA ratios estimates from their target values was assessed (Figure 15).
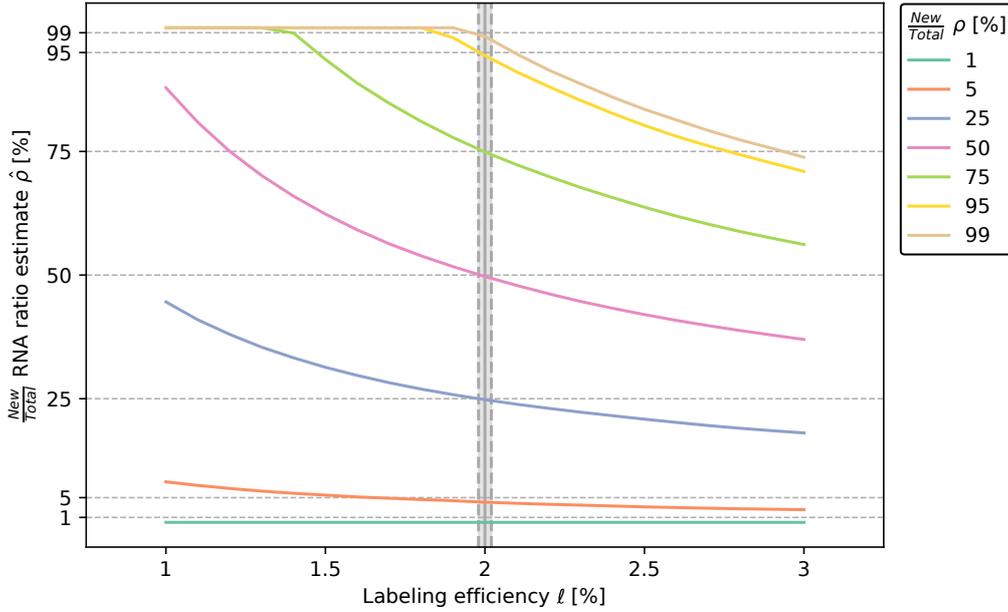


**Figure 15: Variation of the $\frac{New}{Total}$ RNA Ratio in dependence on the 4sU Incorporation Rate.**
Shown is a range of labeling efficiencies ranging from 1-3% (x-axis) used to estimate specific target $\frac{New}{Total}$ RNA ratios $\rho$ ranging from 1-99%. However, T>C conversion counts for each ratio were inferred from a 'true' labeling efficiency of 2%. 1,000 transcripts were simulated for each $\frac{New}{Total}$ RNA ratio $\rho$, each with a coverage of 100 reads. The median estimate of the $\frac{New}{Total}$ RNA ratios $\hat{\rho}$ of these transcripts were plotted as colored lines across the range of labeling efficiencies (y-axis). The gray vertical area depicts the 1% relative error region of the 'true' labeling efficiency of 2%. T>C conversions were generated by binomial models accounting for T>C sequencing errors ($\epsilon_+$) and C>X sequencing errors ($\epsilon_-$) of 0.1% each.

The $\frac{New}{Total}$ RNA ratios are sensitive to variations in the labeling efficiency (Figure 15). Assuming a 'true' labeling efficiency of 2%, slight deviations of this efficiency lead to comparably high deviations of the $\frac{New}{Total}$ RNA ratio estimates from their target ratios. The ratios are generally overestimated if the labeling efficiency falls below the 'true' labeling efficiency of 2% and vice versa. Notably, at a ratio of 1% the estimates tend to fall to 0%, likely due to the small numbers of reads from the new fraction. On the contrary, at high ratios of 95-99%, the estimates tend to increase to 100% if the labeling efficiency is underestimated. However, if the labeling efficiency is overestimated, the ratios decrease

rapidly. Based on this result, it was concluded that the labeling efficiency estimate should deviate with less than 1% relative error from its 'true' quantity.

Next, the simulation was scaled up to test the performance of the presented EM with and without read pre-selection. 15,000 transcripts were generated from annotated 3'UTRs of GRCh38 (Data and Methods 2.2.4). The coverage of each transcript was sampled from a beta-binomial distribution with a mean of 100 reads (Figure 4), resulting in a total number of roughly 1.5 million reads. Further, each transcript was assigned a random $\frac{New}{Total}$ RNA ratio between 1-99% drawn from a uniform distribution. Subsequently, T>C conversions for each read were generated using binomial models, accounting for T>C sequencing errors ($\epsilon_+$) and C>X sequencing errors ($\epsilon_-$). Both $\epsilon_+$ and $\epsilon_-$ were set to 0.1%. The labeling efficiency ranged from 0.1-5%. 5 replicates were generated for each parameter combination. First, the impact of the labeling site threshold on the accuracy of labeling efficiency estimation was assessed using this data (Figure 16).
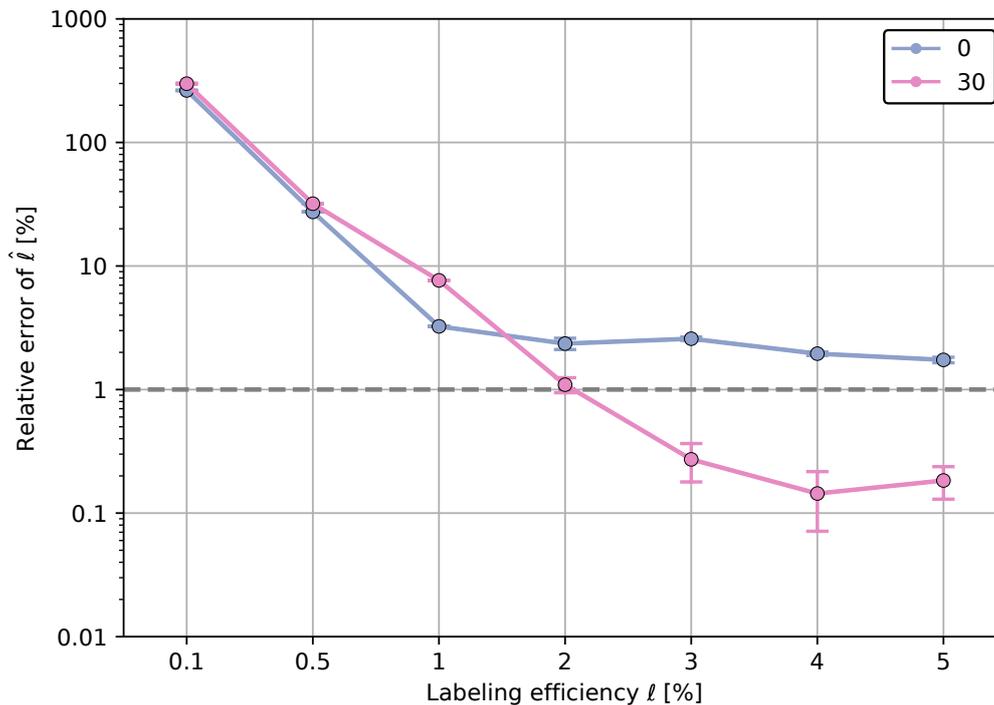
**Figure 16: Impact of Read Pre-selection on Labeling Efficiency.** Different labeling efficiencies $\ell$ are shown, ranging from 0.1-5% (x-axis), which were estimated by the new EM with and without read pre-selection. The threshold was set to $\geq 0$ (which effectively includes all reads; blue line) or $\geq 30$ (pink line). The relative errors of the respective labeling efficiency estimates $\hat{\ell}$ are plotted on the y-axis. The horizontal dashed line depicts the 1% relative error threshold. 5 replicates were simulated for each labeling efficiency. Points represent the means of the respective replicates and error bars the corresponding standard error. 15,000 transcripts were simulated for each replicate. A coverage was randomly sampled from a beta-binomial distribution with a mean of 100. A random $\frac{New}{Total}$ RNA ratio ranging between 1% and 99% was uniformly assigned to each transcript. T>C conversions were generated by binomial models using the target labeling efficiency $\ell$ ranging from 0.1-5%, while accounting for T>C sequencing errors ($\epsilon_+$) and C>X sequencing errors ($\epsilon_-$) of 0.1% each. Note that due to the logarithmic y-axis, the standard errors of smaller estimates appear larger than those of higher estimates.

A pre-selection of reads with at least 30 potential labeling sites yields more accurate labeling efficiency estimates than no pre-selection ($\geq 0$ threshold, Figure 16). Suppose the labeling efficiency is 2%. In that case, the relative error of its corresponding estimates is in the range of 1% upon read pre-selection. In contrast, without pre-selection, the relative error was roughly 3% (compare Figure 16). In the case of even higher labeling efficiencies ($\geq 3$%), the error dropped below 1% upon pre-selection. In contrast, the error always exceeded the 1% threshold if no read pre-selection was performed. If the labeling efficiency was set to 0.1%, 0.5%, or 1%, the relative errors of the labeling efficiency estimates were higher with pre-selection than without. A possible explanation is the small number

of reads that pass the labeling site filter. Further, the background sequencing errors $\epsilon_+$ and $\epsilon_-$ have a higher impact on the estimation precision as they have a similar magnitude as the labeling efficiency itself. However, actual SLAM-seq data typically has higher coverage and labeling efficiency [68]. Therefore, 30 potential labeling sites per read can be considered a reasonable pre-selection threshold.

Next, the performance of the new EM algorithm and GRAND-SLAM were compared on this simulated data. GRAND-SLAM performs a fundamentally different read pre-selection step than the here presented method. Instead of selecting reads that exhibit a certain minimum number of potential labeling sites, GRAND-SLAM pre-selects reads whose numbers of T>C conversions exceed those expected to be observed from background sequencing errors [72]. For comparison, the presented EM algorithm and the corresponding implementation of GRAND-SLAM were applied to estimate a wide range of labeling efficiencies from 0.1-5% (Figure 17).
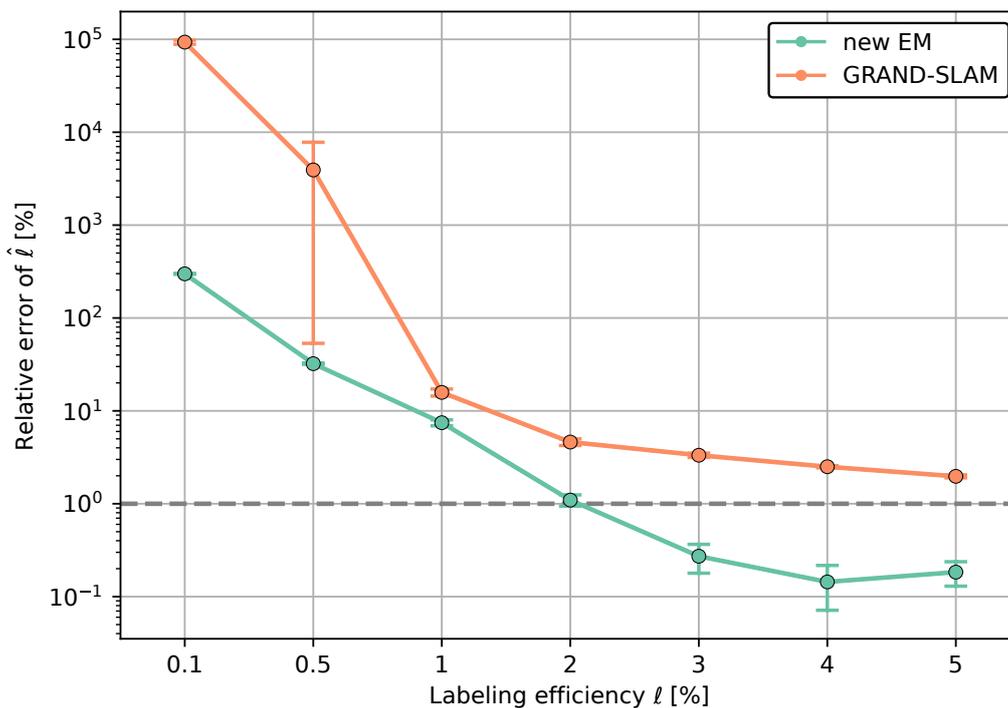
**Figure 17: Performance of the New EM and GRAND-SLAM in Estimating the Labeling Efficiency.**
Shown are different labeling efficiencies $\ell$ ranging from 0.1-5% (x-axis), which were to be estimated
by either the new EM (green) or GRAND-SLAM (orange). The relative error of the respective labeling
efficiency estimates $\hat{\ell}$ is plotted on the y-axis. The horizontal dashed line depicts the 1% relative er-
ror threshold. 5 replicates were simulated for each labeling efficiency. Points represent the means
of the respective replicates and error bars the corresponding standard error. 15,000 transcripts
were simulated for each replicate. T>C conversions were generated by binomial models using a
target labeling efficiency $\ell$ ranging from 0.1-5%, while accounting for T>C sequencing errors ($\epsilon_+$)
and C>X sequencing errors ($\epsilon_-$) of 0.1% each. Note that due to the logarithmic y-axis, the standard
errors of smaller estimates appear larger than those of higher estimates.

Assuming realistic error rates ($\epsilon_+ = 0.1$% and $\epsilon_- = 0.1$%) [117], both EM algorithms pro-
vide labeling efficiency estimates with relative errors less than 10% if the labeling effi-
ciency is at least 2% (compare Figure 17). At a labeling efficiency of 1%, GRAND-SLAM's
estimate has a relative error of roughly 20%, whereas the error of the presented EM is
below 10%. Notably, GRAND-SLAM computed estimates with more than 1000% relative
error if the labeling efficiency was set to 0.1% or 0.5%. This observation is because only
a small number of reads (less than 10,000) would fulfill the preselection criteria. There-
fore, GRAND-SLAM's EM estimates the labeling efficiency either at the starting value of
100% (see Data and Methods 2.2.4) or at a slightly lower value. In theory, GRAND-SLAM
would even stop the computation, as at least 10,000 reads are required for the estima-
tion (noted by the authors [72]). However, the implementation was still run for complete-

ness. In contrast, the presented EM algorithm could recover a labeling efficiency of 0.5% with less than 40% relative error. If the labeling efficiency was set to the same value as the sequencing errors of 0.1%, the relative error increased to roughly 300%. This result demonstrates that the presented EM outperforms GRAND-SLAM regarding the labeling efficiency estimation.

Complementary to the labeling efficiency estimation, both methods were used to determine the $\frac{New}{Total}$ RNA ratio of each simulated transcript. Subsequently, the deviation of these estimates from their respective 'true' ratios was assessed (Figure 18).
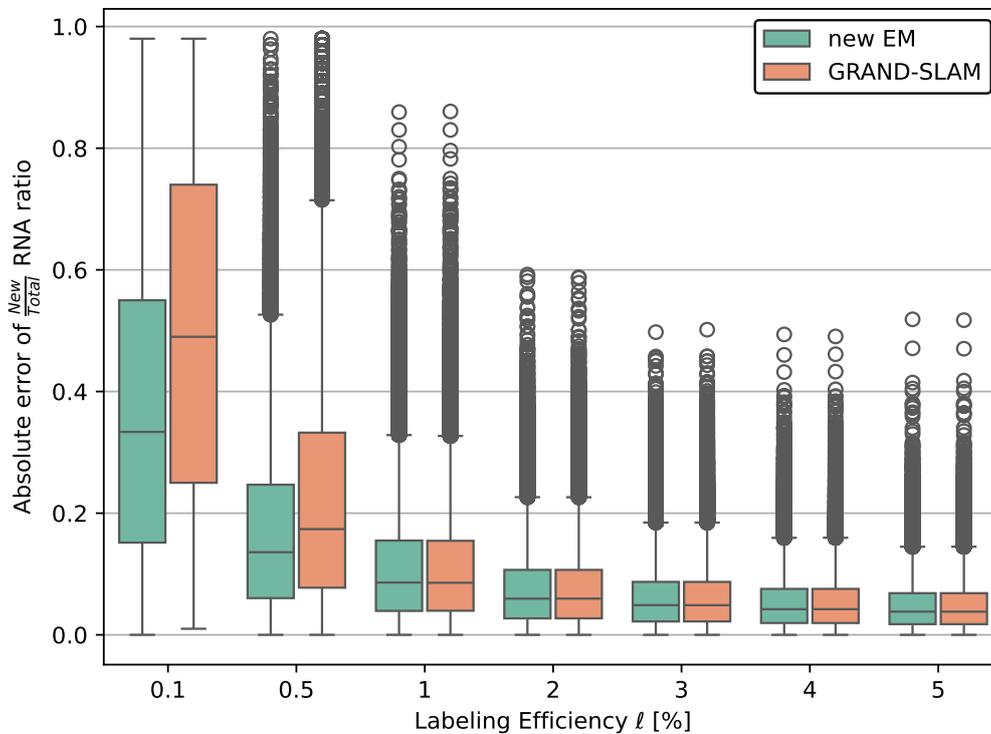


**Figure 18: Performance of the New EM and GRAND-SLAM in Estimating the $\frac{New}{Total}$ RNA Ratio.** Shown is a range of labeling efficiencies $\ell$ (x-axis), which were to be estimated by either the presented EM (green) or GRAND-SLAM (orange). 5 replicates with 15,000 transcripts were simulated for each labeling efficiency. A coverage was randomly sampled from a beta-binomial distribution with a mean of 100. A random $\frac{New}{Total}$ RNA ratio ranging between 1% and 99% was uniformly assigned to each transcript. T>C conversions were generated by binomial models using a target labeling efficiency $\ell$ ranging from 0.1-5%, while accounting for T>C sequencing errors ($\epsilon_+$) and C>X sequencing errors ($\epsilon_-$) of 0.1% each. The absolute errors of the $\frac{New}{Total}$ RNA ratios across all 5 replicates are plotted on the y-axis.

As shown by the boxplots in Figure 18, both the presented EM and GRAND-SLAM compute $\frac{New}{Total}$ RNA ratios with median absolute errors of roughly 10% and inter-quartile ranges (IQR) of roughly 5-15%, if the underlying labeling efficiency is set to at least 1%. With increasing labeling efficiency, the median absolute errors of the $\frac{New}{Total}$ RNA ratios drop from roughly 10% (at 1% labeling efficiency) to about 4% (at 5% labeling efficiency) with both methods. At the same time, these errors of both methods converge to a detection limit. This convergence is expected, as the fraction of newly synthesized reads was sampled from a binomial distribution, using the 'true' $\frac{New}{Total}$ RNA ratio as the event probability. This procedure introduces some variability to the data according to the variance $np(1-p)$ of the binomial distribution, which is challenging to capture by both methods. For the labeling efficiency scenarios of 0.1% and 0.5%, the presented EM computes estimates with median absolute errors of roughly 30% (IQR: roughly 15-55%) and 10% (IQR: roughly 5-25%), respectively. Despite the high variability, the EM outperforms GRAND-SLAM in these scenarios (compare Figure 18). In theory, GRAND-SLAM would have stopped computation in these scenarios due to its read pre-selection cut-off of at least 10,000 reads.

In summary, the results of the simulations show that the EM algorithm with a global optimization approach achieves a high degree of accuracy in estimating the labeling efficiency. Combined with a pre-selection of reads with at least 30 potential 4sU labeling sites, the presented EM algorithm performs with higher precision than the corresponding EM algorithm of the widely used tool GRAND-SLAM. In direct comparison, it also outperforms GRAND-SLAM even in scenarios with low labeling efficiency, which could occur at early time points of a metabolic labeling time series. However, the variability of labeling efficiency estimates (and the subsequent $\frac{New}{Total}$ RNA ratio) is still comparatively high when the labeling efficiency falls below 1%. Therefore, it is advisable to critically evaluate the estimates when such magnitudes are observed in experimental data. But, with a typical labeling efficiency in the range of 2%, this problem should affect SLAM-seq data to less extent. Overall, the simulations demonstrated that the presented EM algorithm is a promising method to help analyze metabolic sequencing data.

# 3.2  RNA Metabolism along the Cell Cycle

The performance of the presented EM algorithm was evaluated on simulated data in the Results section 3.1. Simulations help investigate a particular method's performance and limitations, but they remain simplifications.  The modeling assumptions on which they are based may not capture all the technical or biological effects that could occur in actual experimental data. If the assumptions are not chosen appropriately, a method in question may even fail with such data sets in the worst case.  Therefore, the EM presented here was tested as part of the modeling framework on data from two SLAM-seq time series measured in HeLa-S3 cells generated exclusively for this project (Data and Methods 2.1.1).

An essential feature of the two new SLAM-seq time-series datasets generated for this work is that the cells were synchronized with respect to the cell cycle.  This synchronization contrasts with the data we used in our previous study, which was obtained from unsynchronized cells [80].  Such data provides insight into bulk RNA dynamics but only limited information about changes in transcript metabolism during the cell cycle.  Therefore, the two SLAM-seq time series were generated from cells in G1 and Mitosis to assess putative cell cycle-specific fluctuations in RNA metabolism. Before applying the two-compartment model, the data was pre-processed, including thorough quality control of sequencing reads, followed by read alignment and T>C conversion count for each time-point of a respective time-series (see Data and Methods 2.1.1 and Data and Methods 2.2.5-2.2.7).

### 3.2.1  Estimating Cell Compartment-specific $\frac{New}{Total}$ RNA Ratios

After the SLAM-seq time series had been pre-processed, the EM algorithm was applied to estimate the labeling efficiency for each time point of a respective time series. In G1, the labeling efficiency ranged from 2.76-5.82% labeling efficiency in the nucleus and 1.58-5.81% in the cytosol. Correspondingly, it ranged from 2.77-4.98% in the nucleus and 2.12-5.23% in the cytosol during Mitosis.  Subsequently, these labeling efficiency estimates were used by the modeling framework to infer $\frac{New}{Total}$ RNA ratios for each detected 3'UTR (Data and Methods 2.2.1).  In total, estimates for 15,779 3'UTRs were captured in G1 and 16,517 in Mitosis.

The $\frac{New}{Total}$ RNA ratios are expected to increase over time due to the duration of 4sU labeling. Such an increase was monitored in each compartment and cell cycle phase (Figure 19).
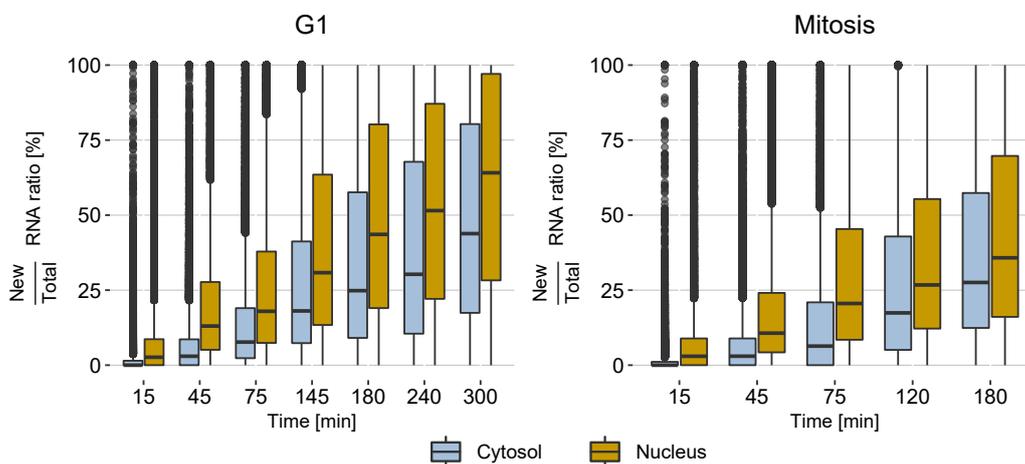


**Figure 19:** $\frac{New}{Total}$ **RNA Ratios estimated by the presented Modeling Framework from the SLAM-seq Cell Cycle Experiments.** Shown are boxplots of estimated $\frac{New}{Total}$ RNA ratios for the nuclear (darkorange) and cytosolic (blue) compartments over time for both G1 (left) and Mitosis (right). Each boxplot represents estimates from 15,779 3'UTRs detected in G1 and 16,517 3'UTRs in Mitosis.

Figure 19 shows that in both time series, the nuclear and cytosolic $\frac{New}{Total}$ RNA ratios increase over time. Notably, the median nuclear ratios exceed the cytosolic ones at each time point, irrespective of the cell cycle phase. Further, considering the median ratios per time-point, the nuclear estimates almost increase linearly over time. In contrast, the cytosolic ratios increase with some delay, which can be explained by the compartmentalization of the cell. Transcription occurs in the nucleus, for which newly synthesized transcripts will be detected in this compartment first. Consequently, a transcript must be exported from the nucleus before it can be measured in the cytosol. The delay further suggests that the nuclear export of RNA does not happen immediately after transcription, likely due to post-transcriptional processing steps. When comparing the median $\frac{New}{Total}$ RNA ratios between G1 and Mitosis over time, the ratios appear slightly elevated in G1. This variability might result from higher transcriptional activity in G1 but could also reflect technical variance. Reassuringly, the variance of the $\frac{New}{Total}$ RNA ratios increase over time, indicating that the model quantifies a wide range of transcript dynamics.

### 3.2.2    An mRNA spends most of its Life in the Nucleus, not the Cytosol

Having determined $\frac{New}{Total}$ RNA ratios, the two-compartment model was fitted to these quantities to obtain a nuclear vanishing rate $(\nu + \tau)$ and a cytosolic degradation rate $\lambda$ for each measured 3'UTR. The fit was performed with a two-dimensional MCMC followed by numerical optimization (Data and Methods 2.2.11). In contrast to our previous implementation, $(\nu + \tau)$ and $\lambda$ were fitted jointly to the $\frac{New}{Total}$ RNA ratios of each respective 3'UTR. Subsequently, the obtained rates were converted to nuclear and cytosolic RNA half-lives by calculating $\frac{ln(2)}{\nu+\tau}$ and $\frac{ln(2)}{\lambda}$, respectively. Additionally, stringent reliability criteria were applied to yield highly reliable RNA half-lives for each compartment and cell cycle phase (Data and Methods 2.2.12). This procedure determined 1,931 highly reliable nuclear and 956 cytosolic estimates in G1. Correspondingly, 2,295 highly reliable nuclear and 907 cytosolic estimates were identified in Mitosis. For each cell cycle phase, the nuclear RNA half-lives were compared with the corresponding cytosolic estimates (Figure 20).
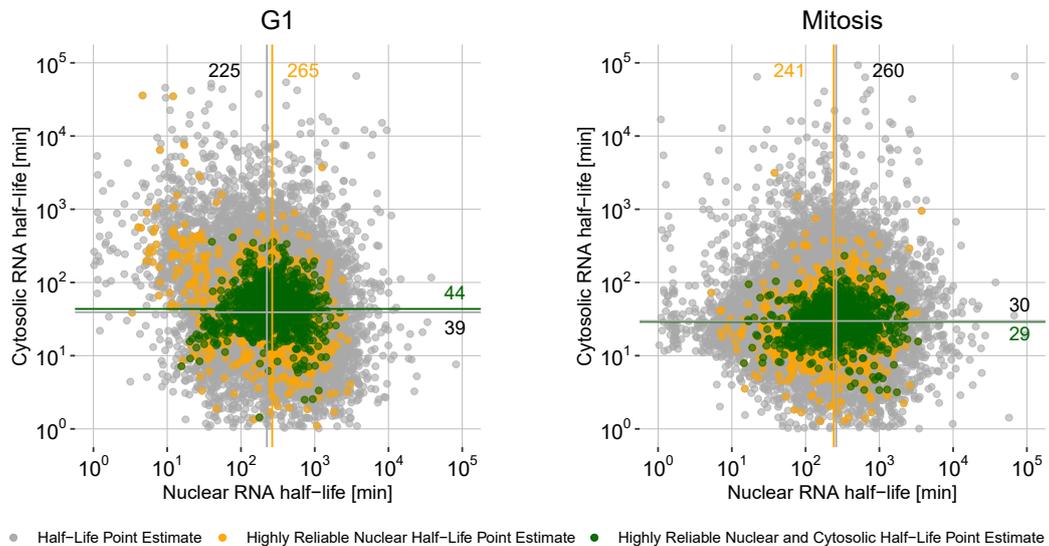


**Figure 20: Nuclear vs Cytosolic RNA Half-life Estimates from SLAM-seq Cell Cycle Experiments.** The nuclear RNA half-lives were plotted against the respective cytosolic estimates for both the G1 and Mitosis cell cycle phases. Yellow points highlight nuclear RNA half-life estimates, which were highly reliable (G1: n=1,931, Mitosis: n=2,295; Data and Methods 2.2.12). The green points highlight estimates found to be highly reliable in both the nucleus and the cytosol (G1: n=956, Mitosis: n=907). Gray points highlight half-life estimates from all other 3'UTRs. The lines and numbers in the respective colors depict median RNA half-lives in minutes.

Considering the reliable estimates, the median nuclear RNA half-life is 265 min (IQR: 140-485 min), whereas the median cytosolic RNA half-life is 44 min (IQR: 29-68 min) in the G1 cell cycle phase (compare Figure 20). In contrast, the median nuclear RNA half-life is 242 min (IQR: 128-432 min), and the cytosolic RNA half-life is 29 min (IQR: 20-43 min), considering the reliable estimates measured in Mitosis. These numbers indicate that the nuclear export of RNA is slow, while cytosolic degradation is fast in both cell cycle phases. Consequently, an average transcript remains in the nucleus at least 6-times as long as it remains in the cytosol. Further, some qualitative differences between the RNA half-lives in G1 and Mitosis can be observed, as the mitotic estimates are lower on average. However, these differences are minor, considering the assumption that nuclear and cytosolic half-life fractions could mix up completely upon nuclear envelope breakdown (NEBD) during Mitosis. In such a scenario, the half-lives of both compartments should have more or less the same magnitude. However, the NEBD was observed to happen within roughly 35 minutes from initial breakdown to nuclear reassembly [118]. Further, subcellular fractionation protocols rely on centrifugation [82] to separate intact nuclei from the cytoplasmic fractions. These two data properties might bear experimental difficulties in monitoring NEBD with subcellular metabolic labeling sequencing experiments. However, the observed RNA half-lives suggest that NEBD affects mitotic estimates to a minor extent. Replicate time series are needed to check whether the mitotic RNA turnover rates are robust, but further validation experiments were out of scope for this study.

The long nuclear RNA half-lives imply that the transcript abundances could be higher in the nucleus than in the cytosol. ERCC RNA spike-ins were used to normalize each sequencing library and, subsequently, calculate $\frac{Cyt}{Nuc}$ RNA abundance ratios for each cell cycle phase and time point to check this hypothesis (Data and Methods 2.2.13; Figure 21).
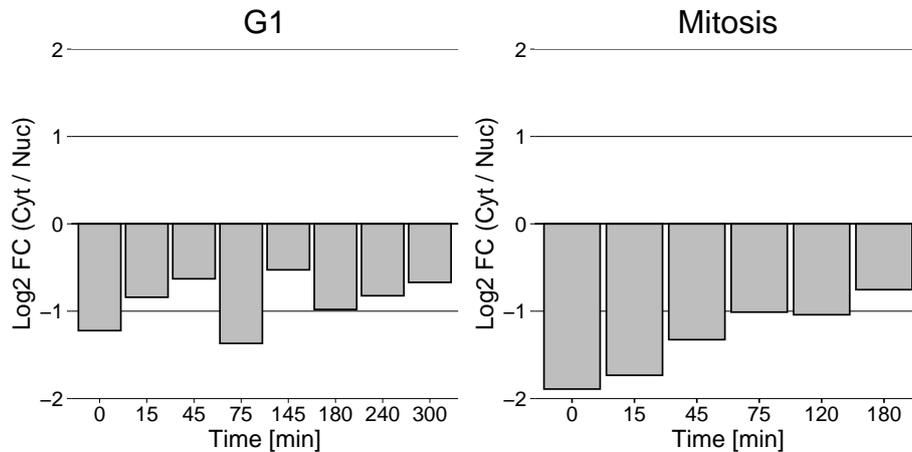
**Figure 21: Cytosolic vs Nuclear RNA Abundance Ratios during the Cell Cycle.** The left plot shows the log2 fold-change of $\frac{Cyt}{Nuc}$ RNA ratios for each time point in the G1 phase, whereas the right plot shows the ratios per time point for the mitotic phase. Negative values indicate higher RNA abundances in the nucleus, whereas positive values represent higher numbers of transcripts in the cytosolic fraction. RNA abundances were calculated with ERCC RNA spike-ins (Data and Methods 2.2.13).

The $\frac{Cyt}{Nuc}$ RNA abundance ratios of each time point in each cell cycle phase indicate higher RNA amounts in the nucleus compared to the cytosol (compare Figure 2.2.13). Interestingly, RNA amounts vary between the time points in both G1 and Mitosis samples. At the start of Mitosis (0 min), there is almost a 4-times higher amount of RNA in the nucleus, but this amount declines to more balanced abundances in the nucleus and cytosol over time. These more balanced ratios can be expected upon NEBD during Mitosis as the nuclear and cytosolic RNA fractions could mix up. Still, roughly 2 times more RNA is in the nucleus at the 180 min time point sample. Further, the $\frac{Cyt}{Nuc}$ RNA abundance ratios fluctuate during the G1 phase (especially at 75 min). Similar to the differences regarding the RNA half-lives from both cell cycle phases (compare Figure 20), these $\frac{Cyt}{Nuc}$ RNA abundance ratios could also result from technical variation. Overall, these ratios provide further evidence that an average mRNA spends most of its life in the nucleus, not the cytosol.

Although the RNA half-lives seem to be highly similar in both G1 and Mitosis, the metabolism of individual transcripts could differ along the cell cycle. The compartment-specific RNA half-lives from G1 and Mitosis were plotted against each other to investigate for such differences (Figure 22).
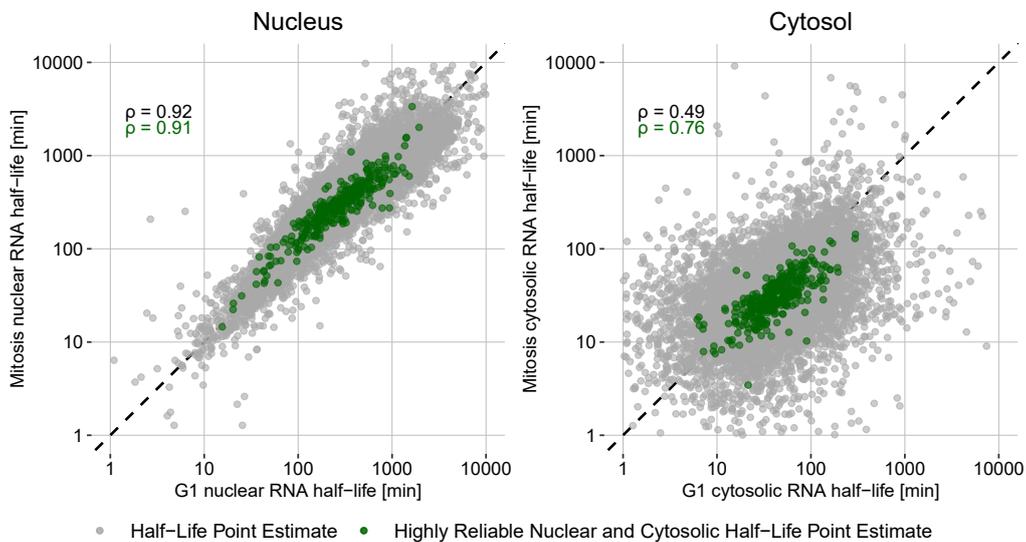
**Figure 22: Comparison Compartment-specific RNA Half-life Estimates between G1 and Mitosis.** Shown are nuclear (left plot) and cytosolic (right plot) RNA half-life estimates from G1 (x-axis) and Mitosis (y-axis) plotted against each other. Green points depict reliable half-life estimates (Data and Methods 2.2.12), whereas gray points reflect all other estimates. Spearman's $\rho$ correlation values of the estimates are indicated in the plots. In total, 14,814 3'UTRs were commonly detected in both cell cycle phases.

The reliable nuclear RNA half-lives from both cell cycle phases are highly correlated (Spearman's $\rho = 0.91$). In contrast, the corresponding cytosolic estimates show a comparably weaker correlation (Spearman's $\rho = 0.76$, compare Figure 22). While the correlation of all estimates is similar to the reliable ones in the nucleus (Spearman's $\rho = 0.92$), the correlation of all estimates is less pronounced in the cytosol (Spearman's $\rho = 0.49$). Although this inconsistency might reflect cell cycle-specific differences in cytosolic degradation rates, it could also be caused by technical variation. A similar discrepancy between actual replicate measurements was observed in our previous study, suggesting that technical variation is the leading cause [80].

### 3.2.3   Cytosolic RNA Half-life Estimates suffer from technical Variance

A higher number of replicate time series could be used to obtain more stable estimates for the cytosolic RNA half-lives. But as already mentioned, these measurements were out of scope for this study. One alternative is to tweak the experimental parameters, for instance, by increasing the coverage. As higher numbers of reads should have higher

chances to capture relatively small $\frac{New}{Total}$ RNA ratios but also lowly abundant RNA (due to high cytosolic degradation rates), a higher coverage might help to account for the technical variability observed in the cytosol.

A simulation was performed to check if an increasing coverage could improve the estimation accuracy of cytosolic RNA half-lives (Data and Methods 2.2.14). Therefore, a wide range of cytosolic RNA half-lives ranging from 10-100 min and different coverages ranging from 30-1,000 reads were used to simulate cytosolic $\frac{New}{Total}$ RNA ratios. Then, the modeling framework was challenged to estimate back the cytosolic RNA half-lives from these ratios to investigate the technical variance. The simulation was bootstrapped 100 times (Figure 23).
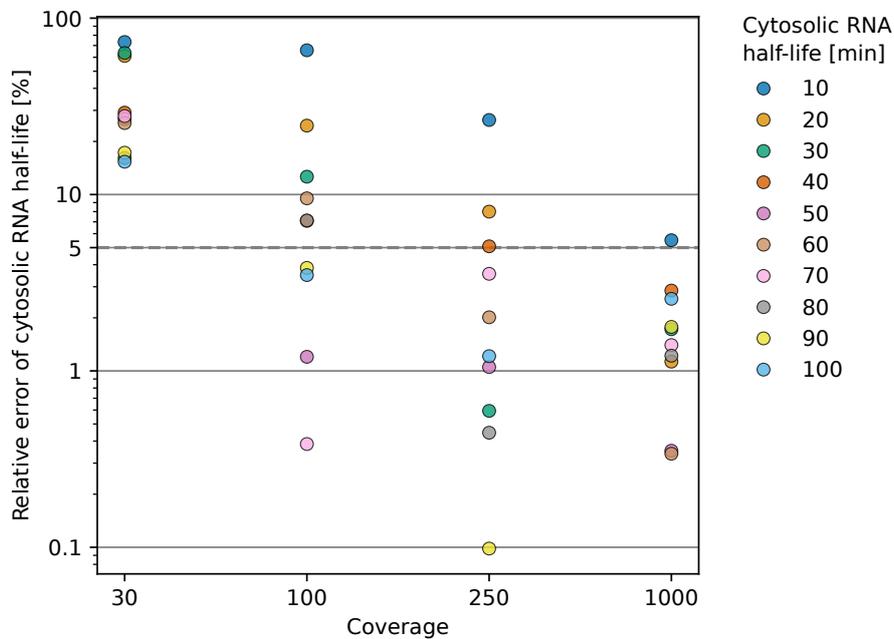


**Figure 23: Reproducibility of Cytosolic RNA Half-Life Estimates.** The point plot shows the relative error of a variety of cytosolic RNA half-lives ranging from 10-100 minutes (see colors) on the y-axis, in dependence of the coverage on the x-axis. Each cytosolic RNA half-life was used to define the proportion of reads from newly synthesized transcripts given a certain coverage ranging from 30-1,000 reads. Then, the two-compartment model was used to estimate the cytosolic RNA half-lives from the respective $\frac{New}{Total}$ RNA ratios. Each simulation was bootstrapped 100 times to yield median point estimates for each cytosolic RNA half-life (Data and Methods 2.2.14).

As shown in Figure 23, the relative errors of the cytosolic RNA half-life estimates shrink with increasing coverage. If the coverage is set to 30 reads, the whole range of distinct cytosolic RNA half-lives from 10-100 minutes is estimated with more than 10% error. Only if

the coverage is set to 1,000 reads, RNA half-lives higher than 20 minutes can be estimated with less than 5% relative error on average. Overall, the simulation results suggest that increasing coverage could help obtain more robust cytosolic RNA half-lives. This finding will help to reduce future experimental costs and resources on replicate time series. At the same time, these results show that handling the variability of cytosolic RNA half-lives remains a challenging task, for which advances in experimental and computational methods are needed in the future.

### 3.2.4   RNA Half-lives are consistent along the Cell Cycle

Although the compartment-specific RNA half-lives were found to be highly correlated between G1 and Mitosis (see Results 3.2.1), some variability could be observed. Therefore, the RNA half-lives obtained from G1 and Mitosis were tested for significant differences using a Wilcoxon rank-sum test. For robustness, only 3'UTRs with reliable parameter estimates in both cell cycle phases were considered (Figure 24).
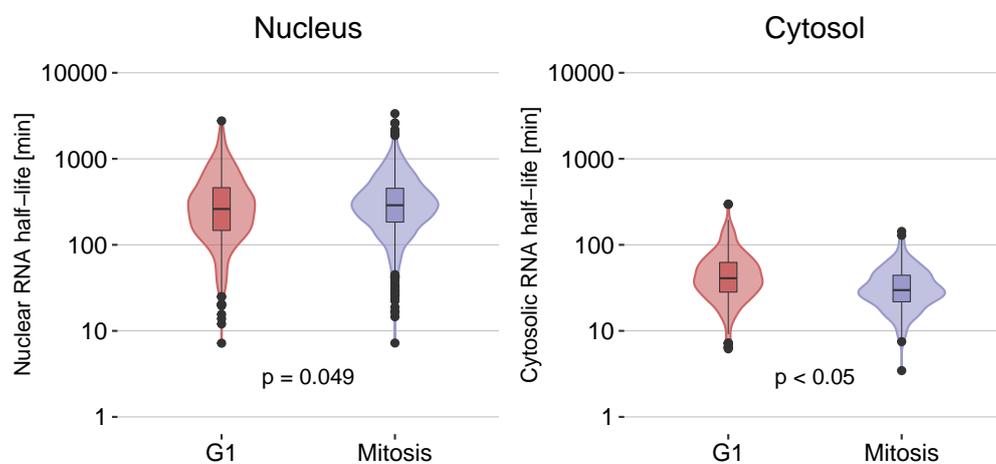


**Figure 24: Assessment of significant Differences between Compartment-specific RNA Half-life Estimates from the SLAM-seq Cell Cycle Experiments.** Shown are violin plots which represent the disributions of reliable RNA half-life estimates (Data and Methods 2.2.12) measured in the nucleus (left plot; $n = 766$) and the cytosol (right plot; $n = 326$). A Wilcoxon rank-sum test was used to test for significant differences between the compartment-specific RNA half-lives of both the G1 and Mitosis cell cycle phases (see p-values in the plots).

The violin plots in Figure 24 elucidate that the mitotic RNA half-lives have less variance than the G1 counterparts in both the nucleus and cytosol. Furthermore, the RNA half-lives

of both cell cycle phases are significantly different in the nucleus ($p = 0.049$; Wilcoxon rank-sum test) and the cytosol ($p < 0.05$; Wilcoxon rank-sum test). However, replicate time series are required to assess whether these differences are due to discrepancies in RNA metabolic rates during the cell cycle. Given the large technical variance in cytosolic estimates and the minor differences in the nucleus (see Results 3.2.2 and 3.2.3), this analysis does not suggest that RNA metabolic rates differ between phases of the cell cycle.

### 3.2.5 RNA Half-lives from synchronized Cells correlate with those from unsynchronized Samples

In the previous sections, cell cycle-specific RNA half-lives were tested against each other. However, they could deviate from estimates obtained from HeLa-S3 cells that are not synchronized regarding their cell cycle. The cell cycle-specific rates were compared to bulk measures from our previous work [80] (Data and Methods 2.1.2) to check the concordance between compartment-specific RNA half-lives from synchronized and unsynchronized cells (Figure 25).

Half−Life Point Estimate   Highly Reliable Nuclear and Cytosolic Half−Life Point Estimate
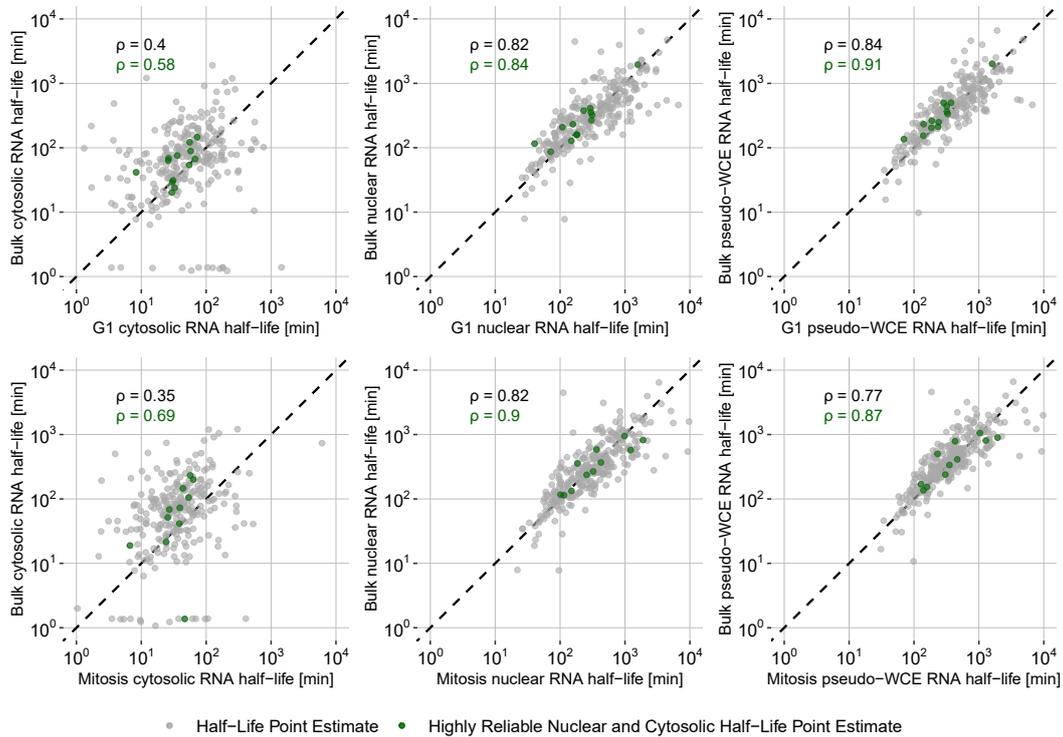
**Figure 25: Comparison of Cell Cycle-specific RNA Half-lives with Estimates from Unsynchronized Cells.** Shown are compartment-specific comparisons of bulk and cell cycle-specific RNA half-lives for the cytosol (left), the nucleus (middle), and pseudo-whole-cell (right) estimates. The top panel depicts the comparison of G1 estimates with the corresponding RNA half-lives from unsynchronized cells. Analogously, the bottom panel shows the comparison of Mitosis and the bulk estimates. Green points ($n = 11$) depict RNA half-lives which passed stringent reliability criteria (Data and Methods 2.2.12), and gray points ($n = 291$) all other parameter estimates. Spearman's $\rho$ correlation values of the estimates are indicated in the plots.

While the cytosolic RNA half-lives of the unsynchronized cells correlate moderately with the corresponding estimates from the cell cycle data (G1: Spearman's $\rho = 0.4$, Mitosis: Spearman's $\rho = 0.35$), they are in high agreement with the nuclear turnover rates (G1: Spearman's $\rho = 0.82$, Mitosis: Spearman's $\rho = 0.82$; Figure 25). This consistency still holds when comparing pseudo-whole-cell extract (pseudo-WCE) estimates, which reflect the sum of the nuclear and cytosolic RNA half-lives (G1: Spearman's $\rho = 0.84$, Mitosis: Spearman's $\rho = 0.77$). Overall, these results indicate that the RNA metabolism of synchronized cells does not differ remarkably from that of unsynchronized cells, at least in the nucleus. In contrast, the variability in cytosolic RNA half-life might reflect differences in the RNA metabolism along the cell cycle. However, this cannot be determined confidently due to the technical variation in cytosolic RNA half-life and the lack of replicate measurements.

### 3.2.6    Cell Cycle-specific RNA Half-lives are consistent with Estimates from the Literature

In the previous sections, the RNA half-life estimates of the modeling framework presented here were only compared internally with its own estimates. However, comparisons with external datasets are essential to evaluate the accuracy and credibility of the respective estimates. Therefore, the presented RNA half-lives were compared to estimates published by Schüler *et al.* (2014) [63] (Data and Methods 2.1.3). This dataset comprises RNA half-lives measured from WCEs of HEK293 and MCF7 cells. The authors applied 4sU labeling followed by biotinylation to separate newly synthesized from pre-existing transcripts. A simple exponential decay model was fitted as described in Schwanhäusser *et al.* (2011) [66] using these fractions. For this comparison, the nuclear and cytosolic RNA half-lives from each cell cycle phase were summarized to pseudo-WCE estimates (Figure 26).
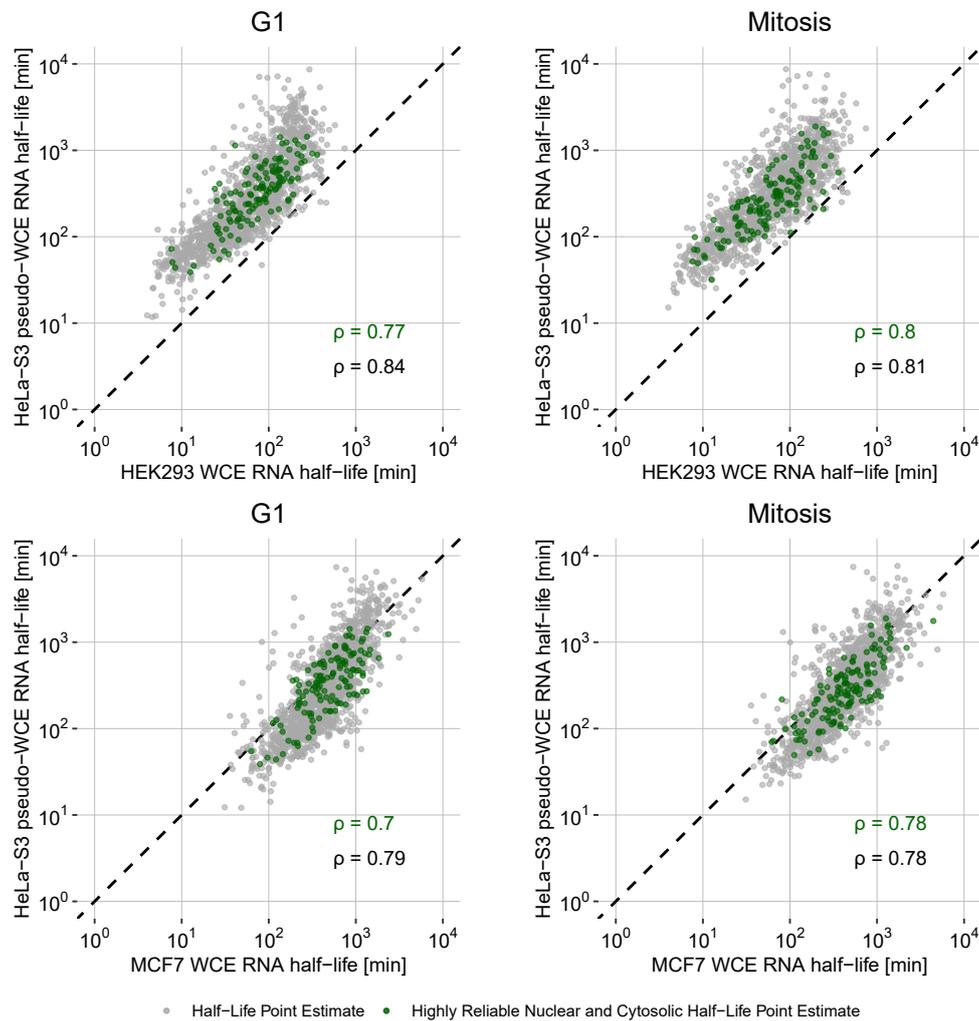
**Figure 26: Comparison of Cell Cycle-specific RNA Half-lives with Estimates from Whole-Cell Extracts.** Nuclear and cytosolic RNA half-lives were summarized to pseudo-WCE estimates for both G1 (left panel) and Mitosis (right panel). The pseudo-WCE estimates were plotted against the WCE RNA half-lives measured from HEK293 (top panel) and MCF7 (bottom panel) cells [63]. Highly reliable parameter estimates are indicated by green points (Data and Methods 2.2.12), whereas the gray points represent all other estimates. Spearman's $\rho$ correlation values of the estimates are indicated in the plots. For HEK293 cells, 1,385 RNA half-lives (reliable: 130) from HeLa-S3 estimates in G1 were compared (top left), and 1,386 (reliable: 135) in Mitosis (top right). For MCF7 cells, 1,357 RNA half-lives (reliable: 130) from HeLa-S3 estimates in G1 were compared (bottom left), and 1,357 (reliable: 137) in Mitosis (bottom right).

The reliable pseudo-WCE RNA half-lives of the HeLa-S3 cells are in high agreement with the WCE estimates obtained from HEK293 (G1: $\rho = 0.77$, Mitosis: $\rho = 0.8$) and MCF7 (G1: $\rho = 0.7$, Mitosis: $\rho = 0.78$) cells (Figure 26). Notably, the HeLa-S3 RNA half-lives are systematically longer than the respective measurements from HEK293 cells. In contrast,

they are systematically shorter than the ones from MCF7 cells. Such shifts might indicate cell line-specific differences in RNA turnover rates and have already been observed in the analyses of the first version of our modeling setup [80]. Nonetheless, these results show that the model presented here calculates compartment-specific estimates whose sums are in high agreement with corresponding measurements from the literature.

### 3.2.7   RNA-binding Proteins might modulate nuclear RNA Turnover Rates

RNA-binding proteins enable essential processing steps of mRNAs, such as alternative splicing and poly-adenylation, but also export from the nucleus [119]. In these processes, the composition of RBPs bound to a particular transcript plays a crucial role in modulating gene expression [120]. It can be assumed that RNAs with similar half-lives exhibit similar RBP compositions. Therefore, the half-lives of RNAs to which a certain RBP binds should differ from the corresponding half-lives of the unbound transcripts. The binding profiles of 120 RBPs were analyzed from eCLIP data deposited on ENCODE [85] (Data and Methods 2.1.4) to elucidate such differences. For each RBP, a set of unbound and bound transcripts was defined. The reliable RNA half-lives of these two groups were then tested for significant differences.

In total, 24 RBPs showed significant effects on nuclear RNA half-life in Mitosis (Supplemental Figure S2). Here, the binding of 2 of these proteins, RBM15 and KHSRP, appears to promote nuclear RNA export (Figure 27).
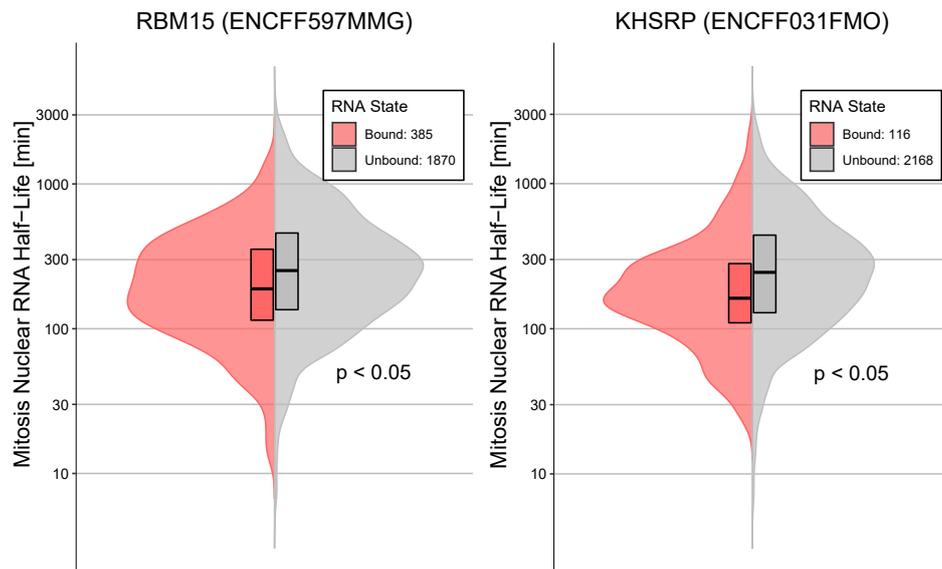
**Figure 27: Differences in mitotic nuclear RNA Half-lives upon binding by RBM15 and KHSRP.**
Each violin plot shows the distributions of the mitotic nuclear half-lives of transcripts that are either bound (red) or unbound (gray) by RBM15 (left plot) or KHSRP (right plot). A Wilcoxon rank-sum test was applied to test for significant differences in RNA half-life between the bound and unbound transcripts. Only those transcripts whose RNA half-lives met stringent reliability criteria (Data and Methods 2.2.12) were considered for this analysis. The ENCODE accession numbers for the corresponding eCLIP data are listed in brackets next to the RBP name.

As the violin plots in Figure 27 depict, the bound RNAs of RBM15 and KHSRP had significantly shorter nuclear turnover rates than the corresponding unbound transcripts. Interestingly, RBM15 has been associated with NXF1-dependent nuclear export of RNA in previous studies [73], [121] and was also found as a potential modulator in our earlier study [80]. In contrast, KHSRP emerged as a new candidate whose binding might enhance nuclear RNA export. The function of this protein has been linked to processes such as mRNA splicing and decay [122], [123]. According to the Uniprot database [124], it might also play a role in mRNA trafficking (Uniprot accession number: Q92945).

In the cytosol, 2 RBPs were found to significantly effect RNA half-life in Mitosis (Supplemental Figure S3). Here, the binding of PUM2 seems to promote cytosolic degradation (Figure 28).
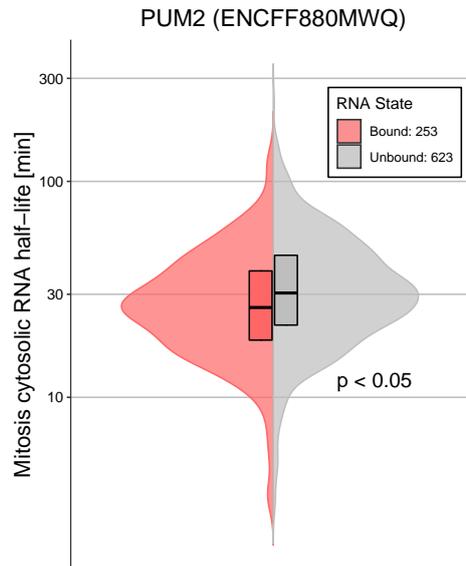
**Figure 28: Differences in mitotic cytosolic RNA Half-life upon binding by PUM2.** The violin plot shows the distributions of the mitotic cytosolic half-lives of transcripts that are either bound (red) or unbound (gray) by PUM2. A Wilcoxon rank-sum test was applied to test for significant differences in RNA half-life between the bound and unbound transcripts. For this analysis, only those transcripts were considered whose half-lives met stringent reliability criteria (Data and Methods 2.2.12). The ENCODE accession number for the corresponding eCLIP data is listed in brackets next to the RBP name.

PUM2 is an RBP that binds to the 3'UTR region of a transcript and recruits a deadeny-lase, which inhibits translation and, subsequently, induces mRNA degradation [125], [126]. Therefore, this function could be reflected by the shorter cytosolic RNA half-lives of its bound transcripts. For all other tested RBPs, nuclear and cytosolic RNA half-lives of the respective bound transcripts were not found to be significantly shorter than those of the unbound transcripts in Mitosis (Supplemental Figures S2-S3).

In the G1 cell cycle phase, 19 RBPs were found to have significant effects on nuclear RNA half-life (Supplemental Figure S4) and 7 RBPs on cytosolic RNA half-life (Supplemental Figure S6). None of these RBPs seems to promote nuclear RNA export. Instead, 2 RBPs seem to support cytosolic RNA turnover. One of these proteins is again PUM2 (Supplemental Figure S5), which also affects to affect the corresponding estimates in Mitosis (compare Figure 28). Additionally, IGF2BP1 emerged as another RBP to affect cytosolic RNA half-life in the G1 cell cycle phase (Figure 29).
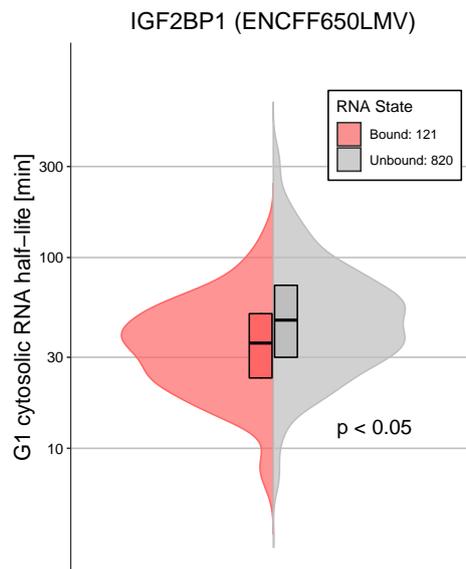
**Figure 29: Differences in G1 Cytosolic RNA Half-life upon binding by IGF2BP1.** The violin plot shows the distributions of the G1 cytosolic half-life of transcripts that are either bound (red) or unbound (gray) by IGF2BP1. A Wilcoxon rank-sum test was applied to test for significant differences in cytosolic RNA half-life between the bound and unbound transcripts. For this analysis, only those transcripts whose half-lives met the reliability criteria described in Data and Methods 2.2.12 were considered.

IGF2BP1 has been associated with RNA stabilization to protect a transcript from miRNA-mediated degradation [127], [128]. This function contradicts a possible shortening effect on cytosolic RNA half-life upon binding this protein. Therefore, IGF2BP1 could be a false-positive hit of this eCLIP analysis. On the one hand, this observation could be due to the comparatively high technical variability of cytosolic RNA half-life. On the other hand, it could also be due to the discrepancy in the cell lines used for the eCLIP experiments. The latter were performed with K562 cells and not with HeLa-S3 cells. The metabolism of the respective transcripts could differ between these cell types. In contrast, the finding that PUM2 was found as a putative modulator of cytosolic degradation in both cell cycle phases supports the analysis for the detection of regulatory RBPs of RNA metabolism.

As KHSRP was previously associated with regulating the stability of transcripts with AU-rich elements (ARE) in their 3'UTRs [123], the RNA half-lives of ARE transcripts defined in the database ARED-Plus [88] (Data and Methods 2.1.5) were further investigated. A comparison of ARE and non-ARE transcripts revealed significant differences regarding the nuclear RNA half-lives during Mitosis (see Figure 30).
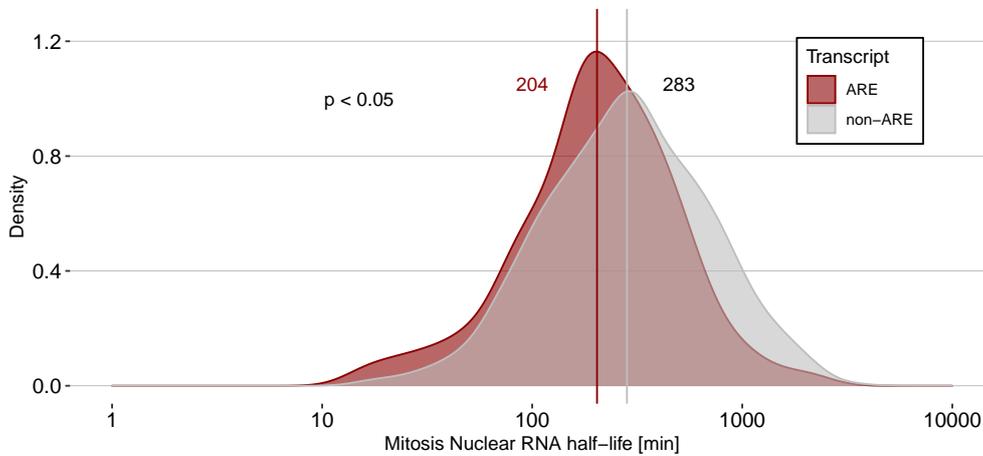
**Figure 30: Mitotic nuclear RNA Half-life Distribution of AU-rich Element Transcripts.** Genes with AU-rich element (ARE) RNAs were downloaded from the ARED-plus database [88]. The half-lives of mitotic nuclear ARE transcripts are indicated in dark red ($n = 187$), whereas all non-ARE RNAs are depicted in gray ($n = 720$). A Wilcoxon rank-sum test was applied to test for significant differences in nuclear RNA half-life between the ARE and non-ARE transcripts. Only those transcripts were considered whose nuclear RNA half-lives passed stringent reliability criteria (Data and Methods 2.2.12).

In Mitosis, the median nuclear half-life of ARE transcripts is 204 minutes, whereas the non-ARE transcripts have a median half-life of 283 minutes. A similar pattern was observed for the nuclear RNA half-lives measured in G1 (Supplemental Figure S7). These results indicates that AU-rich element transcripts spend less time in the nucleus than non-AU-rich element RNAs. They also support the role of KHSRP as a potential modulator of nuclear RNA turnover rates (compare Figure 27). Whether it stabilizes, degrades, or modulates the process of RNA export needs to be validated in future experiments. Additionally, ARE transcripts also had significantly shorter cytosolic half-lives in both the G1 and Mitosis cell cycle phase (Supplemental Figures S8-S9). However, these differences should be interpreted with caution due to the technical variability of cytosolic RNA half-lives. Nonetheless, these results demonstrate that the two-compartment model calculates RNA half-lives that capture the properties of specific RNA binding factors and sequence elements.

### 3.2.8   RNA Half-lives of Protein Subunits vary within Protein Complexes

The stoichiometry of individual subunits in protein complexes is essential to ensure the correct function of the complex [129]. A cell might control this stoichiometry by balancing

protein synthesis and degradation. Yet, the exact mechanism remains unclear [129]. One determining factor could be the overall RNA turnover rate. To check this hypothesis, the variability of RNA half-lives of individual subunits within a protein complex was analyzed using annotation data from the CORUM database [89] (Data and Methods 2.1.6; Figure 31).
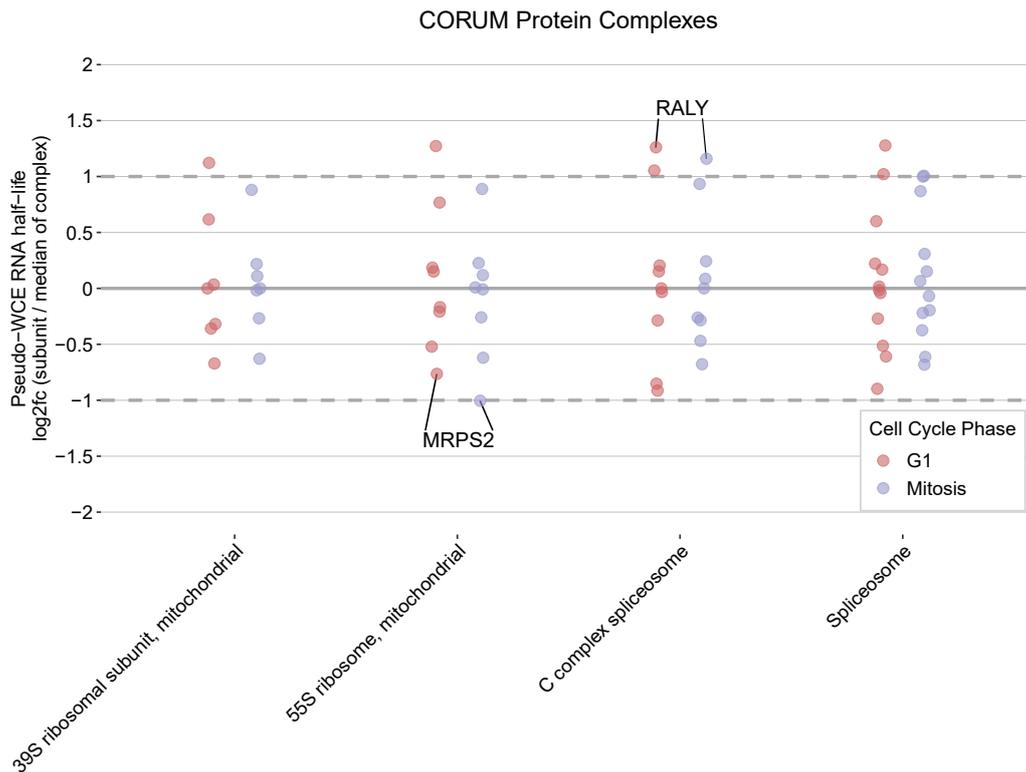


**Figure 31: Variability in RNA Half-life of Protein Complex Subunits.** The point plot depicts the log2 fold-changes between pseudo-WCE RNA half-lives of protein subunits and the corresponding median RNA half-life of the whole complex (Data and Methods 2.1.6). The CORUM database [89] was used to annotate protein complexes. Red points depict RNA half-lives measured in G1, and blue points represent RNA half-lives measured in Mitosis.

Notably, the RBP RALY of the C complex spliceosome [130] is the only protein subunit with a more than 2-times higher RNA half-life than the whole complex in both G1 and Mitosis (compare Figure 31). Besides being part of the C spliceosome, RALY was also found to be part of the Drosha complex, which is involved in the genesis of miRNAs [131], [132]. Possibly due to its regulatory role as an RBP and its role in multiple complexes, the elevated transcript half-life could reflect more stable protein expression.

At the other end of the spectrum, MRPS2, part of the small ribosomal subunit of the 55S mitochondrial ribosome, is the only protein with a more than 2 times lower RNA half-life than the whole complex, at least in Mitosis. In G1, its RNA half-life is not at least 2 times lower, but the lowest compared to the overall complex RNA half-life (compare Figure 31). Interestingly, mutations in *MRPS2* have been linked to deficiencies in the oxidative phosphorylation system of mitochondria as a consequence of impairments in the assembly of the mitochondrial small ribosomal subunit [133]. Therefore, the short RNA half-lives of MRPS2 provide further evidence that this protein might be a limiting factor in mitochondrial ribosome stability. More analyses have to be performed, which would compare, for instance, the half-lives with actual protein expression data obtained by mass-spectrometry to check this hypothesis. Nonetheless, the presented analysis provides a starting point to unravel additional determinants of protein complex stability by incorporating transcript abundances and their dynamics.

### 3.2.9   Motif Search in RNAs with slow and fast Turnover Rates

RNA sequence features determine transcript stability by, for instance, providing binding sites for RBPs [134], [135]. The high number of reliable RNA half-lives returned by the two-compartment model allowed for the grouping of transcripts from fast to slow turnover rates based on quartiles: transcripts with 1) fast turnover rates (0-25%), 2) moderately fast turnover rates (25-50%), 3) moderately slow turnover rates (50-75%), and 4) slow turnover rates (75-100%; Data and Methods 2.2.15). For each compartment and cell cycle phase, a *de novo* motif search was performed with transcripts of these four categories using HOMER [100] (analysis performed by Walter Sandt; Figure 32).
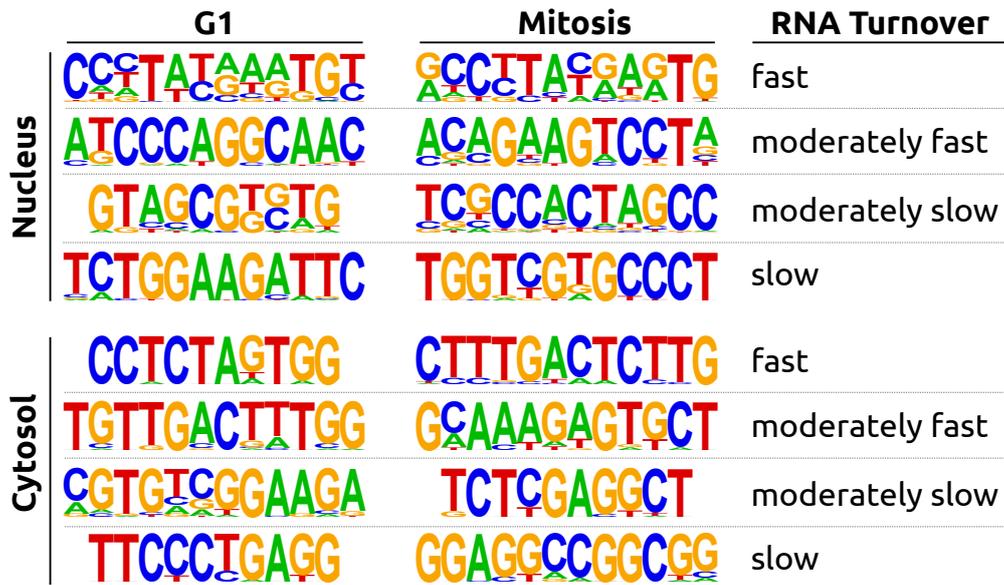
**Figure 32:** *De novo* **Motif Search on Transcripts with different turnover rates (perfomed by Walter Sandt).** The reliable estimates of the nuclear and cytosolic RNA half-lives (Data and Methods 2.2.12) were categorized into 4 different groups based on quartiles: transcripts with 1) fast, 2) moderately fast, 3) moderately slow, and 4) slow turnover rates. The analysis was performed for both the G1 and Mitosis cell cycle phases.

As shown in Figure 32, distinct *de novo* motifs, which could act as potential RBP-binding sites, were returned by HOMER. Note that this is an exploratory analysis for which these motifs must be validated experimentally. To that end, it would be interesting to target these motifs, for instance, using CRISPR, and test the dynamics of the respective transcripts for significant differences. Overall, these motifs represent potential targets for the RNA research community.

# 4 Discussion

In this work, I developed a model that resolves an mRNA's life cycle across cell compart-
ments. A novel EM algorithm that robustly estimates the 4sU incorporation rate was im-
plemented at the core of this framework. Since the $\frac{New}{Total}$ RNA ratios are very sensitive to
variation in the labeling efficiency, the accuracy of the 4sU incorporation rate estimate
is a crucial factor for successfully estimating the transcript dynamics. The presented EM
algorithm not only considers the possibility of false-positive T>C conversions caused by
sequencing errors to ensure this accuracy but also accounts for possible false-negative
C>X conversions, which could mask 4sU-labeling induced T>C conversions. Furthermore,
it performs a read pre-selection for reads whose numbers of potential labeling sites are
greater or equal to 30. This filtering step increases the probability of recognizing T>C
conversions in 4sU-labeled transcripts.

The EM's read pre-selection contrasts with the equivalent step of GRAND-SLAM, a
tool widely applied in analyzing metabolic labeling sequencing data. GRAND-SLAM pre-
selects reads based on the number of observed T>C conversions, not the potential num-
ber of labeling sites. Such a procedure requires the sequencing data to be corrected
from SNPs, RNA editing sites, and potentially other unknown sources of biases. GRAND-
SLAM performs some data cleansing steps, but the pre-selection could still be prone to
enrichment with T>C conversions originating from such error sources that escaped the
pre-processing steps. Although the same sources of error could be present in the data
used in this work, the here presented EM does not enrich for them, as it merely selects
reads based on the potential number of T>C conversions and not the observed ones. Our
modeling framework performs an even more conservative correction for SNPs and editing
sites than GRAND-SLAM. In particular, our framework masks every position with a corre-
sponding rate higher than 5%, whereas GRAND-SLAM applies a milder threshold of 50%.
Consequently, the here presented read pre-selection should have an even smaller chance
of enriching for T>C conversions, which were not induced by 4sU-labeling.

Although the presented EM algorithm outperforms GRAND-SLAM's EM algorithm in
estimating the 4sU incorporation rate, it also has its limitations. As pointed out by the

simulation, computing an accurate estimator becomes increasingly difficult if the labeling efficiency equals or falls below 1%. In such a scenario, the labeling efficiency is slightly higher than the sequencing error rates, approaching theoretical limits for the EM algorithm to differentiate 4sU labeling-induced T>C conversions from sequencing errors. This problem can be partially circumvented by increasing the depth of the sequencing libraries. Additionally, it is advisable to choose a read length that is at least 150 nucleotides since more than 30 potential labeling sites can be expected on average in the reads (assuming that each nucleotide occurs with the same frequency). Longer read lengths would even increase the number of potential labeling sites per read on average, so the threshold of at least 30 sites per read can be further elevated to gain additional confidence in the labeling efficiency estimator. However, efforts should be made to achieve a higher 4sU incorporation rate by, for instance, optimizing the amount and concentration of 4sU for the cell line of choice.

Like in our previous work [80], the analyses revealed that nuclear RNA export is slow, while cytosolic RNA degradation is comparatively fast. Consequently, an average mRNA spends most of its life in the nucleus, not the cytosol. A possible explanation for such high nuclear dwell times is that a cell could regulate protein expression by keeping the number of transcripts low in the cytosol but high in the nucleus. When more elevated amounts of a particular protein are required, the RNA can be rapidly exported from the nucleus to promote translation. Analogously, the cell can rapidly degrade a small number of cytosolic transcripts when fewer amounts of protein are required.

The mere translocation of transcripts through the NPC has been found to happen on the order of milliseconds [136], which seems to contradict the slow nuclear RNA export rates determined in this work. However, these export rates were derived from SLAM-seq data that captures poly-adenylated transcripts. Consequently, the presented nuclear RNA export rates reflect all processes that take place after polyadenylation and not solely the passage through the NPC. The long half-lives of nuclear RNA indicate that the maturation and quality control steps, such as alternative splicing and 5'-end capping, are overall long-lasting processes. The duration of these processes might be a reason why stress-induced transcripts can bypass such quality mechanisms [137].

## Discussion

The comparison of cell cycle-specific RNA half-lives with estimates from bulk measurements indicates that although some variability has been observed in the data, RNA metabolism is consistent along the cell cycle. Notably, this observation was made for both G1 and Mitosis. In the latter, some deviations of the RNA metabolism can be expected as the nuclear envelope breaks down in this phase. As a result, the nuclear and cytosolic RNA fraction could mix up. The fact that high deviations in nuclear export and cytosolic degradation were not observed in this work could imply that NEBD during mitosis does not affect the overall RNA metabolism to a great extent. One possible reason is the short duration of the breakdown, which was observed in live-cell imaging experiments to last roughly 35 minutes [118]. In comparison, the whole cell cycle of in HeLa-S3 cells was reported to be in the range of approximately 22 hours (10 h G1, 8 h S, 3 h G2, 1 h Mitosis) [138]. Since this period is only a fraction of the SLAM-seq time series, its effects may remain undetected. Moreover, possibly not all cells were perfectly synchronized, so NEBD starts with a slight offset in different cells. Therefore, the subcellular fractionation protocol might still capture a fraction of cells with intact nuclei, retaining the long half-lives of nuclear RNA.

Also, how "freely" RNAs can move through the cells remains debatable. A cell is fully packed with proteins, organelles, chromatin, and other structures, which are potential physical obstacles, especially for larger RNAs. It has been shown that mRNPs of different sizes have different diffusion parameters [139]. Depending on the used technique, diffusion parameters of different magnitudes were measured, ranging from $0.005 \frac{\mu m^2}{s}$ to $> 1 \frac{\mu m^2}{s}$ [140]. Consider a diameter of a HeLa-S3 cell of $17.1\,\mu m$ [141], which equals a radius of $8.55\,\mu m$. Then, we can calculate the travel time of the RNA from the center of the cell to the inner membrane in three-dimensional space with the formula $T = (d^2)/(6 \cdot D)$, where $d$ is the distance and $D$ the diffusion coefficient. Plugging in the values for $d = 8.55\,\mu m$ and $D = 0.005 - 1 \frac{\mu m^2}{s}$ results in diffusion times ranging from $0.2 - 41\,min$. These naive estimators indicate that certain transcripts could theoretically diffuse through the entire cell during NEBD. However, mRNAs were shown to move not only by diffusion but also by active transport via the cytoskeleton [142]–[144]. These findings lead to the question of whether RNA in the nucleus is actively transported through the cell as soon as the nuclear envelope breaks down. As there is evidence that certain

RNAs are distributed asymmetrically between mother and daughter cell [145] and some RNAs were found to be associated with chromatin during mitosis [146], it could be possible that nuclear and cytosolic RNAs do not necessarily have to mix up during Mitosis. Therefore, it will be interesting to study the fate of nuclear RNA during NEBD in live cell imaging experiments in the future. However, the presented results suggest that Mitosis does not affect bulk mRNA metabolism to a noticeable extent.

The analyses of the eCLIP experiments revealed that the putative binding of PUM2 promotes cytosolic degradation. This finding is supported by previous evidence that PUM2 regulates the degradation of transcripts by recruiting a deadenylase [125], [126]. Further, 2 proteins were identified that possibly modulate nuclear RNA export, namely RBM15 and KHSRP. Notably, RBM15 was also found in the corresponding analysis of our previous study, but KHSRP emerged as a new candidate. RBM15 has been associated with NXF1-dependent RNA export and regulation of mRNA stability [73], [121], [147], but its exact function remains unclear. KHSRP has been related to the regulation of AU-rich element transcripts stability [123]. Such ARE transcripts are less stable than non-ARE transcripts, which was also confirmed by their comparatively shorter nuclear and cytosolic RNA half-lives. However, these analyses are only correlative and not causal. For instance, the eCLIP data were acquired from WCE of K562 cells and not HeLa-S3 cells. The metabolism of certain transcripts could vary not only between these cell lines, but also the RBP function. Previously, cell line-specific differences in RBP function have been identified in the regulation of RNA decay between K562 and HepG2 cells [148]. Such effects could lead to the detection of false-positive RBPs in this eCLIP data analysis, for which further validation experiments are needed to pinpoint the roles of RBPs in RNA metabolism. Nonetheless, these findings show that the RNA half-lives computed by the two-compartment model could reflect the influence of certain sequence elements and regulatory proteins.

Collectively, the modeling framework captures critical aspects of eukaryotic mRNA metabolism with high precision. The underlying EM algorithm accounts for various sequencing errors to provide a global parameter of the 4sU incorporation rate. This labeling efficiency estimate is robust against variability from spurious, non-constant gene expression, which might bias the estimator's accuracy. I could show that this EM algorithm out-

performs the corresponding methodology of the currently widely applied tool GRAND-SLAM. As a result, the here presented framework delivers precise $\frac{New}{Total}$ RNA ratios for every measured transcript from which RNA half-lives can be derived for nuclear and cytosolic cell compartments.

# 5 Outlook

The two-compartment model is a robust framework that quantifies RNA metabolism on a genome-wide scale. In the future, this framework could be expanded to gain even higher accuracy. For instance, the 4sU incorporation into newly synthesized transcripts is modeled by a binomial distribution with a fixed labeling efficiency. It can be assumed that this rate is not necessarily constant, be it for spatial accessibility of RNA Pol II to 4sU, differences in RNA Pol II elongation speed [149], or transcriptional errors. A possible way to account for such effects is to model T>C conversions by a beta-binomial model. In this model, the labeling efficiency is not a constant probability but is sampled from a beta distribution. This way, the beta-binomial distribution captures some overdispersion in binomial-type distributed data. As SLAM-seq data is noisy, an EM algorithm based on a beta-binomial distribution is a promising model to make the 4sU incorporation rate estimation more robust and, therefore, increase its accuracy.

Furthermore, the two-compartment model could be used to study the effects of certain RBPs on RNA metabolism. For instance, an RBP involved in nuclear RNA export or degradation pathways could be knocked out. Then, the impact of the loss of this protein on RNA metabolism could be assessed. One attractive candidate for such an experiment is RBM15. As reported previously, this protein was associated with RNA export and regulation of transcript stability [73], [121], [147], but its exact role remains unclear. The two-compartment model could be fitted to two individual SLAM-seq time series obtained from control and RBM15 knock-out cell lines. Subsequently, the RBM15-associated transcripts could be tested for significant differences in nuclear export rates to elucidate whether RBM15 mainly acts as an export factor.

Since the two-compartment model captures RNA metabolism in steady state, this condition must also be fulfilled in a knock-out cell line of interest. This condition can be achieved when knocking out a non-essential gene and letting the cell line grow under optimal conditions, but not if the gene of interest is essential for cell viability. The two-compartment model needs to be extended to non-stationary conditions to study immediate cellular responses upon rapid depletion of a particular protein. To that end, the two-

compartment model could be expanded by time-dependent RNA metabolic rates. Previously, such expansions have been developed to model time-dependent RNA synthesis and degradation [53], [69]. Using the temporal and spatial information provided by subcellular fractionation combined with SLAM-seq, RNA export could also be modeled time-dependent. Such an expansion might capture the influence of biological processes, such as the bypass of RNA quality export followed by rapid export [137]. Further, it can also be used to study nuclear degradation. We assessed nuclear degradation's contribution to mRNA metabolism and found its influence to be minor [80], whereas Smalec *et al.* (2022) [76] report that a substantial amount of transcripts undergoes nuclear degradation. Since two putative regulators of RNA stability emerged in the presented work, namely RBM15 and KHSRP, performing a refined analysis of nuclear degradation will be interesting. To that end, a protein of the nuclear RNA degradation machinery such as hDIS3 [150] could be rapidly depleted using, for instance, an auxin-inducible degron fused to the respective protein. This depletion would lead to immediate changes in nuclear RNA half-life if nuclear decay plays a major role in RNA metabolism.

With technical advances in quantifying RNA metabolism, we have refined our understanding of how RNA flows through a cell. Since early quantification attempts, a simple exponential decay is used as a model to quantify RNA turnover rates. In general, exponential models assume constant degradation (or growth) rates over time and unlimited resources. When thinking of a single cell, these assumptions do not necessarily apply. For instance, the availability of specific protein subunits of the exosome could limit the capacity of RNA degradation. The same principle applies to nuclear RNA export. If a certain RBP is not available to ensure the correct protein composition of a particular RNA, it might not be exported. In general, the regulatory processes of distinct transcript subtypes could lead to notable differences in their metabolism. These examples emphasize that more flexible, alternative models for RNA turnover might be worth investigating in the future. For instance, logistic decay models, which involve a logistic function that accounts for the carrying capacity of the system in question, give an alternative to exponential decay. One could think of such a carrying capacity as the maximum transcript abundance determined by the cellular resources. A generalized version of the logistic model is the hyperbolastic model, which is an appropriate model for stem cell growth [151]. Importantly, this model

can also describe decay [152]. Unlike classic logistic models, it can not only fit sigmoidal curves but also biphasic behavior. In the case of cytosolic degradation, such behavior could reflect bursty RNA release from the nucleus, which changes the carrying capacity of cytosolic RNA. Therefore, these models represent promising alternatives that could overcome the limitations of the classic exponential decay models due to their higher flexibility.

So far, RNA metabolism has been quantified at the gene level rather than specifically for the corresponding RNA isoforms. Such transcript variants can lead to the production of protein variants with different functions. A cell can generate a diverse pool of such isoforms, for instance, by alternative splicing and the use of alternative transcription start or stop sites [153], [154]. Due to the amount of different processing steps and regulatory roles, it can also be expected that the metabolism of these transcripts differs between variants. Their quantification is challenging because of the high sequence similarity between variants and low abundance levels of specific isoforms [155], [156]. Further, short-read RNA-seq protocols (as used in SLAM-seq) are of limited use for isoform quantification due to short read lengths that cannot recover full-length transcripts. In contrast, long-read sequencing methods can recover full-length transcripts but currently suffer from high sequencing errors and do not provide as high sequencing depth as short-read sequencing methods [157]. However, this precision and depth are required to infer RNA dynamics. With advancements in long-read sequencing, combining these methods with the SLAM-seq protocol will be interesting to detect and quantify RNA isoforms and their underlying dynamics.

In conclusion, the multitude of suggestions mentioned above for improving the two-compartment model and for future studies on RNA metabolism in general shows how challenging this field of research is. I am convinced that the model presented here offers new opportunities to unravel the dynamics of the complex life cycle of eukaryotic mRNA and, therefore, will help understand how a eukaryotic cell regulates gene expression.

# References

[1] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 1970, ISSN: 1476-4687. DOI: 10.1038/227561a0.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, "From DNA to RNA," in *Molecular Biology of the Cell. 4th edition*, Garland Science, 2002.

[3] K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech, "Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena," *Cell*, vol. 31, no. 1, pp. 147–157, Nov. 1, 1982, ISSN: 0092-8674. DOI: 10.1016/0092-8674(82)90414-7.

[4] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman, "The RNA moiety of ribonuclease p is the catalytic subunit of the enzyme," *Cell*, vol. 35, no. 3, pp. 849–857, Dec. 1, 1983, ISSN: 0092-8674. DOI: 10.1016/0092-8674(83)90117-4.

[5] P. A. Sharp, "On the origin of RNA splicing and introns," *Cell*, vol. 42, no. 2, pp. 397–400, Sep. 1, 1985, ISSN: 0092-8674. DOI: 10.1016/0092-8674(85)90092-3.

[6] N. R. Pace and T. L. Marsh, "Rna catalysis and the origin of life," *Origins of life and evolution of the biosphere*, vol. 16, no. 2, pp. 97–116, Jun. 1, 1985, ISSN: 1573-0875. DOI: 10.1007/BF01809465.

[7] R. Lewin, "RNA catalysis gives fresh perspective on the origin of life," *Science*, vol. 231, no. 4738, pp. 545–546, Feb. 7, 1986. DOI: 10.1126/science.231.4738.545.

[8]     M. P. Robertson and G. F. Joyce, "The origins of the RNA world," *Cold Spring Harbor Perspectives in Biology*, vol. 4, no. 5, a003608, May 2012, ISSN: 1943-0264. DOI: `10.1101/cshperspect.a003608`.

[9]     W. Gilbert, "Origin of life: The RNA world," *Nature*, vol. 319, no. 6055, pp. 618–618, Feb. 1986, ISSN: 1476-4687. DOI: `10.1038/319618a0`.

[10]    X. Zhang, W. Wang, W. Zhu, *et al.*, "Mechanisms and functions of long non-coding RNAs at multiple regulatory levels," *International Journal of Molecular Sciences*, vol. 20, no. 22, p. 5573, Nov. 8, 2019, ISSN: 1422-0067. DOI: `10.3390/ijms20225573`.

[11]    K. Plath, S. Mlynarczyk-Evans, D. A. Nusinow, and B. Panning, "Xist RNA and the mechanism of x chromosome inactivation," *Annual Review of Genetics*, vol. 36, no. 1, pp. 233–278, 2002. DOI: `10.1146/annurev.genet.36.042902.092433`.

[12]    Z. Han and W. Li, "Enhancer RNA: What we know and what we can achieve," *Cell Proliferation*, vol. 55, no. 4, e13202, 2022, ISSN: 1365-2184. DOI: `10.1111/cpr.13202`.

[13]    P.-F. Tsai, S. Dell'Orso, J. Rodriguez, *et al.*, "A muscle-specific enhancer RNA mediates cohesin recruitment and regulates transcription in trans," *Molecular Cell*, vol. 71, no. 1, 129–141.e8, Jul. 5, 2018, ISSN: 1097-2765. DOI: `10.1016/j.molcel.2018.06.008`.

[14]    K. R. Kukurba and S. B. Montgomery, "RNA sequencing and analysis," *Cold Spring Harbor protocols*, vol. 2015, no. 11, pp. 951–969, Apr. 13, 2015, ISSN: 1940-3402. DOI: `10.1101/pdb.top084970`.

[15]    S. Sainsbury, C. Bernecky, and P. Cramer, "Structural basis of transcription initiation by RNA polymerase II," *Nature Reviews Molecular Cell Biology*,

## References

vol. 16, no. 3, pp. 129–143, Mar. 2015, ISSN: 1471-0080. DOI: `10.1038/nrm3952`.

[16] D. B. Nikolov and S. K. Burley, "RNA polymerase II transcription initiation: A structural view," *Proceedings of the National Academy of Sciences*, vol. 94, no. 1, pp. 15–22, Jan. 7, 1997. DOI: `10.1073/pnas.94.1.15`.

[17] X. Liu, D. A. Bushnell, and R. D. Kornberg, "RNA polymerase II transcription: Structure and mechanism," *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, RNA polymerase II Transcript Elongation, vol. 1829, no. 1, pp. 2–8, Jan. 1, 2013, ISSN: 1874-9399. DOI: `10.1016/j.bbagrm.2012.09.003`.

[18] J. Saba, X. Y. Chua, T. V. Mishanina, *et al.*, "The elemental mechanism of transcriptional pausing," *eLife*, vol. 8, J. L. Workman and J. L. Manley, Eds., e40981, Jan. 8, 2019, ISSN: 2050-084X. DOI: `10.7554/eLife.40981`.

[19] D. L. Bentley, "Coupling mRNA processing with transcription in time and space," *Nature reviews. Genetics*, vol. 15, no. 3, pp. 163–175, Mar. 2014, ISSN: 1471-0056. DOI: `10.1038/nrg3662`.

[20] J. Guhaniyogi and G. Brewer, "Regulation of mRNA stability in mammalian cells," *Gene*, vol. 265, no. 1, pp. 11–23, Mar. 7, 2001, ISSN: 0378-1119. DOI: `10.1016/S0378-1119(01)00350-X`.

[21] C. L. Will and R. Lührmann, "Spliceosome structure and function," *Cold Spring Harbor Perspectives in Biology*, vol. 3, no. 7, a003707, Jul. 2011, ISSN: 1943-0264. DOI: `10.1101/cshperspect.a003707`.

[22] L. Cui, R. Ma, J. Cai, *et al.*, "RNA modifications: Importance in immune cell biology and related diseases," *Signal Transduction and Targeted Therapy*, vol. 7, p. 334, Sep. 22, 2022, ISSN: 2095-9907. DOI: `10.1038/s41392-022-01175-9`.

[23]  H. Shi, J. Wei, and C. He, "Where, when, and how: Context-dependent functions of RNA methylation writers, readers, and erasers," *Molecular Cell*, vol. 74, no. 4, pp. 640–650, May 16, 2019, ISSN: 1097-2765. DOI: `10.1016/j.molcel.2019.04.025`.

[24]  O. Gabay, Y. Shoshan, E. Kopel, *et al.*, "Landscape of adenosine-to-inosine RNA recoding across human tissues," *Nature Communications*, vol. 13, no. 1, p. 1184, Mar. 4, 2022, ISSN: 2041-1723. DOI: `10.1038/s41467-022-28841-4`.

[25]  X. Jiang, B. Liu, Z. Nie, *et al.*, "The role of m6a modification in the biological functions and diseases," *Signal Transduction and Targeted Therapy*, vol. 6, no. 1, pp. 1–16, Feb. 21, 2021, ISSN: 2059-3635. DOI: `10.1038/s41392-020-00450-x`.

[26]  B. L. Bass, "RNA editing by adenosine deaminases that act on RNA," *Annual Review of Biochemistry*, vol. 71, no. 1, pp. 817–846, 2002. DOI: `10.1146/annurev.biochem.71.110601.135501`.

[27]  K. Nishikura, "A-to-i editing of coding and non-coding RNAs by ADARs," *Nature Reviews Molecular Cell Biology*, vol. 17, no. 2, pp. 83–96, Feb. 2016, ISSN: 1471-0080. DOI: `10.1038/nrm.2015.4`.

[28]  G. Singh, G. Pratt, G. W. Yeo, and M. J. Moore, "The clothes make the mRNA: Past and present trends in mRNP fashion," *Annual review of biochemistry*, vol. 84, pp. 325–354, 2015, ISSN: 0066-4154. DOI: `10.1146/annurev-biochem-080111-092106`.

[29]  T. Glisovic, J. L. Bachorik, J. Yong, and G. Dreyfuss, "RNA-binding proteins and post-transcriptional gene regulation," *FEBS letters*, vol. 582, no. 14, pp. 1977–1986, Jun. 18, 2008, ISSN: 0014-5793. DOI: `10.1016/j.febslet.2008.03.004`.

## References

[30] M. Okamura, H. Inose, and S. Masuda, "RNA export through the NPC in eukaryotes," *Genes*, vol. 6, no. 1, pp. 124–149, Mar. 20, 2015, ISSN: 2073-4425. DOI: 10.3390/genes6010124.

[31] J. C. Zinder and C. D. Lima, "Targeting RNA for processing or destruction by the eukaryotic RNA exosome and its cofactors," *Genes & Development*, vol. 31, no. 2, pp. 88–100, Jan. 15, 2017, ISSN: 0890-9369. DOI: 10.1101/gad.294769.116.

[32] A. Łabno, R. Tomecki, and A. Dziembowski, "Cytoplasmic RNA decay pathways - enzymes and mechanisms," *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, vol. 1863, no. 12, pp. 3125–3147, Dec. 1, 2016, ISSN: 0167-4889. DOI: 10.1016/j.bbamcr.2016.09.023.

[33] N. L. Garneau, J. Wilusz, and C. J. Wilusz, "The highways and byways of mRNA decay," *Nature Reviews Molecular Cell Biology*, vol. 8, no. 2, pp. 113–126, Feb. 2007, ISSN: 1471-0080. DOI: 10.1038/nrm2104.

[34] J. Coller and R. Parker, "Eukaryotic mRNA decapping," *Annual Review of Biochemistry*, vol. 73, no. 1, pp. 861–890, 2004. DOI: 10.1146/annurev.biochem.73.011303.074032.

[35] T. Saldi, M. A. Cortazar, R. M. Sheridan, and D. L. Bentley, "Coupling of RNA polymerase II transcription elongation with pre-mRNA splicing," *Journal of molecular biology*, vol. 428, no. 12, pp. 2623–2635, Jun. 19, 2016, ISSN: 0022-2836. DOI: 10.1016/j.jmb.2016.04.017.

[36] M. Brugiolo, L. Herzel, and K. M. Neugebauer, "Counting on co-transcriptional splicing," *F1000Prime Reports*, vol. 5, p. 9, Apr. 2, 2013, ISSN: 2051-7599. DOI: 10.12703/P5-9.

[37] S. McCracken, N. Fong, E. Rosonina, *et al.*, "5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal do-

main of RNA polymerase II," *Genes & Development*, vol. 11, no. 24, pp. 3306–3318, Dec. 1997, ISSN: 0890-9369.

[38]  S. McCracken, N. Fong, K. Yankulov, *et al.*, "The c-terminal domain of RNA polymerase II couples mRNA processing to transcription," *Nature*, vol. 385, no. 6614, pp. 357–361, Jan. 1997, ISSN: 1476-4687. DOI: 10.1038/385357a0.

[39]  T. Misteli and D. L. Spector, "RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo," *Molecular Cell*, vol. 3, no. 6, pp. 697–705, Jun. 1999, ISSN: 1097-2765. DOI: 10.1016/s1097-2765(01)80002-2.

[40]  Y. Hirose, R. Tacke, and J. L. Manley, "Phosphorylated RNA polymerase II stimulates pre-mRNA splicing," *Genes & Development*, vol. 13, no. 10, pp. 1234–1239, May 15, 1999, ISSN: 0890-9369, 1549-5477.

[41]  K. Sträßer, S. Masuda, P. Mason, *et al.*, "TREX is a conserved complex coupling transcription with messenger RNA export," *Nature*, vol. 417, no. 6886, pp. 304–308, May 2002, ISSN: 1476-4687. DOI: 10.1038/nature746.

[42]  H. O. Nousiainen, M. Kestilä, N. Pakkasjärvi, *et al.*, "Mutations in mRNA export mediator GLE1 result in a fetal motoneuron disease," *Nature Genetics*, vol. 40, no. 2, pp. 155–157, Feb. 2008, ISSN: 1546-1718. DOI: 10.1038/ng.2007.65.

[43]  J. A. Hurt and P. A. Silver, "mRNA nuclear export and human disease," *Disease Models & Mechanisms*, vol. 1, no. 2, pp. 103–108, Sep. 26, 2008, ISSN: 1754-8403. DOI: 10.1242/dmm.000745.

[44]  F. Gros, H. Hiatt, W. Gilbert, C. G. Kurland, R. W. Risebrough, and J. D. Watson, "Unstable ribonucleic acid revealed by pulse labelling of escherichia coli," *Nature*, vol. 190, no. 4776, pp. 581–585, May 1961, ISSN: 1476-4687. DOI: 10.1038/190581a0.

# References

[45] H. A. John, M. L. Birnstiel, and K. W. Jones, "RNA-DNA hybrids at the cytological level," *Nature*, vol. 223, no. 5206, pp. 582–587, Aug. 1969, ISSN: 1476-4687. DOI: 10.1038/223582a0.

[46] J. G. Gall and M. L. Pardue, "Formation and detection of rna-dna hybrid molecules in cytological preparations*," *Proceedings of the National Academy of Sciences*, vol. 63, no. 2, pp. 378–383, Jun. 1969. DOI: 10.1073/pnas.63.2.378.

[47] P. R. Langer-Safer, M. Levine, and D. C. Ward, "Immunological method for mapping genes on drosophila polytene chromosomes.," *Proceedings of the National Academy of Sciences*, vol. 79, no. 14, pp. 4381–4385, Jul. 1982. DOI: 10.1073/pnas.79.14.4381.

[48] R. H. Singer and D. C. Ward, "Actin gene expression visualized in chicken muscle tissue culture by using in situ hybridization with a biotinated nucleotide analog.," *Proceedings of the National Academy of Sciences*, vol. 79, no. 23, pp. 7331–7335, Dec. 1982. DOI: 10.1073/pnas.79.23.7331.

[49] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich, "Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 51 Pt 1, pp. 263–273, 1986, ISSN: 0091-7451. DOI: 10.1101/sqb.1986.051.01.032.

[50] P. Maiuri, A. Knezevich, A. De Marco, *et al.*, "Fast transcription rates of RNA polymerase II in human cells," *EMBO reports*, vol. 12, no. 12, pp. 1280–1285, Dec. 2011, ISSN: 1469-221X. DOI: 10.1038/embor.2011.196.

[51] A. Raghavan, R. L. Ogilvie, C. Reilly, *et al.*, "Genome-wide analysis of mRNA decay in resting and activated primary human t lymphocytes," *Nucleic Acids Research*, vol. 30, no. 24, pp. 5529–5538, Dec. 15, 2002, ISSN: 1362-4962. DOI: 10.1093/nar/gkf682.

[52] R. D. Alexander, J. D. Barrass, B. Dichtl, *et al.*, "RiboSys, a high-resolution, quantitative approach to measure the in vivo kinetics of pre-mRNA splicing and 3⍰-end processing in saccharomyces cerevisiae," *RNA*, vol. 16, no. 12, pp. 2570–2580, Dec. 2010, ISSN: 1355-8382. DOI: 10.1261/rna.2162610.

[53] A. Zeisel, W. J. Köstler, N. Molotski, *et al.*, "Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli," *Molecular Systems Biology*, vol. 7, p. 529, Sep. 13, 2011, ISSN: 1744-4292. DOI: 10.1038/msb.2011.62.

[54] M. Ratnadiwakara, S. K. Archer, C. I. Dent, *et al.*, "SRSF3 promotes pluripotency through nanog mRNA export and coordination of the pluripotency gene expression program," *eLife*, vol. 7, J. Valcárcel, Ed., e37419, May 9, 2018, ISSN: 2050-084X. DOI: 10.7554/eLife.37419.

[55] T. Trcek, D. R. Larson, A. Moldón, C. C. Query, and R. H. Singer, "Single-molecule mRNA decay measurements reveal promoter- regulated mRNA stability in yeast," *Cell*, vol. 147, no. 7, pp. 1484–1497, Dec. 23, 2011, ISSN: 0092-8674. DOI: 10.1016/j.cell.2011.11.051.

[56] V. Pelechano and J. E. Pérez-Ortín, "The transcriptional inhibitor thiolutin blocks mRNA degradation in yeast," *Yeast*, vol. 25, no. 2, pp. 85–92, 2008, ISSN: 1097-0061. DOI: 10.1002/yea.1548.

[57] W. T. Melvin, H. B. Milne, A. A. Slater, H. J. Allen, and H. M. Keir, "Incorporation of 6-thioguanosine and 4-thiouridine into RNA," *European Journal of Biochemistry*, vol. 92, no. 2, pp. 373–379, 1978, ISSN: 1432-1033. DOI: 10.1111/j.1432-1033.1978.tb12756.x.

[58] L. Dölken, Z. Ruzsics, B. Rädle, *et al.*, "High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and

decay," *RNA*, vol. 14, no. 9, pp. 1959–1972, Sep. 1, 2008, ISSN: 1355-8382, 1469-9001. DOI: `10.1261/rna.1136108`.

[59] M. D. Cleary, C. D. Meiering, E. Jan, R. Guymon, and J. C. Boothroyd, "Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay," *Nature Biotechnology*, vol. 23, no. 2, pp. 232–237, Feb. 2005, ISSN: 1087-0156. DOI: `10.1038/nbt1061`.

[60] C. Miller, B. Schwalb, K. Maier, *et al.*, "Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast," *Molecular Systems Biology*, vol. 7, no. 1, p. 458, Jan. 2011, ISSN: 1744-4292. DOI: `10.1038/msb.2010.112`.

[61] M. Sun, B. Schwalb, D. Schulz, *et al.*, "Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation," *Genome Research*, vol. 22, no. 7, pp. 1350–1359, Jul. 1, 2012, ISSN: 1088-9051, 1549-5469. DOI: `10.1101/gr.130161.111`.

[62] M. Rabani, J. Z. Levin, L. Fan, *et al.*, "Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells," *Nature Biotechnology*, vol. 29, no. 5, pp. 436–442, May 2011, ISSN: 1546-1696. DOI: `10.1038/nbt.1861`.

[63] M. Schueler, M. Munschauer, L. H. Gregersen, *et al.*, "Differential protein occupancy profiling of the mRNA transcriptome," *Genome Biology*, vol. 15, no. 1, R15, 2014, ISSN: 1465-6906. DOI: `10.1186/gb-2014-15-1-r15`.

[64] B. Schwalb, M. Michel, B. Zacher, *et al.*, "TT-seq maps the human transient transcriptome," *Science (New York, N.Y.)*, vol. 352, no. 6290, pp. 1225–1228, Jun. 3, 2016, ISSN: 1095-9203. DOI: `10.1126/science.aad9841`.

[65] L. Windhager, T. Bonfert, K. Burger, *et al.*, "Ultrashort and progressive 4su-tagging reveals key characteristics of RNA processing at nucleotide resolution," *Genome Research*, vol. 22, no. 10, pp. 2031–2042, Oct. 2012, ISSN: 1549-5469. DOI: 10.1101/gr.131847.111.

[66] B. Schwanhäusser, D. Busse, N. Li, *et al.*, "Global quantification of mammalian gene expression control," *Nature*, vol. 473, no. 7347, pp. 337–342, May 2011, ISSN: 1476-4687. DOI: 10.1038/nature10098.

[67] E. E. Duffy, M. Rutenberg-Schoenberg, C. D. Stark, R. R. Kitchen, M. B. Gerstein, and M. D. Simon, "Tracking distinct RNA populations using efficient and reversible covalent chemistry," *Molecular Cell*, vol. 59, no. 5, pp. 858–866, Sep. 3, 2015, ISSN: 1097-2765. DOI: 10.1016/j.molcel.2015.07.023.

[68] V. A. Herzog, B. Reichholf, T. Neumann, *et al.*, "Thiol-linked alkylation of RNA to assess expression dynamics," *Nature Methods*, vol. 14, no. 12, pp. 1198–1204, Dec. 2017, ISSN: 1548-7105. DOI: 10.1038/nmeth.4435.

[69] S. de Pretis, T. Kress, M. J. Morelli, *et al.*, "INSPEcT: A computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4su-seq time course experiments," *Bioinformatics*, vol. 31, no. 17, pp. 2829–2835, Sep. 1, 2015, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv288.

[70] K. Burger, B. Mühl, M. Kellner, *et al.*, "4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response," *RNA biology*, vol. 10, no. 10, pp. 1623–1630, Oct. 2013, ISSN: 1555-8584. DOI: 10.4161/rna.26214.

[71] T. Neumann, V. A. Herzog, M. Muhar, *et al.*, "Quantification of experimentally induced nucleotide conversions in high-throughput sequencing datasets," *BMC Bioinformatics*, vol. 20, no. 1, p. 258, May 20, 2019, ISSN: 1471-2105. DOI: 10.1186/s12859-019-2849-7.

## References

[72]  C. Jürges, L. Dölken, and F. Erhard, "Dissecting newly transcribed and old RNA using GRAND-SLAM," *Bioinformatics (Oxford, England)*, vol. 34, no. 13, pp. i218–i226, Jul. 1, 2018, ISSN: 1367-4811. DOI: `10 . 1093 / bioinformatics/bty256`.

[73]  B. Zuckerman, M. Ron, M. Mikl, E. Segal, and I. Ulitsky, "Gene architecture and sequence composition underpin selective dependency of nuclear export of long RNAs on NXF1 and the TREX complex," *Molecular Cell*, vol. 79, no. 2, 251–267.e6, Jul. 16, 2020, ISSN: 1097-2765. DOI: `10.1016/j.molcel. 2020.05.013`.

[74]  Y. Finkel, A. Gluck, A. Nachshon, *et al.*, "SARS-CoV-2 uses a multipronged strategy to impede host protein synthesis," *Nature*, vol. 594, no. 7862, pp. 240–245, Jun. 2021, ISSN: 1476-4687. DOI: `10 . 1038 / s41586 – 021 – 03610–3`.

[75]  T. Fisher, A. Gluck, K. Narayanan, *et al.*, "Parsing the role of NSP1 in SARS-CoV-2 infection," *Cell Reports*, vol. 39, no. 11, Jun. 14, 2022, ISSN: 2211-1247. DOI: `10.1016/j.celrep.2022.110954`.

[76]  B. M. Smalec, R. Ietswaart, K. Choquet, E. McShane, E. R. West, and L. S. Churchman, *Genome-wide quantification of RNA flow across subcellular compartments reveals determinants of the mammalian transcript life cycle*, Aug. 21, 2022. DOI: `10.1101/2022.08.21.504696`.

[77]  I. Vukovic, S. Baranda, J. W. Ruffin, J. Karlin, and S. B. McMahon, *Non-redundant roles for the human mRNA decapping cofactor paralogs DCP1a and DCP1b*, Oct. 19, 2023. DOI: `10.1101/2023.10.19.563039`.

[78]  L. Fishman, G. Nechooshtan, F. Erhard, A. Regev, J. A. Farrell, and M. Rabani, *Single-cell temporal dynamics reveals the relative contributions of transcrip-*

*tion and degradation to cell-type specific gene expression in zebrafish embryos*, Apr. 21, 2023. DOI: `10.1101/2023.04.20.537620`.

[79]  A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977, ISSN: 0035-9246.

[80]  J. M. Müller, K. Moos, T. Baar, K. Zumer, and A. Tresch, *Nuclear export is a limiting factor in eukaryotic mRNA metabolism*, May 4, 2023. DOI: `10.1101/2023.05.04.539375`.

[81]  P. Moll, M. Ante, A. Seitz, and T. Reda, "QuantSeq 3' mRNA sequencing for RNA quantification," en, *Nature Methods*, vol. 11, no. 12, pp. i–iii, Dec. 2014, Number: 12 Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: `10.1038/nmeth.f.376`.

[82]  T. Nojima, T. Gomes, M. Carmo-Fonseca, and N. J. Proudfoot, "Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide," *Nature Protocols*, vol. 11, no. 3, pp. 413–428, Mar. 2016, ISSN: 1750-2799. DOI: `10.1038/nprot.2016.012`.

[83]  D. Karolchik, A. S. Hinrichs, T. S. Furey, *et al.*, "The UCSC table browser data retrieval tool," *Nucleic Acids Research*, vol. 32, pp. D493–D496, suppl_1 Jan. 1, 2004, ISSN: 0305-1048. DOI: `10.1093/nar/gkh103`.

[84]  E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, *et al.*, "Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)," *Nature Methods*, vol. 13, no. 6, pp. 508–514, Jun. 2016, ISSN: 1548-7105. DOI: `10.1038/nmeth.3810`.

# References

[85] Y. Luo, B. C. Hitz, I. Gabdank, *et al.*, "New developments on the encyclopedia of DNA elements (ENCODE) data portal," *Nucleic Acids Research*, vol. 48, p. D882, D1 Jan. 1, 2020. DOI: `10.1093/nar/gkz1062`.

[86] A. Frankish, M. Diekhans, A.-M. Ferreira, *et al.*, "GENCODE reference annotation for the human and mouse genomes," *Nucleic Acids Research*, vol. 47, pp. D766–D773, D1 Jan. 8, 2019, ISSN: 0305-1048. DOI: `10.1093/nar/gky955`.

[87] C.-Y. A. Chen and A.-B. Shyu, "AU-rich elements: Characterization and importance in mRNA degradation," *Trends in Biochemical Sciences*, vol. 20, no. 11, pp. 465–470, Nov. 1, 1995, ISSN: 0968-0004. DOI: `10.1016/S0968-0004(00)89102-1`.

[88] T. Bakheet, E. Hitti, and K. S. A. Khabar, "ARED-plus: An updated and expanded database of AU-rich element-containing mRNAs and pre-mRNAs," *Nucleic Acids Research*, vol. 46, pp. D218–D220, D1 Jan. 4, 2018, ISSN: 0305-1048. DOI: `10.1093/nar/gkx975`.

[89] G. Tsitsiridis, R. Steinkamp, M. Giurgiu, *et al.*, "CORUM: The comprehensive resource of mammalian protein complexes–2022," *Nucleic Acids Research*, vol. 51, pp. D539–D545, D1 Jan. 6, 2023, ISSN: 0305-1048. DOI: `10.1093/nar/gkac1015`.

[90] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-b: Fortran subroutines for large-scale bound-constrained optimization," *ACM Transactions on Mathematical Software*, vol. 23, no. 4, pp. 550–560, 1997, ISSN: 0098-3500. DOI: `10.1145/279232.279236`.

[91] "Babraham bioinformatics - FastQC a quality control tool for high throughput sequence data." (), [Online]. Available: `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/` (visited on 12/13/2023).

[92] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, no. 1, pp. 10–12, May 2, 2011, ISSN: 2226-6089. DOI: `10.14806/ej.17.1.200`.

[93] F. J. Sedlazeck, P. Rescheneder, and A. von Haeseler, "NextGenMap: Fast and accurate read mapping in highly polymorphic genomes," *Bioinformatics*, vol. 29, no. 21, pp. 2790–2791, Nov. 1, 2013, ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btt468`.

[94] S. T. Sherry, M.-H. Ward, M. Kholodov, *et al.*, "dbSNP: The NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, Jan. 1, 2001, ISSN: 0305-1048.

[95] P. Bromiley and N. Thacker, "The effects of an arcsin square root transform on a binomial distributed quantity.," Jan. 1, 2002.

[96] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, Apr. 1, 1970, ISSN: 0006-3444. DOI: `10.1093/biomet/57.1.97`.

[97] K. M. Mullen, D. Ardia, D. L. Gil, D. Windover, and J. Cline, "DEoptim: An r package for global optimization by differential evolution," *Journal of Statistical Software*, vol. 40, pp. 1–26, Apr. 7, 2011, ISSN: 1548-7660. DOI: `10.18637/jss.v040.i06`.

[98] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, Jan. 1, 1965, ISSN: 0010-4620. DOI: `10.1093/comjnl/7.4.308`.

[99] F. J. Martin, M. R. Amode, A. Aneja, *et al.*, "Ensembl 2023," *Nucleic Acids Research*, vol. 51, pp. D933–D941, D1 Jan. 6, 2023, ISSN: 0305-1048. DOI: `10.1093/nar/gkac958`.

## References

[100]   S. Heinz, C. Benner, N. Spann, *et al.*, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities," *Molecular Cell*, vol. 38, no. 4, pp. 576–589, May 28, 2010, ISSN: 1097-4164. DOI: 10.1016/j.molcel.2010.05.004.

[101]   G. van Rossum and J. de Boer, "Interactively testing remote servers using the python programming language," *CWI Quarterly*, vol. 4, no. 4, pp. 283–304, Dec. 1, 1991, ISSN: 0922-5366.

[102]   The pandas development team, *Pandas-dev/pandas: Pandas*, version 1.4.3, Feb. 2020. DOI: 10.5281/zenodo.3509134.

[103]   W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[104]   P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "SciPy 1.0: Fundamental algorithms for scientific computing in python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2.

[105]   C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, ISSN: 1476-4687. DOI: 10.1038/s41586-020-2649-2.

[106]   P. J. A. Cock, T. Antao, J. T. Chang, *et al.*, "Biopython: Freely available python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, Jun. 1, 2009, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp163.

[107]   R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020.

[108]   H. Li, B. Handsaker, A. Wysoker, *et al.*, "The sequence alignment/map format and SAMtools," *Bioinformatics (Oxford, England)*, vol. 25, no. 16, pp. 2078–2079, Aug. 15, 2009, ISSN: 1367-4811. DOI: `10 . 1093 / bioinformatics/btp352`.

[109]   L. T. Johnson and C. J. Geyer, "Variable transformation to obtain geometric ergodicity in the random-walk metropolis algorithm," *The Annals of Statistics*, vol. 40, no. 6, pp. 3050–3076, Dec. 2012, ISSN: 0090-5364, 2168-8966. DOI: `10.1214/12-AOS1048`.

[110]   M. Corporation and S. Weston, *Doparallel: Foreach parallel adaptor for the 'parallel' package*, R package version 1.0.17, 2022.

[111]   J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, May 2007, ISSN: 1558-366X. DOI: `10. 1109/MCSE.2007.55`.

[112]   M. L. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, Apr. 6, 2021, ISSN: 2475-9066. DOI: `10.21105/joss.03021`.

[113]   H. Wickham, "Reshaping data with the reshape package," *Journal of Statistical Software*, vol. 21, pp. 1–20, Nov. 13, 2007, ISSN: 1548-7660. DOI: `10.18637/jss.v021.i12`.

[114]   H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016, ISBN: 978-3-319-24277-4.

[115]   A. Kassambara, *Ggpubr: 'ggplot2' based publication ready plots*, R package version 0.6.0, 2023.

[116]   B. Auguie, *Gridextra: Miscellaneous functions for "grid" graphics*, R package version 2.3, 2017.

## References

[117] N. Stoler and A. Nekrutenko, "Sequencing error profiles of illumina sequencing instruments," *NAR Genomics and Bioinformatics*, vol. 3, no. 1, lqab019, Mar. 1, 2021, ISSN: 2631-9268. DOI: `10.1093/nargab/lqab019`.

[118] N. Belaadi, L. Pernet, J. Aureille, *et al.*, "SUN2 regulates mitotic duration in response to extracellular matrix rigidity," *Proceedings of the National Academy of Sciences*, vol. 119, no. 45, e2116167119, Nov. 8, 2022. DOI: `10.1073/pnas.2116167119`.

[119] M. W. Hentze, A. Castello, T. Schwarzl, and T. Preiss, "A brave new world of RNA-binding proteins," *Nature Reviews Molecular Cell Biology*, vol. 19, no. 5, pp. 327–341, May 2018, ISSN: 1471-0080. DOI: `10.1038/nrm.2017.130`.

[120] W. Wende, P. Friedhoff, and K. Sträßer, "Mechanism and regulation of co-transcriptional mRNP assembly and nuclear mRNA export," in *The Biology of mRNA: Structure and Function*, ser. Advances in Experimental Medicine and Biology, M. Oeffinger and D. Zenklusen, Eds., Cham: Springer International Publishing, 2019, pp. 1–31, ISBN: 9783030314347. DOI: `10.1007/978-3-030-31434-7_1`.

[121] S. Lindtner, A. S. Zolotukhin, H. Uranishi, *et al.*, "RNA-binding motif protein 15 binds to the RNA transport element RTE and provides a direct link to the NXF1 export pathway *," *Journal of Biological Chemistry*, vol. 281, no. 48, pp. 36 915–36 928, Dec. 1, 2006, ISSN: 0021-9258, 1083-351X. DOI: `10.1074/jbc.M608745200`.

[122] H. Min, C. W. Turck, J. M. Nikolic, and D. L. Black, "A new regulatory protein, KSRP, mediates exon inclusion through an intronic splicing enhancer," *Genes & Development*, vol. 11, no. 8, pp. 1023–1036, Apr. 15, 1997, ISSN: 0890-9369. DOI: `10.1101/gad.11.8.1023`.

[123] R. Gherzi, K.-Y. Lee, P. Briata, *et al.*, "A KH domain RNA binding protein, KSRP, promotes ARE-directed mRNA turnover by recruiting the degradation machinery," *Molecular Cell*, vol. 14, no. 5, pp. 571–583, Jun. 4, 2004, ISSN: 1097-2765. DOI: 10.1016/j.molcel.2004.05.002.

[124] The UniProt Consortium, "UniProt: The universal protein knowledgebase in 2023," *Nucleic Acids Research*, vol. 51, pp. D523–D531, D1 Jan. 6, 2023, ISSN: 0305-1048. DOI: 10.1093/nar/gkac1052.

[125] G. Lu and T. M. T. Hall, "Alternate modes of cognate RNA recognition by human PUMILIO proteins," *Structure (London, England: 1993)*, vol. 19, no. 3, pp. 361–367, Mar. 9, 2011, ISSN: 1878-4186. DOI: 10.1016/j.str.2010.12.019.

[126] J. Van Etten, T. L. Schagat, J. Hrit, *et al.*, "Human pumilio proteins recruit multiple deadenylases to efficiently repress messenger RNAs," *The Journal of Biological Chemistry*, vol. 287, no. 43, pp. 36 370–36 383, Oct. 19, 2012, ISSN: 1083-351X. DOI: 10.1074/jbc.M112.373522.

[127] S. Zhu, J.-Z. Wang, D. Chen, *et al.*, "An oncopeptide regulates m6a recognition by the m6a reader IGF2bp1 and tumorigenesis," *Nature Communications*, vol. 11, no. 1, p. 1685, Apr. 3, 2020, ISSN: 2041-1723. DOI: 10.1038/s41467-020-15403-9.

[128] H. Huang, H. Weng, W. Sun, *et al.*, "Recognition of RNA n6-methyladenosine by IGF2bp proteins enhances mRNA stability and translation," *Nature Cell Biology*, vol. 20, no. 3, pp. 285–295, Mar. 2018, ISSN: 1476-4679. DOI: 10.1038/s41556-018-0045-z.

[129] J. C. Taggart, H. Zauber, M. Selbach, G.-W. Li, and E. McShane, "Keeping the proportions of protein complex components in check," *Cell Systems*,

## References

vol. 10, no. 2, pp. 125–132, Feb. 26, 2020, ISSN: 2405-4712. DOI: `10.1016/j.cels.2020.01.004`.

[130]  P. Coltri, K. Effenberger, R. J. Chalkley, A. L. Burlingame, and M. S. Jurica, "Breaking up the c complex spliceosome shows stable association of proteins with the lariat intron intermediate," *PLoS ONE*, vol. 6, no. 4, e19061, Apr. 19, 2011, ISSN: 1932-6203. DOI: `10.1371/journal.pone.0019061`.

[131]  R. I. Gregory, K.-p. Yan, G. Amuthan, *et al.*, "The microprocessor complex mediates the genesis of microRNAs," *Nature*, vol. 432, no. 7014, pp. 235–240, Nov. 2004, ISSN: 1476-4687. DOI: `10.1038/nature03120`.

[132]  L. Sun, A. Wan, Z. Zhou, *et al.*, "RNA-binding protein RALY reprogrammes mitochondrial metabolism via mediating miRNA processing in colorectal cancer," *Gut*, vol. 70, no. 9, pp. 1698–1712, Sep. 1, 2021, ISSN: 0017-5749, 1468-3288. DOI: `10.1136/gutjnl-2020-320652`.

[133]  T. Gardeitchik, M. Mohamed, B. Ruzzenente, *et al.*, "Bi-allelic mutations in the mitochondrial ribosomal protein MRPS2 cause sensorineural hearing loss, hypoglycemia, and multiple OXPHOS complex deficiencies," *The American Journal of Human Genetics*, vol. 102, no. 4, pp. 685–695, Apr. 5, 2018, ISSN: 0002-9297, 1537-6605. DOI: `10.1016/j.ajhg.2018.02.012`.

[134]  J. Cheng, K. C. Maier, Ž. Avsec, P. Rus, and J. Gagneur, "Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast," *RNA*, vol. 23, no. 11, pp. 1648–1659, Nov. 1, 2017, ISSN: 1355-8382, 1469-9001. DOI: `10.1261/rna.062224.117`.

[135]  R. Shalgi, M. Lapidot, R. Shamir, and Y. Pilpel, "A catalog of stability-associated sequence elements in 3' UTRs of yeast mRNAs," *Genome Biology*, vol. 6, no. 10, R86, Sep. 30, 2005, ISSN: 1474-760X. DOI: `10.1186/gb-2005-6-10-r86`.

[136]   D. Grünwald and R. H. Singer, "In vivo imaging of labelled endoge-
        nous β-actin mRNA during nucleocytoplasmic transport," *Nature*, vol. 467,
        no. 7315, pp. 604–607, Sep. 30, 2010, ISSN: 0028-0836. DOI: `10 . 1038 /
        nature09438`.

[137]   G. Zander, A. Hackmann, L. Bender, *et al.*, "mRNA quality control is by-
        passed for immediate export of stress-responsive transcripts," *Nature*,
        vol. 540, no. 7634, pp. 593–596, Dec. 2016, ISSN: 1476-4687. DOI: `10.1038/
        nature20572`.

[138]   M. Deplanche, R. A. E.-A. Filho, L. Alekseeva, *et al.*, "Phenol-soluble mod-
        ulin α induces g2/m phase transition delay in eukaryotic HeLa cells," *The
        FASEB Journal*, vol. 29, no. 5, pp. 1950–1959, 2015, ISSN: 1530-6860. DOI:
        `10.1096/fj.14-260513`.

[139]   J. Sheinberger and Y. Shav-Tal, "The dynamic pathway of nuclear RNA in
        eukaryotes," *Nucleus*, vol. 4, no. 3, pp. 195–205, May 1, 2013, ISSN: 1949-
        1034. DOI: `10.4161/nucl.24434`.

[140]   M. Oeffinger and D. Zenklusen, "To the pore and through the pore: A story
        of mRNA export kinetics," *Biochimica et Biophysica Acta (BBA) - Gene Regu-
        latory Mechanisms*, Nuclear Transport and RNA Processing, vol. 1819, no. 6,
        pp. 494–506, Jun. 1, 2012, ISSN: 1874-9399. DOI: `10.1016/j.bbagrm.2012.
        02.011`.

[141]   A. B. Borle, "Kinetic analyses of calcium movements in HeLa cell cultures
        : I. calcium influx," *Journal of General Physiology*, vol. 53, no. 1, pp. 43–56,
        Jan. 1, 1969, ISSN: 0022-1295. DOI: `10.1085/jgp.53.1.43`.

[142]   Y. Ben-Ari, Y. Brody, N. Kinor, *et al.*, "The life of an mRNA in space and time,"
        *Journal of Cell Science*, vol. 123, no. 10, pp. 1761–1774, May 15, 2010, ISSN:
        0021-9533. DOI: `10.1242/jcs.062638`.

## References

[143]  D. Fusco, N. Accornero, B. Lavoie, *et al.*, "Single mRNA molecules demonstrate probabilistic movement in living mammalian cells," *Current Biology*, vol. 13, no. 2, pp. 161–167, Jan. 21, 2003, ISSN: 0960-9822. DOI: `10.1016/S0960-9822(02)01436-7`.

[144]  A. W. Lifland, C. Zurla, J. Yu, and P. J. Santangelo, "Dynamics of native β-actin mRNA transport in the cytoplasm," *Traffic*, vol. 12, no. 8, pp. 1000–1011, 2011, ISSN: 1600-0854. DOI: `10.1111/j.1600-0854.2011.01209.x`.

[145]  Y. Shlyakhtina, K. L. Moran, and M. M. Portal, "Asymmetric inheritance of cell fate determinants: Focus on RNA," *Non-Coding RNA*, vol. 5, no. 2, p. 38, May 9, 2019, ISSN: 2311-553X. DOI: `10.3390/ncrna5020038`.

[146]  W. Shen, Y. Zhang, M. Shi, *et al.*, "Profiling and characterization of constitutive chromatin-enriched RNAs," *iScience*, vol. 25, no. 11, p. 105 349, Nov. 18, 2022, ISSN: 2589-0042. DOI: `10.1016/j.isci.2022.105349`.

[147]  L.-M. Appel, V. Franke, J. Benedum, *et al.*, "The SPOC domain is a phosphoserine binding module that bridges transcription machinery with co- and post-transcriptional regulators," *Nature Communications*, vol. 14, no. 1, p. 166, Jan. 11, 2023, ISSN: 2041-1723. DOI: `10.1038/s41467-023-35853-1`.

[148]  J. Lin and Z. Ouyang, "Large-scale analysis of the position-dependent binding and regulation of human RNA binding proteins," *Quantitative Biology*, vol. 8, no. 2, pp. 119–129, 2020, ISSN: 2095-4697. DOI: `10.1007/s40484-020-0206-5`.

[149]  L. Muniz, E. Nicolas, and D. Trouche, "RNA polymerase II speed: A key player in controlling and adapting transcriptome composition," *The EMBO Journal*, vol. 40, no. 15, e105740, Aug. 2, 2021, ISSN: 0261-4189. DOI: `10.15252/embj.2020105740`.

[150]   R. Tomecki, M. S. Kristiansen, S. Lykke-Andersen, *et al.*, "The human core exosome interacts with differentially localized processive RNases: hDIS3 and hDIS3L," *The EMBO Journal*, vol. 29, no. 14, pp. 2342–2357, Jul. 2010, Num Pages: 2357 Publisher: John Wiley & Sons, Ltd, ISSN: 0261-4189. DOI: `10.1038/emboj.2010.121`.

[151]   M. A. Tabatabai, Z. Bursac, W. M. Eby, and K. P. Singh, "Mathematical modeling of stem cell proliferation," *Medical & Biological Engineering & Computing*, vol. 49, no. 3, pp. 253–262, Mar. 1, 2011, ISSN: 1741-0444. DOI: `10.1007/s11517-010-0686-y`.

[152]   M. Tabatabai, D. K. Williams, and Z. Bursac, "Hyperbolastic growth models: Theory and application," *Theoretical Biology and Medical Modelling*, vol. 2, no. 1, p. 14, Mar. 30, 2005, ISSN: 1742-4682. DOI: `10.1186/1742-4682-2-14`.

[153]   T. W. Nilsen and B. R. Graveley, "Expansion of the eukaryotic proteome by alternative splicing," *Nature*, vol. 463, no. 7280, pp. 457–463, Jan. 2010, ISSN: 1476-4687. DOI: `10.1038/nature08909`.

[154]   A. Reyes and W. Huber, "Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues," *Nucleic Acids Research*, vol. 46, no. 2, pp. 582–592, Jan. 25, 2018, ISSN: 0305-1048. DOI: `10.1093/nar/gkx1165`.

[155]   J. Westoby, M. S. Herrera, A. C. Ferguson-Smith, and M. Hemberg, "Simulation-based benchmarking of isoform quantification in single-cell RNA-seq," *Genome Biology*, vol. 19, no. 1, p. 191, Nov. 7, 2018, ISSN: 1474-760X. DOI: `10.1186/s13059-018-1571-5`.

[156]   J. Westoby, P. Artemov, M. Hemberg, and A. Ferguson-Smith, "Obstacles to detecting isoforms using full-length scRNA-seq data," *Genome Biology*,

vol. 21, no. 1, p. 74, Mar. 23, 2020, ISSN: 1474-760X. DOI: 10.1186/s13059-020-01981-w.

[157]  Y. Hu, L. Fang, X. Chen, J. F. Zhong, M. Li, and K. Wang, "LIQA: Long-read isoform quantification and analysis," *Genome Biology*, vol. 22, no. 1, p. 182, Jun. 17, 2021, ISSN: 1474-760X. DOI: 10.1186/s13059-021-02399-8.

# Supplements
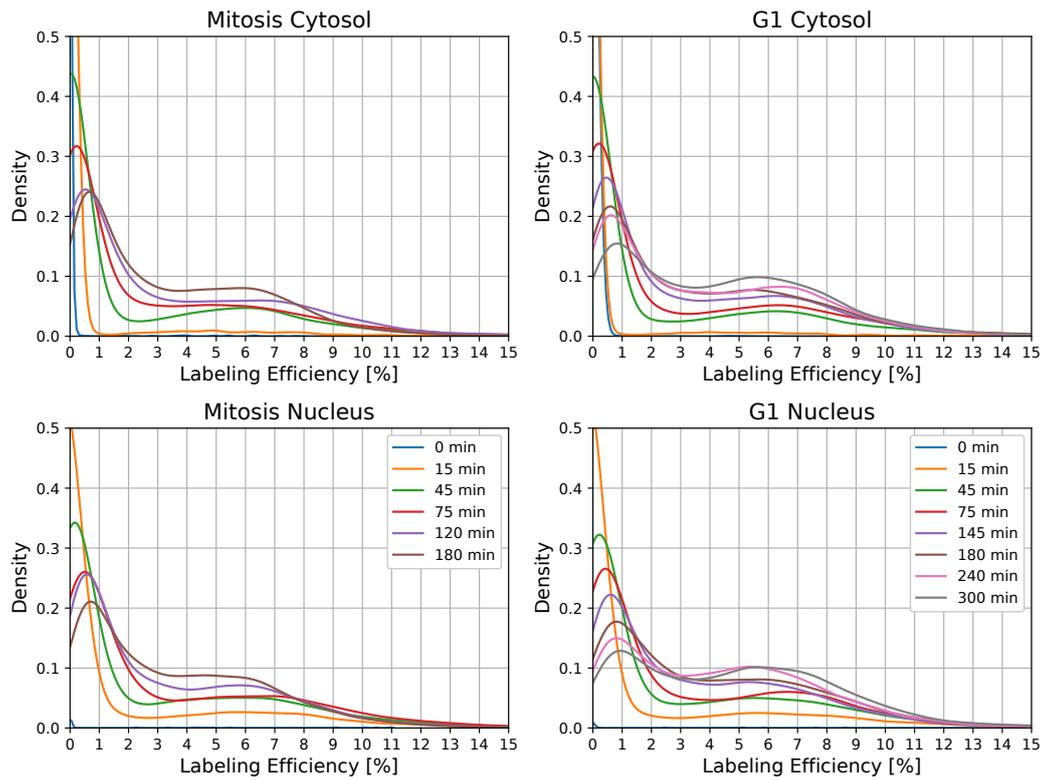


**Supplemental Figure S1: Labeling Efficiency Estimates of Cell Cycle Data obtained by the first Implementation of our Modeling Framework [80].** The density plots show the distributions of the labeling efficiency estimates computed for each time point (see color) of the nuclear and cytosolic compartments from the G1 and Mitosis cell cycle phases.
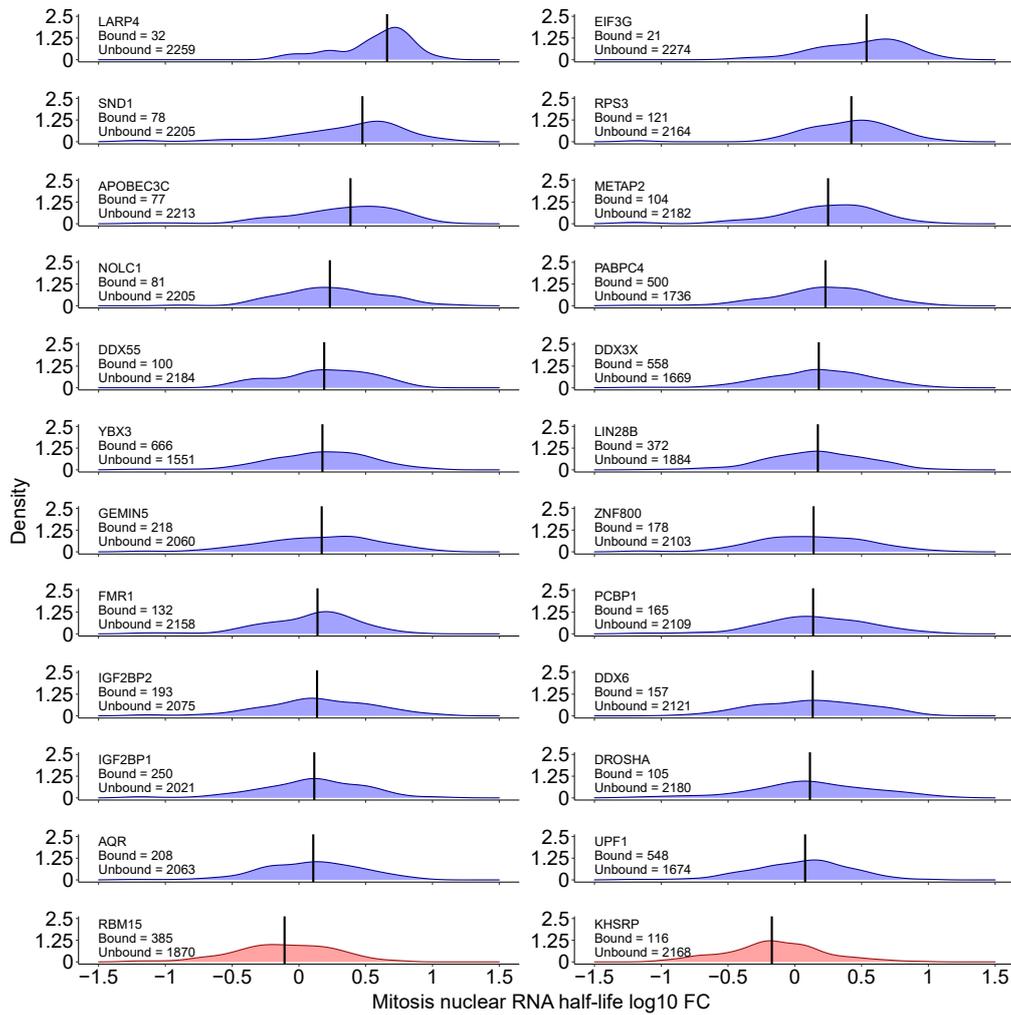
**Supplemental Figure S2: Density Plots for distinct RBPs with significant Effects on nuclear RNA Half-life in Mitosis.** The density plots depict the distributions of the mitotic nuclear half-lives from RNAs putatively bound to a specific RBP. Log10 fold-changes were calculated by taking the log10 of the nuclear RNA half-lives and subtracting the log10 of the global median (the median of all bound and unbound RNA half-lives) from these values. Blue density plots indicate distributions with a higher median RNA half-life of the bound transcripts as the global median. Red densities indicate a shorter median RNA half-life of the bound transcripts compared to the global median. The black bars indicate the median values of the respective distributions.

**Supplemental Figure S3: Density Plots for distinct RBPs with significant Effects on cytosolic RNA Half-life in Mitosis.** The density plots depict the distributions of the mitotic cytosolic half-lives from RNAs putatively bound to a specific RBP. Log10 fold-changes were calculated by taking the log10 of the cytosolic RNA half-lives and subtracting the log10 of the global median (the median of all bound and unbound RNA half-lives) from these values. Blue density plots indicate distributions with a higher median RNA half-life of the bound transcripts as the global median. Red densities indicate a shorter median RNA half-life of the bound transcripts compared to the global median. The black bars indicate the median values of the respective distributions.

**Supplemental Figure S4: Density Plots for distinct RBPs with significant Effects on nuclear RNA Half-life in G1.** The density plots depict the distributions of the G1 nuclear half-lives from RNAs putatively bound to a specific RBP. Log10 fold-changes were calculated by taking the log10 of the nuclear RNA half-lives and subtracting the log10 of the global median (the median of all bound and unbound RNA half-lives) from these values. Blue density plots indicate distributions with a higher median RNA half-life of the bound transcripts as the global median. Red densities indicate a shorter median RNA half-life of the bound transcripts compared to the global median. The black bars indicate the median values of the respective distributions.
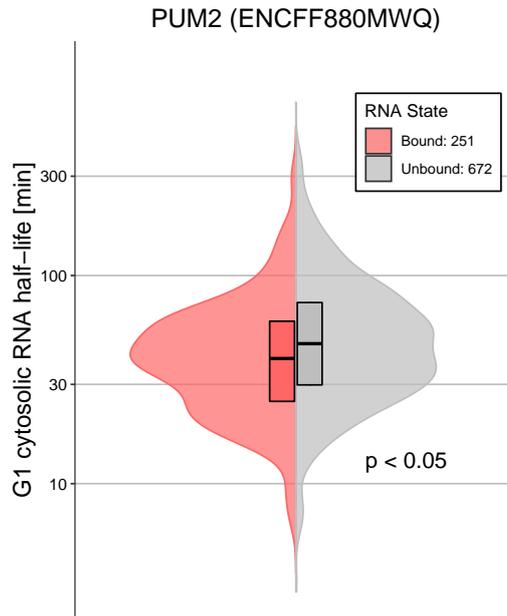
**Supplemental Figure S5: Differences in G1 cytosolic RNA Half-life upon binding by PUM2.** The Violin plot shows the distributions of the cytosolic RNA half-life of transcripts that are bound (red) or unbound (gray) by PUM2. A Wilcoxon rank-sum test was applied to test for significant differences in cytosolic RNA half-life between the bound and unbound transcripts. For this analysis, only those transcripts whose half-lives met the reliability criteria described in Methods 2.2.12 were considered.
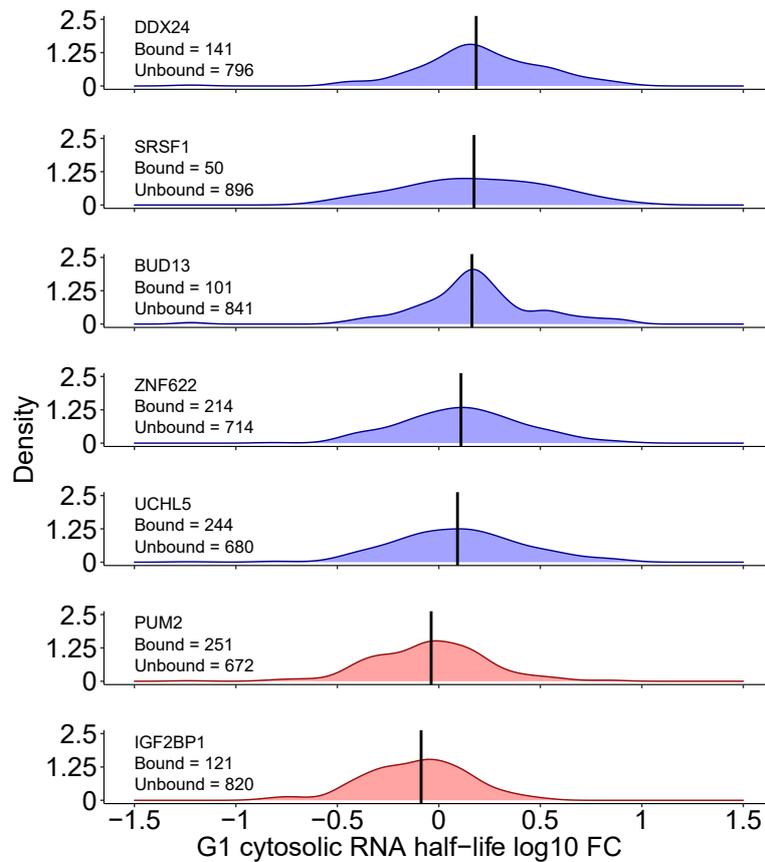
**Supplemental Figure S6: Density Plots for distinct RBPs with significant Effects on cytosolic RNA Half-life in G1.** The density plots depict the distributions of the G1 cytosolic half-lives from RNAs putatively bound to a specific RBP. Log10 fold-changes were calculated by taking the log10 of the cytosolic RNA half-lives and subtracting the log10 of the global median (the median of all bound and unbound RNA half-lives) from these values. Blue density plots indicate distributions with a higher median RNA half-life of the bound transcripts as the global median. Red densities indicate a shorter median RNA half-life of the bound transcripts compared to the global median. The black bars indicate the median values of the respective distributions.

**Supplemental Figure S7: G1 nuclear RNA Half-life Distribution of AU-rich Element Transcripts.**
Genes with AU-rich element (ARE) RNAs were downloaded from the ARED-plus database [88]. The
half-lives of G1 nuclear ARE transcripts are indicated in dark red ($n = 220$), whereas all non-ARE
RNAs are colored in gray ($n = 736$). A Wilcoxon rank-sum test was applied to test for significant dif-
ferences in nuclear RNA half-life between the ARE and non-ARE transcripts. Only those transcripts
whose nuclear RNA half-lives passed stringent reliability criteria were considered (Data and Meth-
ods 2.2.12).
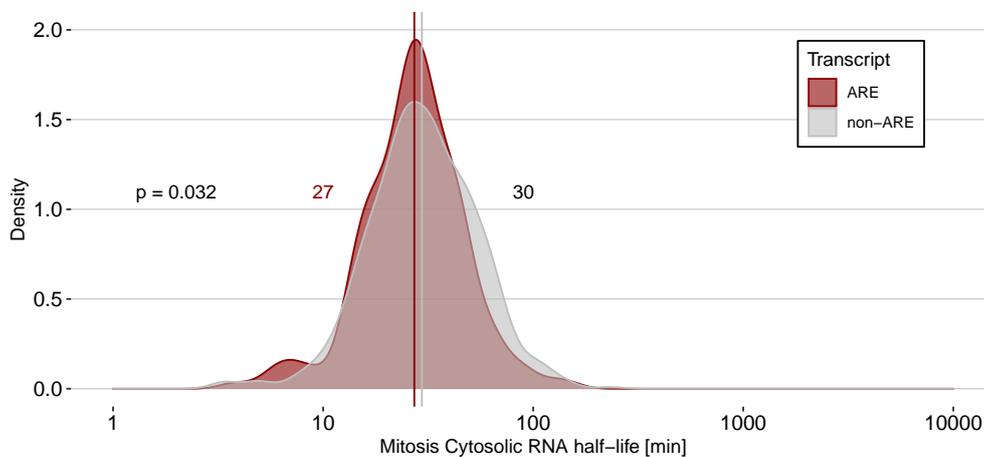


**Supplemental Figure S8: Mitotic cytosolic RNA Half-life Distribution of AU-rich Element Tran-
scripts.** Genes with AU-rich element (ARE) RNAs were downloaded from the ARED-plus database
[88]. The half-lives of mitotic cytosolic ARE transcripts are indicated in dark red ($n = 187$), whereas
all non-ARE RNAs are depicted gray ($n = 720$). A Wilcoxon rank-sum test was applied to test for
significant differences in cytosolic RNA half-life between the ARE and non-ARE transcripts. For this
analysis, only those transcripts whose cytosolic RNA half-lives passed stringent reliability criteria
were considered (Data and Methods 2.2.12).
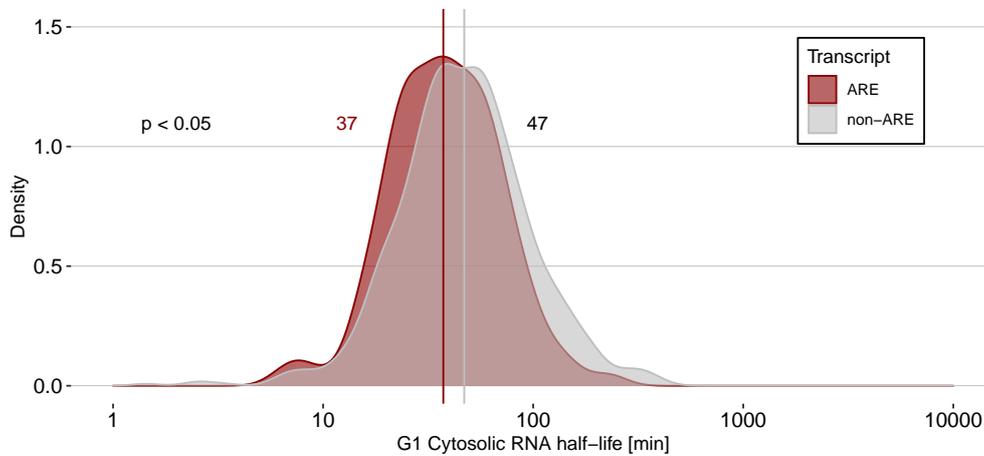
**Supplemental Figure S9: G1 cytosolic RNA Half-life Distribution of AU-rich Element Transcripts.** Genes with AU-rich element (ARE) RNAs were downloaded from the ARED-plus database [88]. The half-lives of G1 cytosolic ARE transcripts are depicted in dark red ($n = 220$), whereas all non-ARE RNAs are shown in gray ($n = 736$). A Wilcoxon rank-sum test was applied to test for significant differences in cytosolic RNA half-life between the ARE and non-ARE transcripts. Only those transcripts whose cytosolic RNA half-lives passed stringent reliability criteria were considered (Data and Methods 2.2.12).

# Data Availability Statement

The SLAM-seq cell cycle datasets were exclusively generated for this work by Elisabeth Altendorfer and Susanne Freier from the Mayer Lab at the MPI for Molecular Genetics (Berlin). The dataset is available upon request from Achim Tresch. For correspondence, please contact achim.tresch@uni-koeln.de.

# Acknowledgements

At the end of this work, I would like to express my immense gratitude to all the people who have accompanied me during my PhD studies and my entire life.

First and foremost, I thank my supervisor, Prof. Dr. Achim Tresch, for enabling me to work on this great research project. It was an instructive and exciting journey, with all kinds of ups and downs - and I enjoyed every day I could spend on it. Achim, thank you for always supporting me. You have encouraged me to explore different research paths and provided me with the opportunity to make it known in the RNA research community. I do not take this for granted. Thank you for your loyalty and your trust in me. It has been an honor for me to learn from you.

I would also like to thank Prof. Dr. Andreas Beyer, not only as my second supervisor of this thesis but also as a former supervisor of my Master's thesis and as a member of my thesis advisory committee. The research in your group sparked my interest in algorithmic bioinformatics and method development in general.

Further, I would like to thank Prof. Dr. Stanislav Kopriva for paving the path of my early scientific career. Even though I was an early bachelor's student without any laboratory experience, you welcomed me into your lab, which enabled me to get my first insights into biological research. Also, you encouraged me to follow the path of quantitative biology, which was the best career decision I made so far. It is a pleasure to have you on my dissertation defense committee.

I thank the minutes taker of my dissertation defense committee, Dr. Till Baar, for being an awesome office companion and system administrator.

Special thanks are due to my experimental collaborators who were involved in this work: Dr. Elisabeth Altendorfer, Susanne Freier, and Dr. Andreas Mayer. I have always enjoyed our fruitful discussions about this project.

Next, I would like to express my gratitude to Dr. Katharina Moos. When I joined this project, you always had an open door for asking questions, explaining details, giving me confidence, or just chatting. You did a great job, Katha.

Also, I would like to thank Prof. Dr. Kay Hofmann for giving me the opportunity to write my bachelor's thesis in the field of computational biology. I learned a lot during this period. Also, I always had a great time in your group.

I thank Dr. Niklas Kleinenkuhnen for being an excellent colleague and minister of joy in our group. It is funny that we happened to become colleagues during my bachelor's degree and subsequent doctoral studies. Let us see where we will end up together in the future.

I would also like to thank all the people of the Tresch research group I worked with during my studies: Karin, Zahra, Thanina, Vlada, Mohammad, Sophia, and Walter. I always enjoyed spending time with you during work but also at private gatherings where we philosophized about science and the world. I also especially thank Walter for being a great student assistant.

Furthermore, I would like to thank all members of the Cologne Graduate School of Ageing Research (CGA) and the RRZK for providing a great research infrastructure for my studies.

My heart-warm gratitude goes to all my great friends who accompanied me since day 1. Here, I especially thank Jasmin Engel, Marius Hermann, Ricardo Guerreiro, Fiona Edenhofer, Florian Orbach, Manuel Lentzen, Leonie Koroll, Niklas Hielscher, Dennis Gadalla, and Tobias Brünger for their loyalty and support in every situation.

Zu guter Letzt möchte ich mich bei meiner Familie bedanken. Allen voran bei meinen Eltern Martina und Erwin Müller, sowie meiner Schwester Sophie Müller. Ohne Eure bedingungslose Unterstützung hätte ich diese Zeilen niemals schreiben können. Danke, dass Ihr mich durch alle Lebenslagen begleitet und ich mich immer auf Euch verlassen kann. Ihr mögt das alles für selbstverständlich halten, aber ich weiß, dass es das nicht ist.

# Erklärung zur Dissertation

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

_____     _____

Ort, Datum                          Name, Unterschrift