A Dynamic Model of mRNA Metabolism

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von:

Katharina Moos

aus Freiburg im Breisgau

26.02.2022

Erstgutachter: Prof. Dr. Achim Tresch

Zweitgutachter: Prof. Dr. Andreas Beyer

Datum der mündlichen Prüfung: 27.04.2022

Acknowledgement

There are many people who accompanied me during this project, and even more who accompanied me throughout my life. Thank you for being a part of this :) this acknowledgement goes to all of you.

First of all, I want to thank Achim Tresch for giving me the opportunity to do my Ph.D. in his group, for learning a lot during this time, and for being a kind and empathic group leader who created a happy group atmosphere. I warmly thank Kristina Zumer and Patrick Cramer for the good collaboration and providing this project with great data and biological knowledge.

I warmly thank my Ph.D. group, the AG Tresch:

Till Baar for his highly competent company in our joint teaching sessions and for being a fabulous game master for our group's Pen & Paper adventures,
Rafael Campos-Martin and Zahra Sadat Hajseyed Nasrollah for the exciting scientific discussions and teaching each other Spanish, Persian and German words,
Lihao Chen and Julia Sasse for the funny board game evenings,
Arijit Das for inviting the group over and cooking Indian dinner for all of us at each year's Diwali,
Sebastian Dümke for his scientific advice,
Mohammad Hussainy for sharing his special Syrian coffee with everyone,
Josefine Joisten for the good conversations,
Niklas Kleinenkuhnen for being an amazing minister of fun & entertainment and organizing our team events, and for watering my plants when I was on holidays,
Vlada Milchevskaya for showing us the nicest coffee place near the office and our 'Today-is-not-Wednesday' excursions in summer,
Jason Müller for his kind help in the running project and taking over for the future, I am happy that this project found you as a successor,
Karin Sablonsky and Ursula Höhne for keeping the group alive, for always having the overview of everything and for being the persons one could rely on whenever there was an issue,
Sophia Schmickler for her company in our joint office room,
Achim Tresch for inviting the group over to his yearly carnival parties,
and everyone for being very supportive and taking care of each other. Thank you for the amazing time I had with you :)

I warmly thank:

Those of my school teachers, and especially Frank Müller, who have always been motivated and supportive towards their students,
Ralf Erens and Thomas Schonhardt for spending their free time to give pupils a deeper insight into natural sciences in the Freiburg Seminar,
Florian Raible for hosting me in a practical course to gain insight about how research is done in a wet lab,
Johannes Normann as well as the kind ladies from the examination office for always having a friendly ear for the biology students,
Stefan Rensing for helping me to prepare for my bioinformatics master and for always finding time whenever I needed help,
Wolfgang Maier and Ekkehard Schulze for supervising me during my bachelor thesis, for teaching me Python and telling me stories about the history of computer science,
Maximilian Ulbrich and Nicole Gensch for spending their free time to guide and support our iGEM team,
Björn Voss for hosting and supervising me in his group for my master thesis,
All the kind people I collaborated with in other projects during my Ph.D. time: Jan Becker, Thorsten Buch, Pietro Cicalese, Brandon Ginley, Brendon Lutnick, Hien Nguyen, Pinaki Sarder;
The AG Börries, my new research group in Freiburg, who warmly welcomed me and made me feel like I have always been part of the team from the first day on: Melanie Börries, Miriam von Scheibner, Geoffroy Andrieux, Geritt Batt, Andreas Blaumeiser, Eyleen Corrales, Antoine Devism, Severin Dicks, Elham Bavafaye Haghighi, Sylvia Herter, Maria Hess, Anselm Hoppmann, Georg Kohnke, Silke Kowar, Ariane Lehmann, Nikolai Lontke, Ralf Mertes, Patrick Metzger, Franziska Nehlert, Thomas Pauli, Senthilkumar Ramamoorthy, Aránzazu Sáenz, Vincent Schipperges, Frederik Voigt, Ella Levit Zerdoun.

I warmly thank all of my friends for sharing the common time, the funny moments, the silly photos, and many amazing memories. Thank you for being part :):

Viktor Andersson; The Beachvolleyball Crew from Cologne, especially Ulf Böse, Bärbel Gröne and Daniel Kopf; Dennis Blum & Laura Niemeyer, Rafael Campos-Martin, Marcus Degener, Daniel Desiro, Wesley Dobreske, Nabeel Farhan, Prateek Garg, Laura Geigele & Ciaran Steger-Hoey, Yiren Richard Hu, Benjamin Kaiser & Agnes Matysiak, Nils Kickert, Urs & Annette Lange, Christina Laudenbach, Andreas Lazzaro; The Meltdown People from Cologne, especially Theo-Renato Attila Akin, Debora Golditz, Marco Kasongo, Nico Kröger, Katharina Peters, Florian Rathmer, Björn Ritke and Tobias Urhahne; Mahmoud Ali Mohammad, Zahra Sadat Hajseyed Nasrollah, Patrick Niemann, Dariel 'Jake' Rosales, Jonathan Schöck, Denis Schulte, my Volleyball team 'The Tigers' from Cologne, Jake Upwards, Yessica & Andreas Waßmer, Robin Wouters, Fabian Ziegler.

Also, I want to thank my laptop for holding on until I finished the writing of this thesis.

Finally, my special thanks goes to my parents Bettina & Dieter Moos. They enabled me to find out about my interests, to choose whichever way I wanted to take in my life, always had my back and supported me in everything I was doing. My parents take these things for granted, but I know they are not.

I dedicate this thesis to my parents :).

Zusammenfassung

RNA ist eine der Schlüsselkomponenten im zentralen Dogma der Biologie. Die protein-kodierenden mRNA Transkripte werden im Zellkern synthetisiert, zum fertigen Transkript prozessiert, ins Zellplasma exportiert, hier als Vorlage für die Proteinsynthese genutzt, und schließlich wieder abgebaut. Genexpression kann durch die Anpassung von beispielsweise mRNA Synthese- oder Degradationsraten gesteuert werden. Mit Hilfe eines sogenannten ‚Metabolic Labeling' Experiments (metabolisches Markierungsexperiment) können mRNA Stoffwechselraten gemessen werden: Neu synthetisierte RNA wird mit modifizierten Nukleosiden markiert und kann dadurch von RNA unterschieden werden, die bereits vor Beginn des Experiments vorhanden war. Die Auswertung der Daten solcher Markierungsexperimente ist jedoch komplex. Nicht alle neu synthetisierten RNA Transkripte werden markiert, da nur ein gewisser Prozentsatz der nativen Nukleoside durch die modifizierten Nukleoside ersetzt wird. Des Weiteren ermöglicht unseres Wissens nach keine der aktuell verfügbaren Analysemethoden die Erforschung von mRNA Export. Um diese Problemstellungen zu bewältigen habe ich dynamisches Modell des mRNA Stoffwechsels entwickelt. Das Modell unterscheidet zwischen dem Zellkern und Zellplasma (Zwei-Kompartiment Modell) und ermöglicht dadurch die Analyse von RNA Export zusätzlich zu RNA Degradation. Die Datenpräprozessierung ist speziell an metabolische Markierungsexperimente angepasst und Markierungseffizienzen werden in Abhängigkeit der Markierungsdauer berechnet. Wir führten ein RNA-Markierungsexperiment an HeLa-S3 Zellen durch und kombinierten es mit zellulärer Fraktionierung, um nukleäre und zytosolische RNA getrennt zu messen. Dann berechneten wir nukleäre und zytosolische RNA Halbwertszeiten unter Anwendung des Modells. Unsere Ergebnisse zeigen, dass die nukleären Halbwertszeiten länger als die zytosolische Halbwertszeiten sind, woraus zu schließen ist, dass nukleärer Export langsamer ist als zytosolische RNA Degradation. Folglich verbringen mRNA Transkripte die meiste ihrer Lebenszeit im Zellkern, und sind im Vergleich zum Zellplasma im Zellkern angereichert. Wir entdecken eine Gruppe herausragender Gene, die wir ‚Supernova' Gene nennen, welche sich durch einen außergewöhnlich schnellen RNA Export auszeichnen und dadurch auf einen neuartigen, distinkten Exportmechanismus hinweisen.

Abstract

RNA is one of the key components in the central dogma of biology. The protein-coding mRNAs are synthesised in the nucleus, processed to form the mature mRNA molecule, exported to the cytosol where they are translated into the functional protein, and eventually degraded. Gene expression can be regulated by adjusting, for example, mRNA synthesis or degradation rates. These mRNA metabolic rates are measured with RNA metabolic labelling experiments: Newly synthesised RNAs are tagged with modified nucleosides and can thereby be distinguished from pre-existing RNA transcripts. Yet, the analysis of metabolic labelling data is statistically challenging. The tagging of newly synthesised RNA transcripts is incomplete as only a fraction of the native nucleosides is replaced by the modified nucleoside. Also, to our knowledge, no currently available analysis tool provides a framework to investigate mRNA export. To address these challenges, I developed a dynamic model of mRNA metabolism. The model distinguishes between the nuclear and cytosolic compartment (two-compartment model) and thereby allows to analyse nuclear export in addition to cytosolic degradation. The data pre-processing steps are specifically tailored to metabolic labelling data and include the estimation of RNA labelling efficiencies in a time-dependent manner. We performed a metabolic labelling experiment on HeLa-S3 cells combined with cellular fractionation to measure nuclear and cytosolic RNA separately. We then applied my model to estimate nuclear and cytosolic RNA half lives. We find that nuclear half lives are much higher than cytosolic half lives, which leads to the conclusion that mRNA export is slower than mRNA degradation. Consequently, mRNA transcripts spend most of their lifetime in the nucleus and are more abundant in the nucleus than in the cytosol. We discover a group of outstanding genes, called ‘Supernova’ genes, which show an exceptionally fast mRNA export, arguing for a novel and distinct export mechanism.

Table of Contents

1 Introduction

RNA is one of the key components in the central dogma of biology, which describes the flow of information between DNA, RNA and proteins (Crick, 1970). RNA is transcribed from a DNA template, and protein is translated from the RNA molecule. RNA transcripts that encode for protein sequences are commonly called messenger RNAs (mRNAs). Yet, not all DNA templates and respectively RNA transcripts are protein-coding. A large number of non-coding RNA species like transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), long non-coding RNAs (lncRNAs) or microRNAs have been described which carry out metabolic or regulatory functions (Cabili et al., 2015; Eulalio et al., 2009; Persson et al., 2009; Eddy, 1999).

The metabolism of mRNA within eukaryotic cells can be summarised in five steps: (1) mRNA is synthesised in the nucleus, yielding a preliminary pre-mRNA molecule. (2) The pre-mRNA molecule is processed to form the mature mRNA transcript. (3) Incorrectly processed transcripts are degraded in the nucleus, (4) successfully processed transcripts are exported to the cytosol, serving as templates for protein synthesis. (5) Finally, cytosolic mRNA is degraded. Eukaryotes share four RNA polymerases (Pol) which are responsible for the synthesis of the different RNA species (Bunch et al., 2016; Werner, Thuriaux, & Soutourina, 2009; Dieci et al., 2007; Arnold et al., 2012). mRNAs and lncRNAs are transcribed by Pol II, pre-rRNAs are synthesised by Pol I, tRNAs and the 5S rRNA by Pol III, and transcripts encoded on the mitochondrial DNA by the mitochondrial RNA polymerase. microRNAs and other non-coding RNAs are transcribed by both Pol II and III.

The pre-mRNA transcript undergoes several maturation steps, including 5’ end-capping, intron splicing, 3’ end cleavage and polyadenylation (Bentley, 2014). Multiple quality control mechanisms in the nucleus ensure the integrity of the processed mRNA and eventually degrade transcripts that are incompletely or incorrectly processed (Jiao et al., 2013; Hackmann et al., 2014; Galy et al., 2004; Coyle et al., 2011; Soheilypour & Mofrad, 2018). The mature mRNA transcript is then ready to be exported to the cytosol. Export is assisted by multiple RNA binding proteins (RBPs) which align to the mRNA transcript either during or after transcription and form a messenger ribonucleoprotein complex (mRNP) (Sträßer et al., 2002; Masuda et al., 2005; Viphakone et al., 2012; Brennan, Gallouzi, & Steitz, 2000; Topisirovic et al., 2009; Okamura, Inose, & Masuda, 2015; Delaleau & Borden, 2015). Several adaptor proteins within

the mRNP particle are required to either enhance or induce the binding affinity to the export receptors that enable mRNP transfer through the nuclear pore complex (NPC) (Viphakone et al., 2012; Brennan et al, 2000; Topisirovic et al., 2009; Okamura et al., 2015; Delaleau & Borden, 2015).

In mammalian cells, mRNP export is mainly mediated via the two export receptors Nfx1/Tap and Crm1/Xpo1 (Herold, Klymenko, & Izaurralde, 2001; Katahira et al., 2015; Brennan et al., 2000; Watanabe et al., 1999; Okamura et al., 2015; Delaleau & Borden, 2015). Nfx1/Tap and its main adaptor protein complex TREX are conserved from humans to yeast and required for bulk mRNA export (Sträßer et al., 2002; Viphakone et al., 2012; Katahira et al., 1999; Herold et al., 2001; Katahira et al., 2015). In contrast, Crm1/Xpo1 exports a subset of specific mRNAs (Okamura et al., 2015; Delaleau & Borden, 2015).

Within the cytosol, mRNA degradation is usually initialised by deadenylation of the polyA tail. Transcripts are subsequently either degraded from the 3' end or, more commonly, decapped and degraded in 5' to 3' direction (Chen et al., 2009; Wu & Belasco, 2006; Yamashita et al., 2005; Decker & Parker, 1993; Wu & Brewer, 2012). As an alternative, the endoribonucleolytic decay pathway functions by mRNA molecule cleavage and is independent of deadenylation (Wu & Belasco, 2006; Meister et al., 2004; Wu & Brewer, 2012). Degradation does not only serve to regulate gene expression, but also to remove transcripts with premature stop codons (nonsense-mediated decay and ribosome extension–mediated decay), missing stop codons (nonstop decay) or stalled ribosomes (no-go decay) (Kedde et al., 2010; Kong & Liebhaber, 2007; Yamashita et al., 2005; Wu & Brewer, 2012).

Decay can be promoted by the non-coding microRNAs (miRNAs) and small-interfering RNAs (siRNAs) (Chen et al., 2009; Jing et al., 2005; Kedde et al., 2010; Meister et al., 2004; Wu & Brewer, 2012). One single-stranded, non-coding RNA molecule and several proteins are assembled to the so-called RNA-induced silencing complex (RISC). RISC then binds to mRNA transcripts complementary to the incorporated non-coding RNA (Chendrimada et al., 2005; Meister et al., 2004; Wu & Brewer, 2012). Imperfect complementary base-pairing triggers deadenylation-dependent degradation, whereas perfect complementary binding leads to endoribonucleolytic cleavage of the target mRNA (Chen et al., 2009; Wu & Belasco, 2006; Wu & Brewer, 2012).

The single mRNA metabolic steps do not act independently but are highly interconnected. RNA polymerase II does not only execute transcription, but is also required for 5' end-capping,

splicing and the initiation of 3' end polyadenylation (Ho & Shuman, 1999; Hirose & Manley, 1998; Rosonina et al., 2003; McCracken et al., 1997). mRNA splicing occurs co-transcriptionally to a large extend (Oesterreich, Preibisch, & Neugebauer, 2010; Khodor et al., 2011; Girard et al., 2012) and can be influenced by promotor architecture and the usage of transcriptional activators (Cramer et al., 1997; Rosonina et al., 2003). Polyadenylation is linked to many proteins that play a role in different steps of mRNA metabolism, like transcriptional activators, splicing factors, and the export-relevant THO complex (Rosonina et al., 2003; Nagaike et al., 2011; Danckwardt et al., 2007; Saguez et al., 2008). Rpb4p, an RNA polymerase II subunit, was found to be involved in both mRNA transcription and export under cellular stress conditions, but also in cytosolic degradation (Maillet et al., 1999; Shalem et al., 2011; Farago et al., 2003; Lotan et al., 2005). Several proteins that function in cytosolic decay were shown to travel back and forth between the cytosol and the nucleus and stimulate transcription (Haimovich et al., 2013). The interplay between transcription and cytosolic decay buffers mRNA levels: Impairing transcription leads to a reduction in degradation, which stabilizes transcripts, and vice versa (Haimovich et al., 2013; Sun et al., 2012).

Furthermore, mRNA metabolic processes are not executed identically for most mRNAs, but carried out differently for specific groups of mRNA transcripts. This enables the regulation of gene expression and gene function at the level of mRNA metabolism, which thereby crucially contributes to a cell's ability to carry out specific biological processes. Different mRNA transcript isoforms produced as a result of differential pre-mRNA processing (Cramer et al., 1997; Danckwardt et al., 2007) can alter the stability of the mRNA transcript itself and thereby influence gene expression (Boutet et al., 2012), impact the mRNA's subcellular localisation (An et al., 2008), or lead to changes in the translated protein product (Muriel et al., 2005; Lu, Gladden, & Diehl, 2003; Solomon et al., 2003; Takagaki et al., 1996). Alternative transcript isoform usage was shown to depict tissue-specificity (Lianoglou et al., 2013) and to be involved in cell differentiation and morphology (Boutz et al., 2007; Takagaki & Manley, 1998; Ji & Tian, 2009; An et al., 2008; Boutet et al., 2012). Distinct mRNA export pathways can transfer specific subgroups of mRNA transcripts (Wickramasinghe et al., 2013; Culjkovic et al., 2006; Okamura et al., 2015), and various biological processes like cell cycle progression, cell differentiation, stress response and immune response are regulated via mRNA export or export-related proteins (Bretes et al., 2014; Wang et al., 2013; Carney et al., 2009; Chakraborty et al., 2008; Faria et al., 2006; Okamura et al., 2015). Notably, it was observed in yeast that mRNA transcripts of

heat-stress associated proteins, when transcribed upon heat shock, skip nuclear quality control (Zander et al., 2016).

Dysregulation of mRNA metabolism can cause a wide variety of diseases. Malfunction in mRNA processing and mRNA export was shown to be associated with neurodegenerative diseases and cancer in several studies (Mayr & Bartel, 2009; Sun et al., 2015; Shiga et al., 2012; Domínguez-Sánchez et al., 2011; Saito et al., 2013; Guo et al., 2005; Chen et al., 2005; Wang et al., 2005; Adamia et al., 2005) and causative for genetically inherited diseases like the lethal congenital contracture syndrome, myotonic dystrophy, and Osteogenesis Imperfecta Type I (Folkmann et al., 2013; Nousiainen et al., 2008; Lin et al., 2006; Holt et al., 2007; Smith et al., 2007; Johnson, 2000).

Multiple methodologies have been developed to investigate the dynamics of mRNA metabolism. Single-molecule imaging techniques were applied to count mRNA or mRNP particles in the nuclear and cytosolic compartment and track their movement within the nucleus. Fluorescence in situ hybridisation (FISH) has been one strategy to label single mRNA transcripts, determine their abundance, and study the effects of transcriptional bursting (Raj et al., 2006; Mor et al., 2010). Tagging single mRNA particles with fluorescent proteins via the MS2 system served to investigate mRNP movement patterns within the nucleus and to measure the time until an mRNP is translocated to the cytosol (Mor et al., 2010; Shav-Tal et al., 2004). The rates of mRNA synthesis, splicing and decay and their dynamic changes in response to external stimuli have often been studied by mRNA transcription inhibition and subsequent mRNA quantification (Zeisel, 2011; Raghavan et al., 2002; Shalem et al., 2008; Slobodin et al., 2020).

Yet, the above-described have a couple of drawbacks and limitations. Single-molecule imaging techniques are restricted in the number of mRNA or mRNP particles per cell to resolve the visual signals successfully. This makes it impossible to study mRNA on a transcriptome-wide scale. Furthermore, these experiments often use artificial gene constructs to implement transcriptional control via the chosen promoter, to enable probing of the mRNA transcript, and to verify translation in the cytosol (for example, by fusing the sequence information a fluorescent protein). This renders the observations incomparable to the behaviour of native molecules. Experimental approaches that use transcriptional arrest to investigate mRNA metabolic rates may be biased due to the interplay between transcription and decay in order to buffer mRNA levels (Pelechano & Pérez-Ortín, 2008).

An alternative perspective on measuring mRNA metabolism with substantially less perturbation of the metabolism itself was delivered by the principle of RNA metabolic labelling. This approach uses nucleoside analogues which deviate from native nucleosides but are still incorporated into newly synthesised RNA strands during transcription (Cleary et al., 2005). Thereby, newly synthesised RNA is tagged and distinguished from pre-existing RNA. The nucleoside analogue 4-thiouridine (4sU) found broad application in the investigation of mRNA metabolism (Miller et al., 2011; Sun et al., 2012; Eser et al., 2014; Amorim et al., 2010; Rabani et al., 2014; Schwanhäusser et al., 2011; Schueler et al., 2014). 4sU residues integrated into RNA strands can be biotinylated and then pulled down with streptavidin beads, thereby separating the labelled from the unlabelled RNA fraction. RNA was then quantified with the help of, for example, microarrays.

Recently, the 4sU labelling strategy was further improved by Herzog et al. (2017). They exploit the fact that 4sU residues are not interpreted as uridines but cytidines when RNA is converted to cDNA by the reverse transcriptase for followed-up sequencing. Consequently, 4sU labelled RNA transcripts contain T>C conversions in their sequence alignments, distinguishing them from unlabelled RNA transcripts. Herzog et al. alkylate 4sU residues with iodoacetamide to increase the T>C conversion efficiency from around 10-11% to more than 94%, making this approach feasible for labelled RNA detection. Accordingly, they named their method SLAM seq ("thiol-linked alkylation for the metabolic sequencing of RNA"). SLAM seq is less laborious than the above described 4sU labelling procedure, and the data acquired is less noisy.

The invention of new experimental procedures entailed the development of new mathematical tools to analyse the data and enable the determination of mRNA metabolic parameters. Many studies calculated global mRNA decay rates based on whole-cell RNA data, assuming that metabolism is in steady state (metabolic rates are constant) and that decay follows a simple exponential degradation $\frac{dM(t)}{dt} = M_0 \cdot e^{-\lambda \cdot t}$ (Dölken et al., 2008; Amorim et al., 2010; Raghavan et al., 2002; Shalem et al., 2008; Miller et al., 2011; Sun et al., 2012; Schwanhäusser et al., 2011; Schueler et al., 2014). Either using short RNA labelling pulses or given that RNA production and decay complement each other when metabolism is in equilibrium, some studies additionally calculated synthesis rates (Miller et al., 2011; Sun et al., 2012; Schwanhäusser et al., 2011; Slobodin et al., 2020). Other approaches moved away from the steady-state assumption and described non-constant metabolic rates as a function of time (Zeisel, 2011; De Pretis et al., 2015; Rabani et al., 2011; Rabani et al., 2014). They discriminated between pre-

mRNA and mature mRNA using intronic sequence information and estimated synthesis, splicing and decay rates. These time-sensitive models are very helpful to investigate the dynamic changes during cellular response to external stimuli. At the same time, they are likely to overfit the data or not be perfectly accurate. They either require a lot of parameters for fitting the time-dependent functions or compute one estimate per time point, assuming a linear behaviour of the metabolic parameters between two consecutive time points.

One mathematical approach specifically designed to determine mRNA degradation rates from SLAM seq data is GRAND-SLAM by Jürges et al. (2018). Given the observed T>C conversions, they first calculate estimates for the proportion of newly synthesised RNA and then fit a simple exponential decay model to the obtained new-by-total RNA ratios over time. In detail, T>C labelling efficiency is estimated from the subpopulation of reads that harbour enough T>C conversions to discriminate these conversions from sequencing errors, assuming that labelling efficiency is constant over time. As a second step, new-by-total RNA ratios are estimated per time point and mRNA transcript based on the full data. To do so, they apply Bayesian inference to a binomial mixture model that defines the probability of the observed T>C conversions for a newly synthesised and a pre-existing transcript, respectively. Finally, the resulting posterior probabilities for the RNA ratios are described by beta distributions, and the exponential decay model is fit to the new-by-total RNA ratios over time.

Notably, all methods mentioned above to quantitatively access mRNA metabolic rates on a transcriptome-wide scale focus on mRNA decay, splicing, or synthesis rates. To the best of our knowledge, there is currently no method available to calculate mRNA export rates. Additionally, SLAM seq as the probably best-performing RNA labelling procedure to date is only poorly covered by computational approaches to analyse the resulting data. GRAND-SLAM itself harbours a few, potentially not negligible biases: Estimating labelling efficiency on the subpopulation of highly converted reads enriches the reads in fragments exposed to RNA editing, which can lead to efficiency over-estimation. Ratio values classically exhibit small absolute variances for values near 0 and 1 and large absolute variances for values near 0.5, which, if not corrected for, results in the data points being unequally weighted in the model fitting procedure, although equally informative.

To overcome the present limitations, I developed a dynamic model for the estimation of mRNA metabolic rates. The model distinguishes between the nuclear and cytosolic compartment (two-compartment model) and includes nuclear export as one of the mRNA metabolic processes. It

uses RNA labelling data produced by SLAM seq and therefore enables the estimation of metabolic parameters on a transcriptome-wide scale. Further, the data pre-processing steps include the correction for potential RNA editing sites, estimation of a time-dependent labelling efficiency, and variance-stabilizing transformation of the data. Overall, the model facilitates the transcriptome-wide estimation of RNA metabolic parameters while correcting for multiple drawbacks and, for the first time, enables the investigation of nuclear export.

We conducted a SLAM seq experiment on HeLa-S3 cells and combined it with cellular fractionation to measure nuclear and cytosolic RNA separately. Then, we applied my model to the metabolic labelling data to estimate nuclear and cytosolic RNA half lives. We observe that nuclear half lives are much higher than cytosolic half lives and conclude that mRNA export is slower than mRNA degradation. Consequently, mRNA transcripts spend most of their lifetime in the nucleus and are more abundant in the nucleus than in the cytosol. We find a group of outstanding genes, that we call ‘Supernova’ genes, with an exceptionally fast mRNA export, arguing for a novel and distinct export mechanism.

2 Methods

2.1 Data sets

We measured three independent SLAM seq time series in HeLa-S3 cells (Herzog et al., 2017) (Figure 1). The experiments were carried out by Kristina Zumer, who also did the sequencing library preparation. Sequencing was done on a HiSeq2000 sequencer (Illumina) at MPI-BPC. In each experiment, two biological samples were measured across a series of time points after labelling onset. Importantly, the nuclear and cytosolic fraction were separated after harvesting the cells, yielding one nuclear and one cytosolic RNA measurement for each time point. A fixed amount of synthetic RNA spike-ins was added to the samples directly after fractionation. Half of the spike-ins were 4sU labelled, the other half was unlabelled. 3'-sequencing (stranded, single-end) was performed to capture transcripts which had already undergone polyadenylation. The T>C nucleotide conversions introduced by the metabolic labelling on mRNA level translate into A>G conversions in reads aligned to the (-) strand and T>C conversions in reads aligned to the (+) strand.

The first, preliminary experiment was conducted to find suitable time points for the subsequent experiment. This experiment included two unlabelled samples which served as negative controls and are hereafter referred to as control samples. The second, pilot experiment was performed to inspect the adapted choice of time points. It was, besides that, used only to examine the recovery of ambiguously or unmapped reads and is in the following referred to as pilot data set. The final, main experiment consists of a time series of measurements at 15, 30, 45, 60, 90, 120 and 180 minutes after labelling onset (Figure 1).

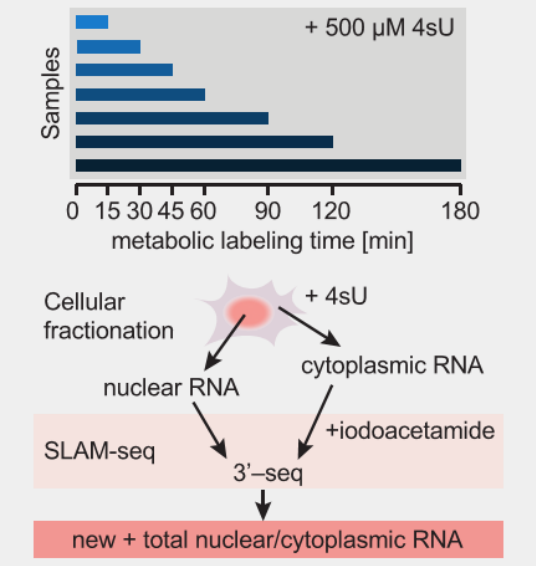


Figure 1 Experimental setup. We performed SLAM seq experiments for the metabolic labelling of RNA in HeLa-S3 cells (Herzog et al., 2017). Cells were treated with the nucleoside analogue 4sU, which is incorporated into newly synthesised RNA transcripts. Cells were harvested after 15, 30, 45, 60, 90, 120, and 180min of treatment (labelling). The nuclear and cytosolic fraction were separated subsequently. Each fraction was treated with iodoacetamide to alkylate the 4sU residues in the newly synthesised RNA transcripts. Alkylated 4sU residues are misinterpreted by the reverse transcriptase, which introduces T>C conversions during cDNA synthesis. Finally, 3'-seq library preparation and sequencing were performed. The sequencing reads measure the nuclear and cytosolic total and newly synthesised RNA.

The experiments and the library preparation were conducted by Kristina Zumer.
Figure by Kristina Zumer, personal communication.

2.2 Read alignment

Reads were mapped to the reference genome hg19 and raw alignments were quality filtered with slamdunk (Neumann et al., 2019) (version 0.3.0, calling *slamdunk map* and *slamdunk filter* with default settings except for resolving reads mapped to multiple genomic loci (multi-mappers) and setting the number of reported alignments per read to 100) (Figure 2, Figure 3, Figure 4A). To compare alignment performance, the control samples were additionally mapped with bowtie2 (Langmead & Salzberg, 2012) (version 2.2.6, default settings) and hisat2 (Kim, Langmead, & Salzberg, 2015) (version 2.0.0-beta, using default settings but setting the mismatch penalty to 1 and disabling spliced alignments) (Figure 2).

2.3 Recovery of multi-mapped and unmapped reads

To assess whether reads mapping to multiple genomic loci (multimappers) and unmapped reads could be recovered, these reads were collected and subjected to a second, modified alignment step described below (Figure 4). To avoid confusion, the original alignment (from which the

multi-mapped and unmapped reads were taken) will be called ‘primary alignment’, and the second, modified step will be called ‘recovery alignment’ in this section.

Multimappers and unmapped reads were filtered from the primary alignment. As the sequencing data is 3’-enriched, recovery mappings were run against a reference consisting of annotated hg19 3’UTR sequences. 3’UTR annotations were retrieved from the UCSC table browser (Karolchik et al., 2004), and overlapping regions were merged irrespective of strand orientation to generate the reference. Multimappers were re-mapped against the 3’UTR reference and considered rescued if there was exactly one reliable alignment (Figure 4B). Unmapped reads were re-mapped against the 3’UTR reference using a degenerated nucleotide alphabet for both the reads and the reference to mask potential labelling conversions. As the sequencing data is stranded, the reference was modified based on 3’UTR strand orientation, converting all G to A (- strand), all C to T (+ strand), or both (unknown strand orientation as a consequence of overlapping 3’UTRs from opposite strands). Reads were degenerated accordingly, yielding one version with all G converted to A and one version with all C converted to T per read. An unmapped read was considered rescued by the recovery mapping if exactly one of its degenerated versions had exactly one reliable alignment (Figure 4B). All recovery mappings were executed and quality filtered with slamdunk (Neumann et al., 2019), using the same version and settings as for the primary mappings (2.2 Read alignment).

The recovery mappings were applied to the pilot experiment. A large fraction of multi-mapped reads could be rescued (Figure 4B). Yet, the recovered reads could not make up for the uniquely mapped reads lost in the nuclear fraction over time (Figure 3A, Figure 4B). At the same time, there is no guarantee that the rescued multi-mapped reads originate from the 3’UTRs they were assigned to in the recovery mapping. Therefore, the recovery mappings were not performed on the final SLAM seq experiment data.

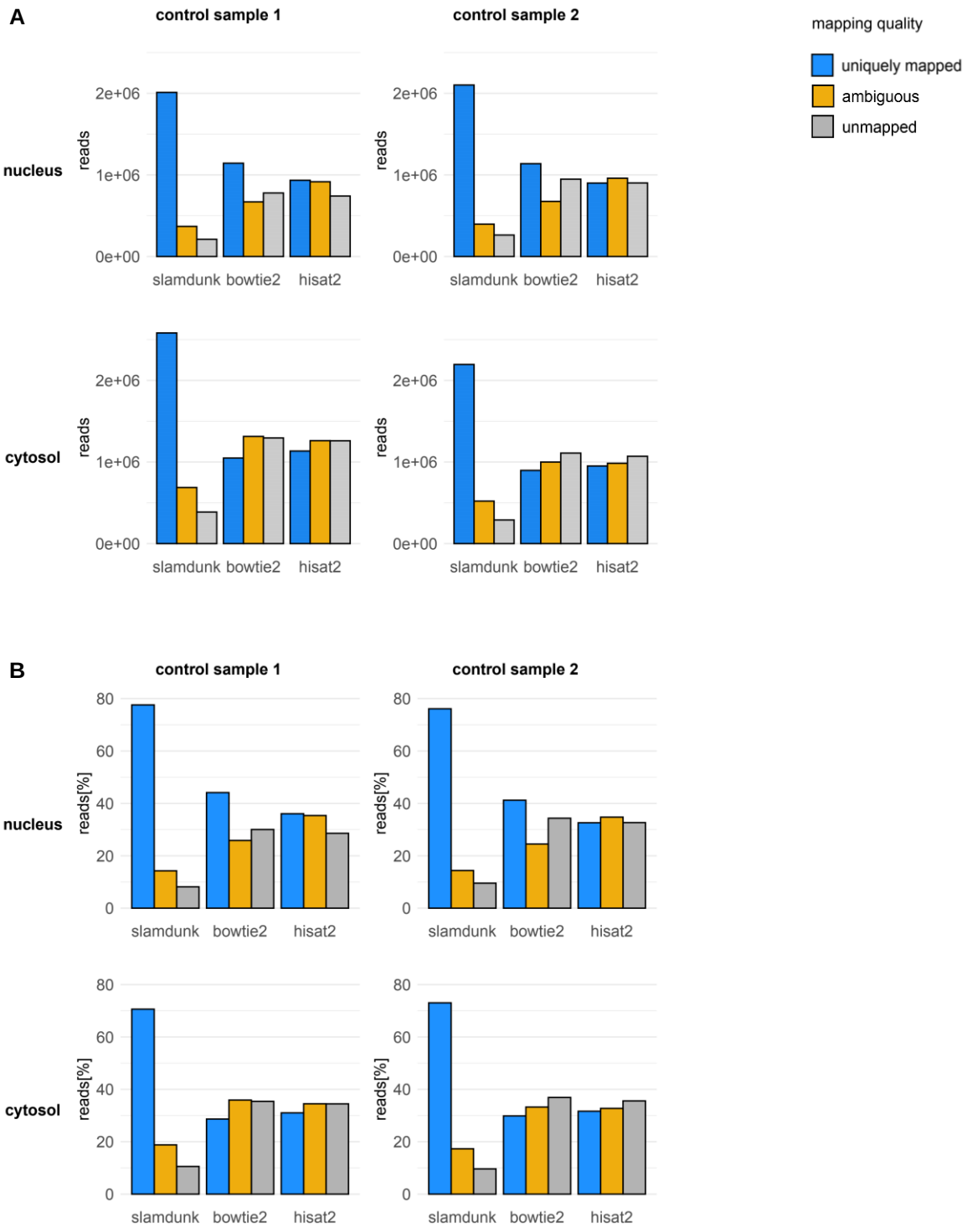


Figure 2 Mapping statistics of the control samples. (A) Absolute numbers. (B) Relative numbers. Data is shown for both sample 1 and 2 and the nuclear and cytosolic fraction, respectively. Data processing with slamdunk includes mapping and alignment quality filtering. The fraction of reads with a specific alignment quality is colour-coded. Uniquely mapped reads: reads with exactly one alignment (bowtie2 and hisat2), reads with exactly one best-scoring, reliable alignment according to quality control (slamdunk). Ambiguous reads: reads with more than one alignment (bowtie2 and hisat2), reads with more than one best-scoring, reliable alignment or only unreliable alignments (slamdunk). Unmapped reads: reads with no alignment (slamdunk, bowtie2, hisat2).

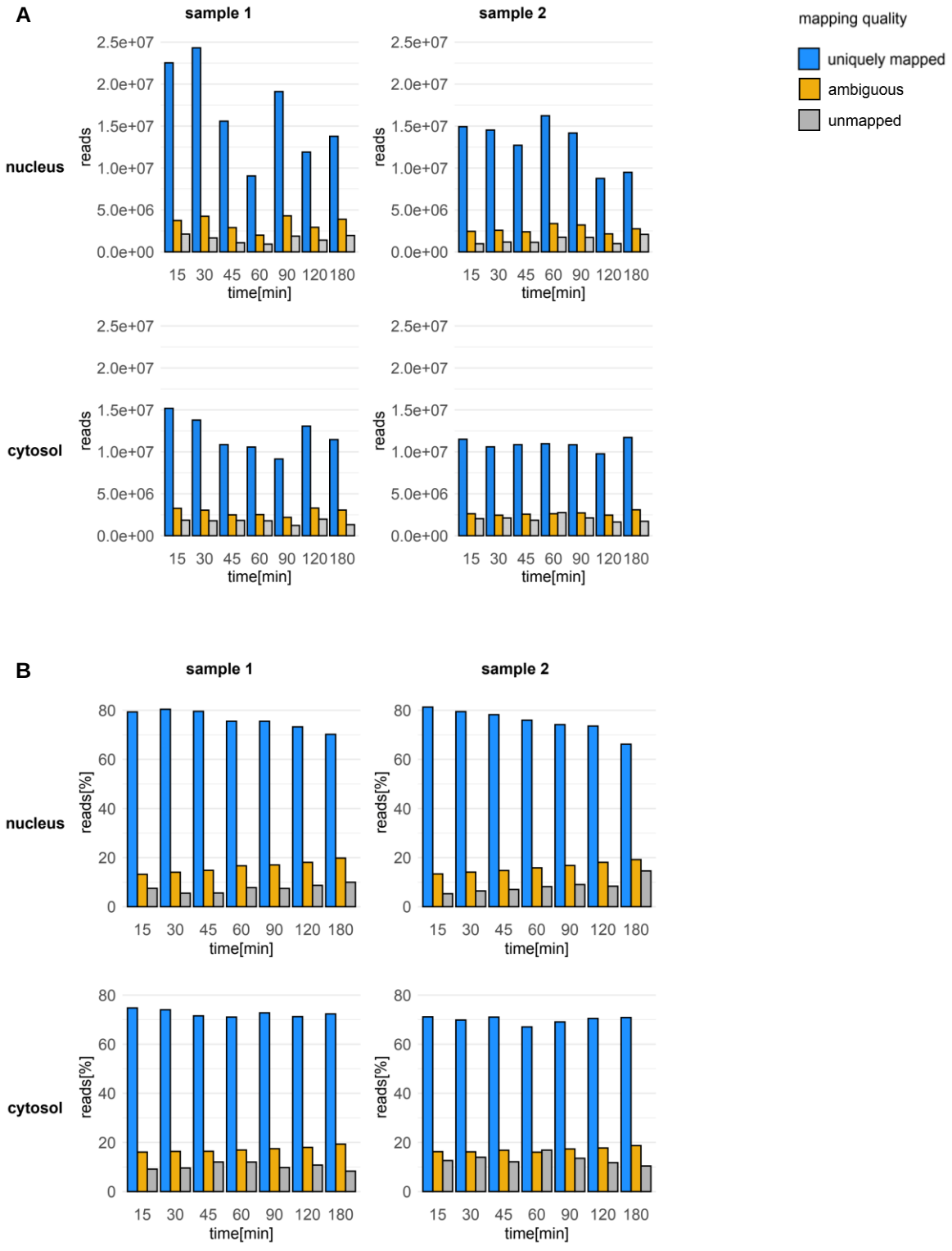
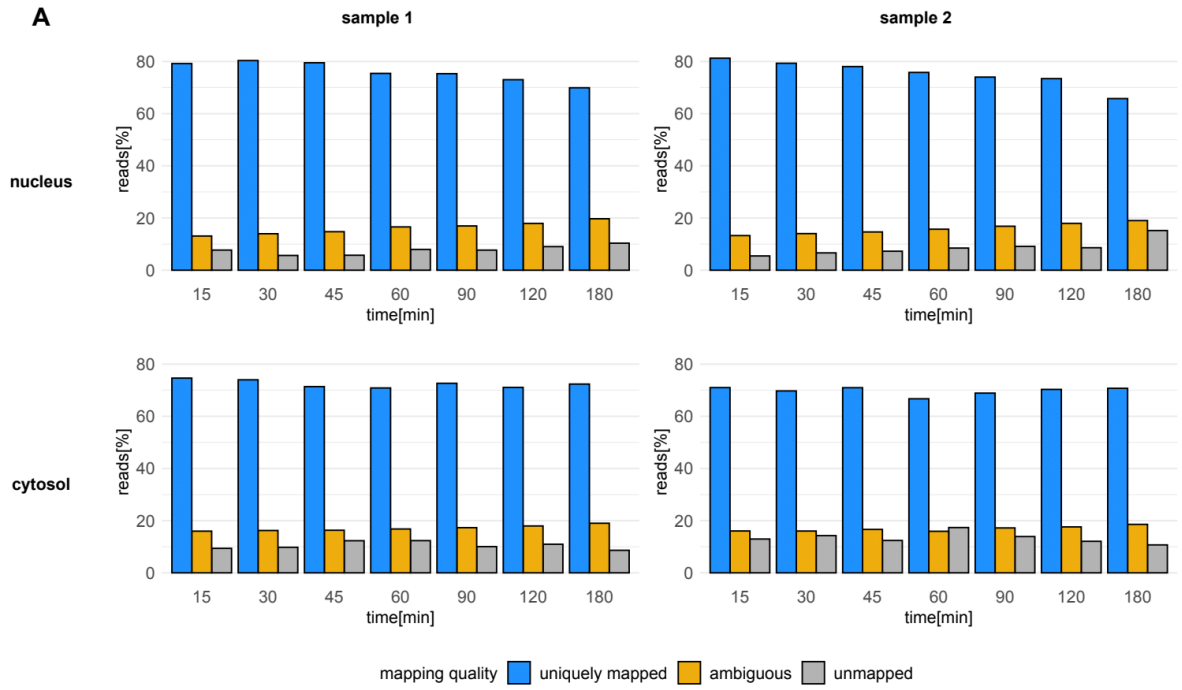


Figure 3 Mapping statistics of the final SLAM seq experiment. (A) Absolute numbers. (B) Relative numbers. Data is shown for both sample 1 and 2 and the nuclear and cytosolic fraction, respectively, resolved by the measured time points. Mapping was performed and quality filtered with slamdunk. The fraction of reads with a specific alignment quality is colour-coded. Uniquely mapped reads: reads with exactly one best-scoring, reliable alignment. Ambiguous reads: reads with more than one best-scoring, reliable alignment or only unreliable alignments. Unmapped reads: reads with no alignment.


A
sample 1
sample 2
nucleus
cytosol
reads[%]
time[min]
15
30
45
60
90
120
180
mapping quality
uniquely mapped
ambiguous
unmapped

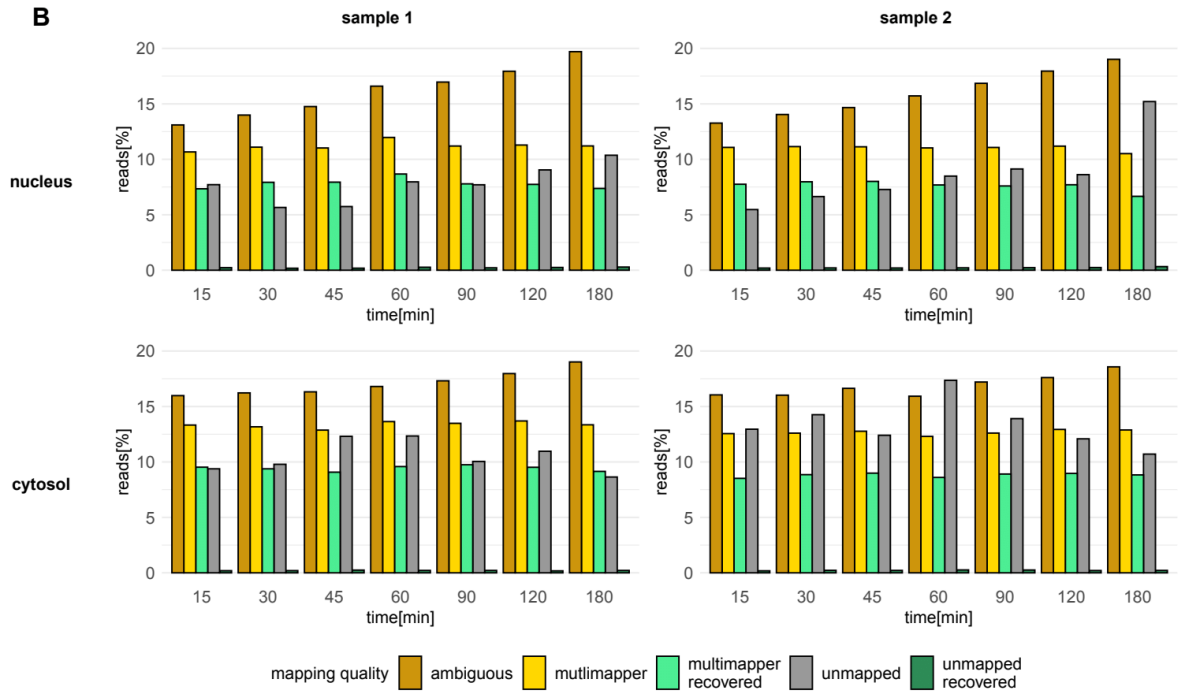

B
sample 1
sample 2
nucleus
cytosol
reads[%]
time[min]
mapping quality
ambiguous
mutlimapper
multimapper recovered
unmapped
unmapped recovered

Figure 4 Mapping statistics for the pilot data set's primary and recovery mappings. Data is shown for both sample 1 and 2 and the nuclear and cytosolic fraction, respectively, resolved by the measured time points. All mappings were performed and quality filtered with slamdunk. The fraction of reads with a specific alignment quality is colour-coded. (A) Primary mapping. Uniquely mapped reads: reads with exactly one best-scoring, reliable alignment. Ambiguous reads: reads with more than one best-scoring, reliable alignment or with only unreliable alignments. Unmapped reads: reads with no alignment. (B) Primary mapping and recovery mappings in comparison. Ambiguous reads: reads with more than one best-scoring, reliable alignment or only unreliable alignments in the primary mapping. Multimappers: reads with more than one best-scoring, reliable alignment in the primary mapping. Multimappers recovered: reads with more than one best-scoring, reliable alignment in the primary mapping, but exactly one best-scoring, reliable alignment using a reference made from annotated 3'UTRs in a recovery mapping step. Unmapped reads: reads with no alignment in the primary mapping. Unmapped recovered reads: reads with no alignment in the primary mapping, but with exactly one best-scoring, reliable alignment using a degenerated 3-nucleotide alphabet for the reference made from annotated 3'UTRs and the reads in a recovery mapping step.

2.4 Assigning reads to annotated 3’UTRs

3'UTR annotations for hg19 were downloaded from the UCSC Table Browser (Karolchik et al., 2004). Overlapping 3'UTRs were merged considering strand orientation. Uniquely mapped reads were assigned to a specific 3'UTR if their alignment overlapped with the UTR with at least one nucleotide position (Figure 5). As a consequence of the stranded sequencing protocol, reads of transcripts encoded on the (-) strand map forwards, and reads of transcripts encoded on the (+) strand map reversely, which was accounted for during assignment. For the sake of simplicity, reads assigned to known, annotated 3’UTRs will be named 3’UTR reads, and reads left unassigned will be named non-3’UTR reads in the following (yet, they might originate from unannotated 3’UTRs).

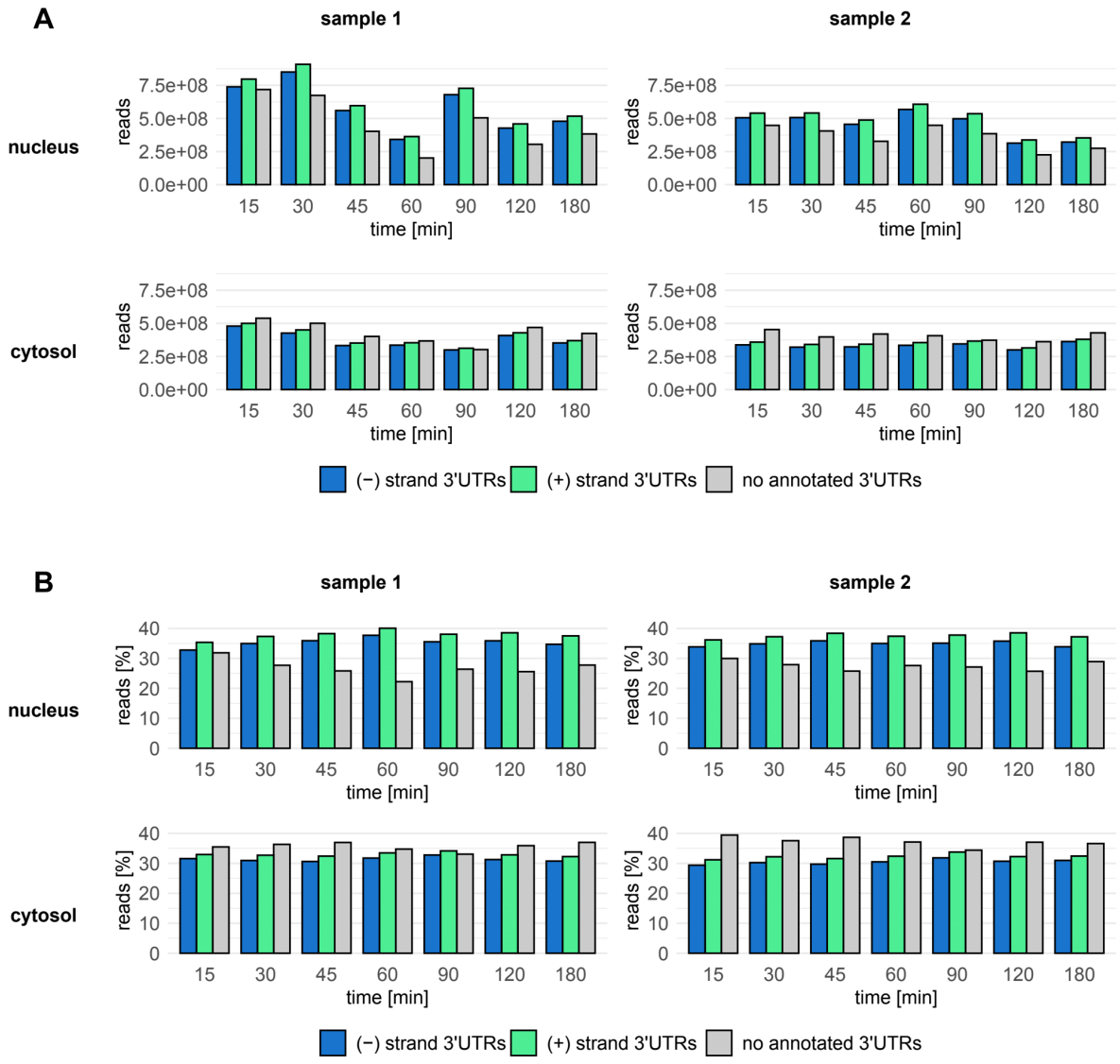


Figure 5 Read assignment to annotated 3'UTRs. Data is shown for both sample 1 and 2 and the nuclear and cytosolic fraction, respectively, resolved by the measured time points. Uniquely mapped reads of the final SLAM seq experiment were assigned to hg19 annotated 3'UTRs. Overlapping 3'UTRs were previously merged considering strand orientation. Read mapping orientation was considered during the assignment process. The number of reads assigned to (-) strand or (+) strand encoded 3'UTRs or without assignment is colour-coded. (A) Absolute read counts. (B) Relative read counts.

2.5 Peak calling

We sought to identify distinct regions of the genome that are covered by dense read clusters. These regions, which we name coverage peaks or simply peaks, are later subjected to metabolic parameter estimation. A peak may overlap with an annotated 3'UTRs and is then assigned to this 3'UTRs. Multiple peaks can be assigned to the same 3'UTR. In that case, they can either derive from A-rich sequences that serve as internal (i.e. non-polyA) priming sites during the sequence amplification process in library preparation, or from different 3'UTR isoforms. Peaks located outside of annotated 3'UTRs are most likely created by internal priming sites. Some of these peaks map to intronic or exonic regions of transcripts. All other peaks could be derived from unknown transcripts, non-coding transcripts, or (if the number of reads of the peak is low) mapping artefacts.

To properly define a peak region, we count the number of reads mapped at each position of the reference genome (per-nucleotide coverage). Then, we examine the coverage as a function of distance to a given central position. Based on the resulting coverage profile, we fix a distance radius that specifies the peak.

2.5.1 Per-nucleotide coverage calculation

Peaks mapped within annotated 3'UTRs are primarily derived from priming against the polyA-tail, whereas peaks located outside of annotated 3'UTRs are mainly derived from internal priming sites or mapping artefacts. Consequently, these two groups of peaks may exhibit different shapes. Peak calling was therefore executed separately for 3'UTR and non-3'UTR reads.

Uniquely mapped reads were pooled across the time points of a time series measurement (but kept separately for sample 1 and 2 as well as the nuclear and cytosolic fraction). The number of reads mapped at each position of the reference genome was calculated, considering strand orientation (per-nucleotide coverage). Only the nucleotide at the 3'end of each read was used for the calculation, as it is closest to the corresponding polyA or internal priming site. Consequently, each read was counted exactly once.

Subsequently, a coverage profile was created to define peak width. Finally, coverage convolution was applied to smooth out noise and call peak centres based on local coverage maxima. The single steps are described in more detail in the following.

2.5.2 Coverage profiles

Coverage profiles describe to which extent, around a nucleotide position that is covered, other nucleotide positions are covered too. In detail, for each nucleotide position with a coverage of at least 1, it is evaluated whether or not the positions within a [-1000, 1000] window around this nucleotide are also covered or not. Each position within the [-1000, 1000] window is scored with a 1 in case it is covered and 0 if it is not. The central window position 0 represents the evaluated nucleotide and is therefore always scored 1. This procedure yields one [-1000, 1000] score vector for each covered nucleotide position. The coverage profile is retrieved by summing all score vectors. The relative profile is created by dividing the coverage profile vector by the central position's score (which is equal to the total number of covered nucleotide positions). Based on the shape of the relative profiles, a coverage peak was defined as the -/+50 nucleotide region around a local coverage maximum (Figure 6). Note that this peak width of, in total, 101 nucleotide positions is not an artefact of read length, as each read is only counted once, namely with the nucleotide at its 3' end.

2.5.3 Coverage convolution and peak calling

To smooth out noise, coverage was convolved with a triangular function. The width of the triangle was set to the peak width (determined based on the coverage profiles in the previous step). Accordingly, the triangle was defined by a 50 nucleotide radius around the central position, resulting in a width of 101 nucleotide positions. All local maxima found within the convolved coverage served as peak centres and were extended by -/+50 nucleotides around the centres to form the called peak regions. This initially identified a total of 98,102 peaks for the 3'UTR reads (3'UTR peaks) and 1,262,263 peaks for the non-3'UTR reads (non-3'UTR peaks). Of course, not all of these peaks are informative, as many originate from mapping artefacts or are generally poorly covered. Therefore, we define a peak detectable if it fulfils the following two conditions: (1) the peak has at least 1 read assigned in each measurement (i.e., each time point measured in each compartment and sample), (2) the peak has at least 30 reads assigned on average across a time series (i.e., per compartment and sample). This yielded 10,170

detectable 3’UTR peaks and 1651 detectable non-3’UTR peaks. The peaks were annotated as described in the following chapter.

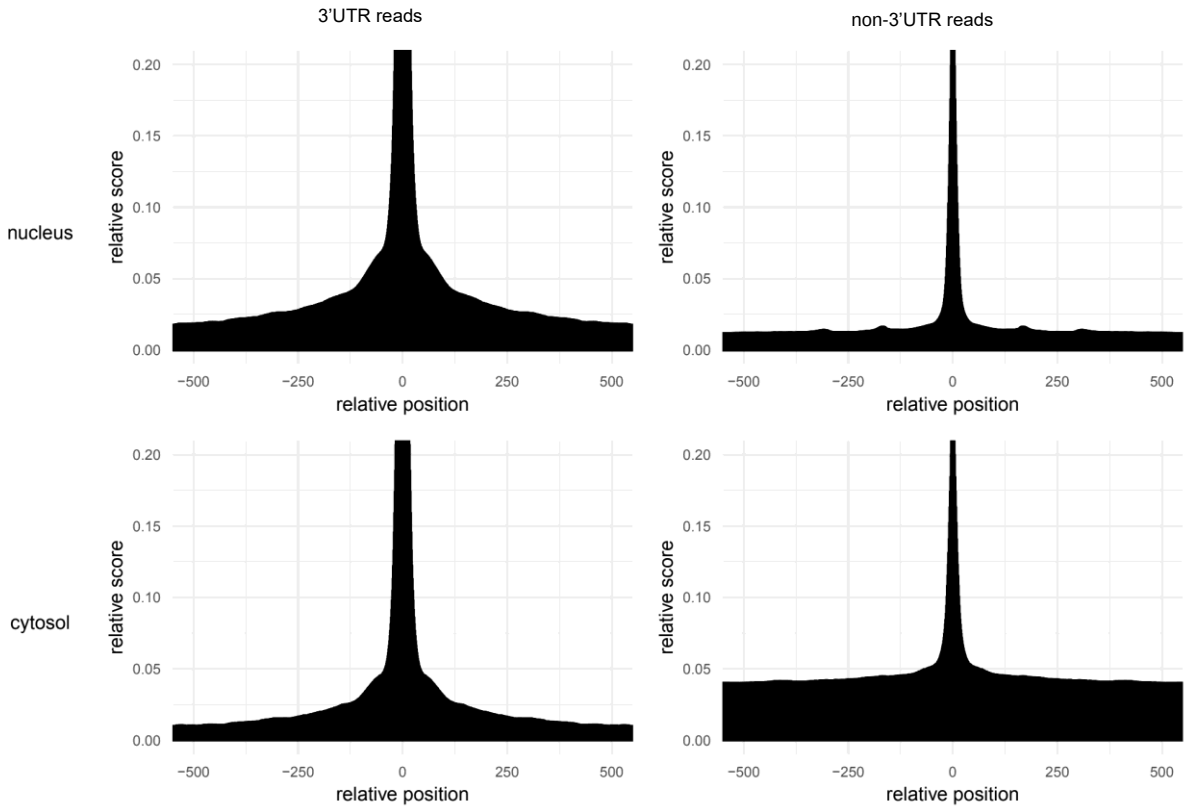


Figure 6 Relative coverage profiles. Profiles were created over a window of 1000 nucleotides both upstream and downstream of a central, covered position. Profiles were evaluated separately for the reads assigned to annotated 3'UTR regions and reads not assigned to such as well as the nuclear and cytosolic fraction. The shown profiles are trimmed on the x- and y-axis to enhance information content; the full profiles can be found in Supplemental Figure 1.

2.6 Annotation of 3’UTR regions and coverage peaks

To assign 3’UTR regions and single coverage peaks to their respective coding genes, the hg19 genome annotation was downloaded from NCBI RefSeq (O'Leary et al., 2016). A 3’UTR region or a coverage peak was assigned to each gene it overlapped with at least one nucleotide position.

3’UTR peaks were additionally assigned to the 3’UTR regions they were overlapping (with at least one nucleotide). In rare cases, a 3’UTR peak was assigned to two 3’UTR regions (20 out of all 10,170 detectable peaks).

Non-3’UTR peaks were further classified. Annotated 5’UTR, exonic and intronic regions were downloaded from the UCSC Table Browser (Karolchik et al., 2004) (additionally to the previously downloaded 3’UTR annotations). 5’UTRs and 3’UTRs are usually also listed as exonic regions. We will distinguish between 5’UTRs, 3’UTRs, and all other exonic regions to avoid confusion. Henceforth, the term ‘exonic region’ excludes 5’UTRs and 3’UTRs. A non-3’UTR coverage peak was assigned to any of these annotated regions if it overlapped with at least one nucleotide position. A peak was classified as exonic (respectively intronic) if it was only assigned to exonic (respectively intronic) regions.

2.7 Mismatch statistics and correction for SNP and editing sites

To estimate sequencing errors, single-nucleotide mismatch rates were calculated for the control samples. Some of the observed mismatches will not be due to technical errors but biological RNA editing (Nishikura, 2016). Edited RNA nucleotides could be misinterpreted and converted by the reverse transcriptase during the library preparation. These editing events may not be equally abundant within the nuclear and cytosolic fraction or within 3’UTR regions and outside of such. Single-nucleotide mismatch rates were therefore calculated separately for the nuclear and cytosolic compartment as well as the 3’UTR and non-3’UTR reads.

Mismatch rates were biased strand-specific for the reads assigned to 3’UTRs (Figure 7). Especially A>G and T>C were enriched on the (+) and respectively (-) strand (note that the SLAM seq experiment combined with the strand-preserving library preparation introduces A>G on the (-) and T>C on the (+) strand). RNA editing events can explain such events. The A>G and T>C bias could be a consequence of A-to-I editing, which occurs in both coding and non-coding RNAs (Nishikura, 2016). Since editing conversions can flaw the count statistics of labelled uridines, it is crucial to correct for potential editing sites. Additionally to the strand-specific bias, mismatch rates were generally higher for the non- 3’UTR reads, especially in the cytosolic fraction (Figure 7). To ensure that spurious mapping artefacts do not mainly introduce the observed mismatch rates of the non-3'UTR reads, only non-3’UTR coverage peaks with a minimum of 5 reads assigned in a compartment and sample were analysed.

To correct the observed mismatch biases, known hg19 SNP positions and potential RNA editing sites were removed from further analysis. Known hg19 SNP positions were obtained from NCBI dbSNP (Sherry et al., 2001). To identify editing sites, the reads of both control samples were pooled. Then, position-wise mismatch rates were calculated separately for the nuclear and cytosolic fraction. Sites with a mismatch rate of more than 5% on either strand were called as potential editing sites. A cut-off of 5% was chosen as it is substantially larger than technical sequencing errors (Pfeiffer et al., 2018) but low enough to capture nucleotide positions that are not frequently modified. As the library sizes of the control samples were smaller than of the final SLAM seq experiment and the correction for biased nucleotide positions is of high importance, no coverage restrictions were applied to this identification of potential editing sites. Finally, the editing sites called for the nuclear and cytosolic fraction were pooled.

Known SNP and potential editing sites were first excluded from the analysis of the control samples. All of the observed biases were substantially reduced particularly after the removal of RNA editing sites, yielding comparable mismatch rates both between strand-specific A>G respectively T>C and other conversions, between the nuclear and cytosolic fraction, and between 3'UTR and non-3'UTR peaks (Figure 7).

SNP and RNA editing site correction was then applied to the final SLAM seq experiment measurements. The mismatch statistics are shown for 3'UTR reads in Figure 8 and non-3'UTR coverage peaks comprising at least 5 reads in Figure 9. Conversions introduced by the metabolic labelling are expected to display as A>G on the (-) and T>C on the (+) strand. The labelling conversions show an increasing trend over time, while the other mismatch rates constantly stay low. Labelling conversions increase less in the cytosol than in the nucleus, as the newly synthesised, labelled RNA transcripts are produced in the nucleus, and their export to the cytosol requires time.

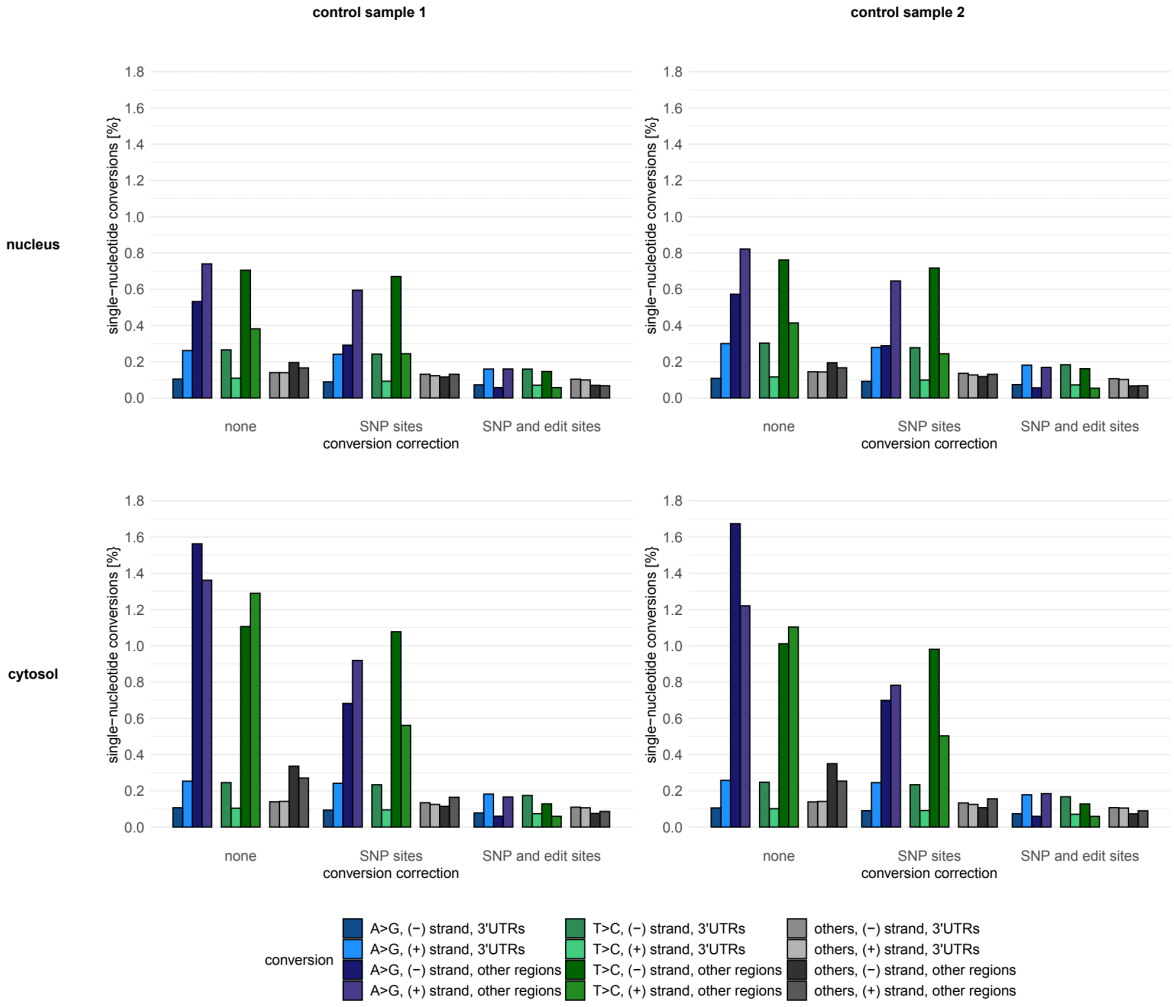


Figure 7 Mismatch statistics of the control samples. Single-nucleotide mismatch rates in the nuclear and cytosolic fraction of control samples 1 and 2. Mismatch rates are shown for different nucleotide conversion correction strategies. The mismatch type is colour-coded. No conversion correction: mismatch rates observed in the alignment. SNP-sites correction: mismatch rates observed in the alignment when excluding all nucleotide positions annotated as known SNP site. SNP and edit-site correction: mismatch rates observed in the alignment when excluding all nucleotide positions annotated as known SNP site or called as potential editing sites. A>G (T>C) conversions: A>G (T>C) conversions observed in the read alignment. Other conversions: average over all non-A>G and non-T>C conversions observed in the read alignment. Conversions of (-) strand ((+) strand) 3'UTRs: conversions of all reads assigned to an annotated (-) strand ((+) strand) 3'UTRs. Conversions of other regions on the (-) strand ((+) strand): conversions of all reads of a non-3'UTR peak comprising at least 5 reads on the (-) strand ((+) strand).

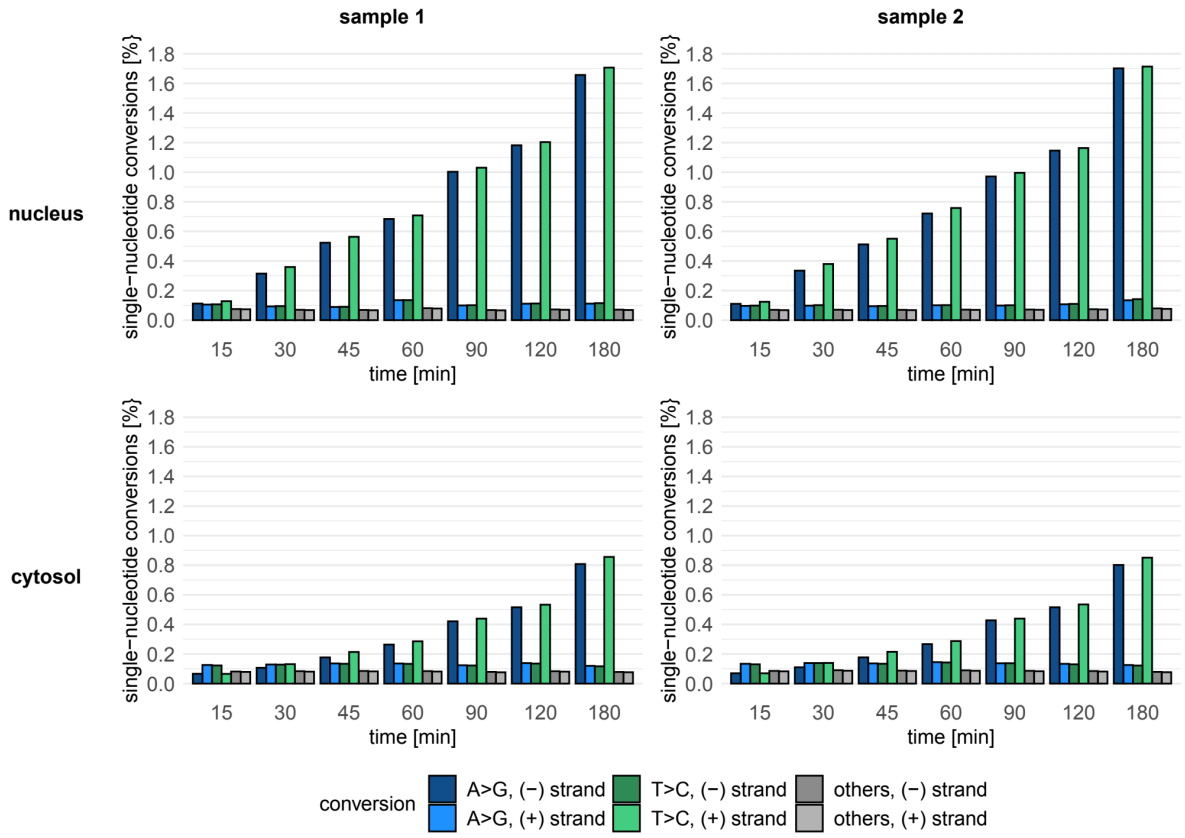


Figure 8 Mismatch statistics of the 3'UTR reads of the final SLAM-seq experiment. Single-nucleotide mismatch rates in 3'UTR reads of the final SLAM seq experiment. Mismatch statistics are calculated separately for the nuclear and cytosolic fraction in sample 1 and 2, and resolved by the measured time points. The mismatch type is colour-coded. Mismatches are corrected for both annotated SNP sites and potential editing sites. A>G (T>C) conversions: A>G (T>C) conversions observed. Other conversions: average over all non-A>G and non-T>C conversions observed. Conversions of the (-) strand ((+) strand): conversions observed for the reads assigned to an annotated (-) strand ((+) strand) 3'UTR.

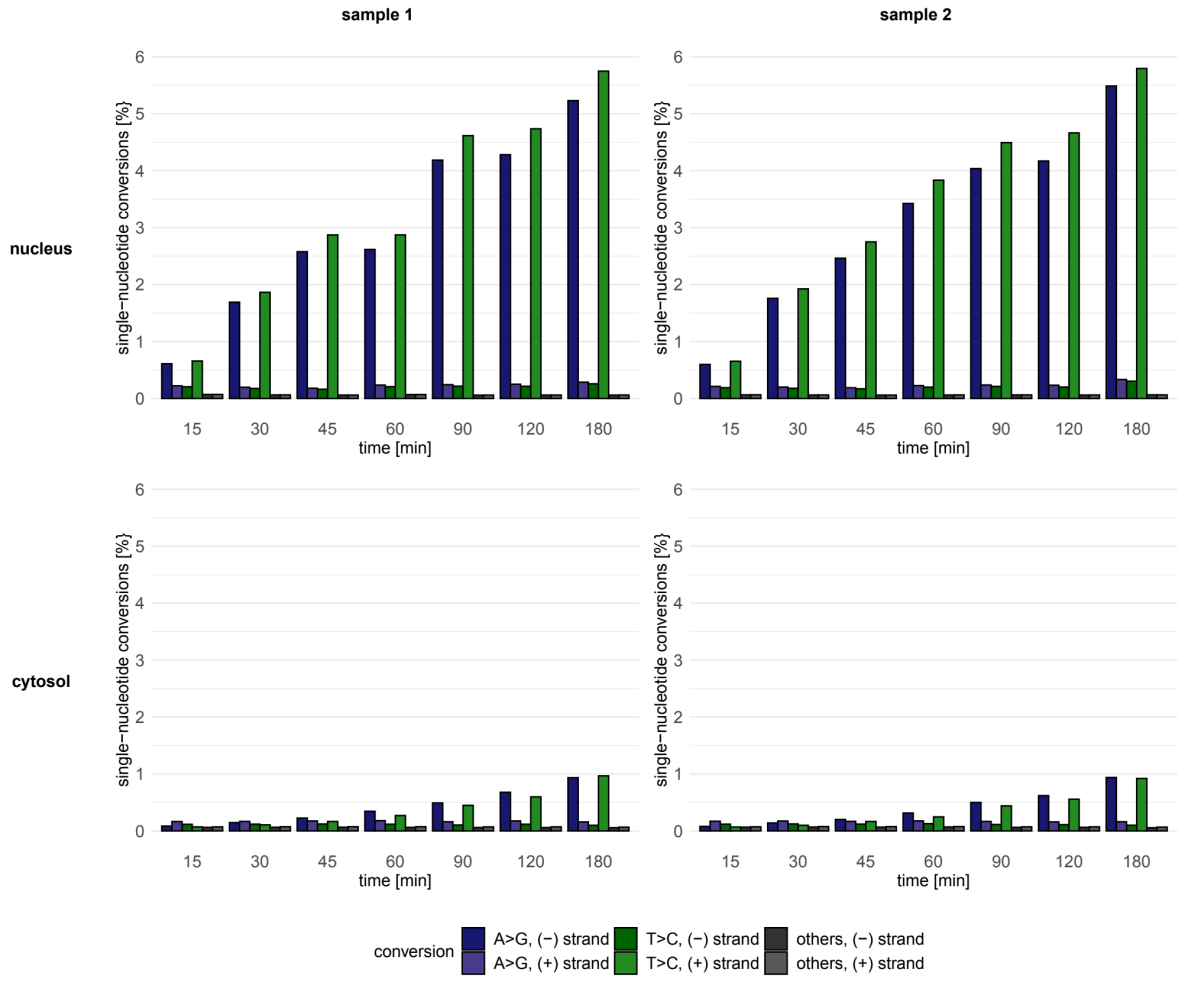


Figure 9 Mismatch statistics of the non-3'UTR reads of the final SLAM-seq experiment. Single-nucleotide mismatch rates in non-3'UTR reads of the final SLAM seq experiment. Mismatch statistics are calculated separately for the nuclear and cytosolic fraction in sample 1 and 2, and resolved by the measured time points. The mismatch type is colour-coded. Statistics are based on non-3'UTR coverage peaks comprising at least 5 reads, and mismatches are corrected for both annotated SNP sites and potential editing sites. A>G (T>C) conversions: A>G (T>C) conversions observed. Other conversions: average over all non-A>G and non-T>C conversions observed. Conversions of the (-) strand ((+) strand): conversions observed for the reads assigned to a (-) strand ((+) strand) non-3'UTR peak.

2.8 Estimation of sequencing errors and labelling time shift

2.8.1 Estimation of sequencing errors

G>non-G and C>non-C sequencing errors can mask the A>G and T>C labelling conversions, which introduces false-negative errors. On the other hand, A>G and T>C sequencing errors mimic the labelling conversions and cause false-positive observations. We calculate false-negative and false-positive sequencing error rates to consider these errors while estimating the labelling efficiency and the newly synthesised RNA ratios. The false-negative error rates can directly be calculated from the final SLAM seq experiments, as the metabolic labelling does not influence the G>non-G and C>non-C mismatch rates. The false-positive error rates have to be inferred with the information from the control samples. The details are described in the following. Note that, due to the strand-preserving library preparation, only A>G and G>non-G conversions on the (-) strand and T>C and C>non-C conversions on the (+) strand are relevant for the error rate estimation.

The error rates were estimated based on the mismatch rates in the control and final SLAM seq experiment after SNP and editing site correction. For the non-3'UTR reads, only reads from coverage peaks comprising at least 5 reads were included in the calculations to exclude mapping artefacts. Mismatch rates were almost identical between both control or SLAM seq experiment samples, respectively (Figure 7, Figure 8, Figure 9). Therefore, mismatch rates were averaged over the two samples of an experiment prior to error rate estimation.

G>non-G and C>non-C sequencing errors (false-negative errors) were obtained from the final SLAM seq experiment. G>non-G mismatch rates on the (-) strand and C>non-C mismatch rates on the (+) strand were averaged over all time point measurements. This yielded one false-negative error rate per cellular compartment and read identity (3'UTR or non-3'UTR) (Supplemental Table 1).

A>G and T>C sequencing errors (false-positive errors) in the final SLAM seq experiment had to be inferred with the information of the control samples. First, we calculated the fold changes between A>G or T>C mismatch rates and all other X>Y mismatch rates in the control experiment. Subsequently, for each sequencing error rate X>Y observed in the final SLAM seq experiment, one A>G and T>C sequencing error rate estimate was computed using the corresponding fold-change from the control data. Finally, the A>G estimates on the (-) strand and the T>C estimates on the (+) strand were averaged over all sequencing errors X>Y and

time point measurements. We thereby obtained one false-positive error rate per cellular compartment and read identity (3'UTR or non-3'UTR) (Supplemental Table 2).

2.8.2 Estimation of the labelling time shift

We observed that the labelling conversions barely increased within the first 15 minutes of the labelling experiment compared to the time window between 15 and 30 minutes (Supplemental Table 3, Supplemental Table 4). This delay in labelling, which we call labelling time shift, can occur because the 4sU needs time to distribute within the cells (i.e., 4sU diffusion). Further, a labelled transcript cannot directly be detected but has to be polyadenylated first; this can also cause a time gap between labelling onset and detection. We estimated this labelling time shift and subtracted it from the experimental time points to retrieve the effective labelling time points for subsequent parameter estimation. The detailed procedure is described in the following.

The labelling time shift was computed separately for the experimental samples to account for technical variation; for 3'UTR and non-3'UTR reads since non-3'UTR reads are mainly derived from internal priming sites and skip the polyadenylation-dependent time shift; and the forward and reverse strand to verify that there is no strand-specific bias. For the non-3'UTR reads, only reads from coverage peaks comprising at least 5 reads were used to exclude mapping artefacts.

In the nucleus, the ratio of newly synthesised transcripts and thereby the number of observed labelling conversions can be assumed to increase linearly in the early time points of the experiment. We estimate the labelling time shift based on this assumption. First, the labelling conversion rate of 0min was set to the sequencing error estimate. Second, the increase in the labelling conversion rate was calculated between the 0min and 15min as well as the 15min and 30min measurements. Then, we calculated the effective labelling time between 0min and 15min as the fraction of labelling conversion rate increase during this time window with respect to the 15min to 30min time window, and multiplied this fraction by the time window size (15min). The labelling time shift is defined by the non-effective labelling time. The forward and reverse strand's time shift estimates were comparable and therefore averaged. This resulted in one estimate per read identity (3'UTR or non-3'UTR reads) and sample (Supplemental Table 3, Supplemental Table 4).

2.9 Estimation of labelling efficiency and newly synthesised RNA ratio

The labelling efficiencies and newly synthesised RNA ratios were estimated jointly using an EM algorithm. The estimation is based on the observed A>G conversions on the (-) strand and T>C conversions on the (+) strand, as well as the previously estimated false-positive and false-negative error rates. Labelling efficiencies were estimated separately for 3'UTR regions and non-3'UTR coverage peaks to examine potential biases between these two groups. The details are described in the following. For better readability, it will hereafter be referred to 3'UTRs and T>C conversions only.

Fix a 3'UTR onto which J reads were mapped. Given a read $j = 1, \dots, J$, let T_j be its number of T-positions in the genomic sequence of the read alignment, and let $o_j \in 0,1,\dots,T_j$ be the number of positions at which we observe T>C conversions. Let $h_j \in \{0,1\}$ be a hidden variable indicating whether read j originates from a pre-existing RNA ($h_j = 0$), or a newly synthesised RNA ($h_j = 1$). Let $\rho \in [0,1]$ be the proportion of newly synthesised RNAs in the RNA population from which read j was drawn. Let $e \in [0,1]$ be the labelling efficiency, i.e., the probability by which a T>C conversion happens in a newly synthesised RNA. Let $b \in [0,1]$ be the probability of a T>C sequencing error (false-positive error). Let $\epsilon \in [0,1]$ be the probability of a C>non-C sequencing error (false-negative error). Probabilities b and ϵ are estimated independently using the control samples and the final SLAM seq experiment (as described in 2.8 Estimation of sequencing errors and labelling time shift). Denote the unknown parameters by $\Theta = (\rho, e)$. Then,

$$P(o_j, h_j ; \Theta) = P(h_j ; \rho) \cdot P(o_j \mid h_j ; e, b, \epsilon) \qquad (1)$$

In the above expression, $\mathrm{P}(\mathrm{h_j} ; \rho) = \mathrm{Bernoulli}(\mathrm{h_j} ; p = \rho)$ and

$$\mathrm{P}(o_j \mid \mathrm{h_j} ; \mathrm{e}, \mathrm{b}, \epsilon) = \begin{cases} \mathrm{Bin}(\mathrm{o_j} ; \mathrm{n} = T_j , \mathrm{p} = b) \text{ if } \mathrm{h_j} = 0 \\ \mathrm{Bin}(\mathrm{o_j} ; \mathrm{n} = T_j , \mathrm{p} = a) \text{ if } \mathrm{h_j} = 1 \end{cases} \qquad (2)$$

where a is the probability of seing a T>C converted nucleotide in a newly synthesised transcript, $a = e(1 - \epsilon) + (1 - e)b = e(1 - \epsilon - b) + b$.

In other words,

$$\log \mathrm{P}\left(\mathrm{o_j}, \mathrm{h_j}\,;\Theta\right)\begin{cases}\log(1-\rho)+\log\binom{T_j}{\mathrm{o_j}}+\mathrm{o_j}\log b+\left(T_j-\mathrm{o_j}\right)\log(1-b) & \text{if } \mathrm{h_j}=0\\ \log\rho\ +\ \log\binom{T_j}{\mathrm{o_j}}+\mathrm{o_j}\log a+\left(T_j-\mathrm{o_j}\right)\log(1-a) & \text{if } \mathrm{h_j}=1\end{cases} \tag{3}$$

2.9.1 Estimation of the labelling efficiency per transcript and measurement

An EM algorithm is performed to estimate the ratio of newly synthesised RNAs ρ and the labelling efficiency e for a 3'UTR.

E-step.

Let $H=(\mathrm{h_j}\,;\ j=1,\dots,J)$, $H_{-1}=(\mathrm{h_j}\,;\ j=2,\dots,J\,)$. Given some parameter set $\Theta'=(\rho',e')$, the function $Q(\Theta\,;\Theta')$ has to be optimized with respect to $\Theta=(\rho,e)$. Here,

$$\begin{aligned}
Q(\Theta\,;\Theta') &:= \mathbb{E}_{P(H\mid O;\,\Theta')}\log P(O,H;\Theta)\\
&= \textstyle\sum_{H\,\in\,\{0,1\}^J} P(H\mid O;\Theta')\cdot\log P(O,H;\Theta)\\
&= \textstyle\sum_{H_{-1}\in\,\{0,1\}^{J-1}}\sum_{h_1\in\,\{0,1\}}\left(P(H_{-1}\mid O_{-1};\Theta')\cdot P(o_1,h_1;\Theta')\right)\cdot\\
&\quad\ (\log P(H_{-1}\mid O_{-1};\Theta)+\log P(o_1,h_1;\Theta))\\
&= \textstyle\sum_{h_1\in\,\{0,1\}} P(h_1\mid o_1;\Theta')\cdot\ \log P(o_1\,,h_1;\Theta)+\\
&\quad\ \textstyle\sum_{H_{-1}\in\,\{0,1\}^{J-1}} P(H_{-1}\mid O_{-1};\Theta')\cdot\log P(H_{-1}\mid O_{-1};\Theta)\\
&\overset{\text{induction}}{=} \textstyle\sum_{j=1}^{J}\sum_{h_j\in\,\{0,1\}} P\left(h_j\mid o_j;\Theta'\right)\cdot\log P\left(o_j,h_j;\Theta\right)
\end{aligned} \tag{4}$$

Let

$$c_{j,h_j}:=P\left(h_j\mid o_j;\Theta'\right)=\frac{P\left(o_j\mid h_j;e\prime\right)\cdot P\left(h_j;\rho\prime\right)}{\sum_{h_j\in\,\{0,1\}}P\left(o_j\mid h_j;\mathrm{e}\prime\right)\cdot P\left(h_j;\rho\prime\right)}\ ,\ j=1,\dots,J \tag{5}$$

Let $C_0=\sum_j c_{j,0}$ and $C_1=\sum_j c_{j,1}$, $A=\sum_j c_{j,1}o_j$, $B=\sum_j c_{j,1}(T_j-o_j)$. Then, equation (4) simplifies to

$$\begin{aligned}
Q(\Theta\,;\Theta') &= \textstyle\sum_j c_{j,0}\left[\log(1-\rho)+\log\binom{T_j}{\mathrm{o_j}}+\mathrm{o_j}\log b+\left(T_j-\mathrm{o_j}\right)\log(1-b)\right]+\\
&\quad\ \textstyle\sum_j c_{j,1}\left[\log\rho\ +\ \log\binom{T_j}{\mathrm{o_j}}+\mathrm{o_j}\log a+\left(T_j-\mathrm{o_j}\right)\log(1-a)\right]\\
&= C_0\log(1-\rho)+C_1\log\rho+A\log a+B\log(1-a)+\text{const}
\end{aligned} \tag{6}$$

M-step.

Taking the partial derivative of Q with respect to ρ and equating this expression to zero yields

$$\begin{aligned} 0 &= \frac{\delta Q(\Theta;\Theta\prime)}{\delta\rho} = -\frac{C_0}{1-\rho} + \frac{C_1}{\rho} \\ \rho &= \frac{C_1}{C_0+C_1} = \frac{C_1}{J} \end{aligned} \quad (7)$$

Recall that $a = e(1-\epsilon-b)+b$, and take the partial derivative of Q with respect to e. Equating the resulting expression to zero yields

$$\begin{aligned} 0 = \frac{\delta Q(\Theta;\Theta\prime)}{\delta e} &= \frac{A}{a}\cdot\frac{\delta a}{\delta e} + \frac{B}{1-a}\cdot\frac{\delta(1-a)}{\delta e} \\ &= \frac{A}{a}\cdot(1-\epsilon-b) - \frac{B}{1-a}\cdot(1-\epsilon-b) \end{aligned} \quad (8)$$

Solving this for a (assuming $1-\epsilon-b \neq 0$) and then solving for e yields

$$\begin{aligned} a &= \frac{B}{A+B} \\ e &= \frac{a-b}{1-\epsilon-b} = \frac{\left(\frac{B}{A+B}-b\right)}{1-\epsilon-b} \end{aligned} \quad (9)$$

2.9.2 Estimation of one labelling efficiency $e(t)$ and transcript-specific ratios $\rho_g(t)$

Given the measurement t of one time point in one compartment and sample. The above procedure yields one estimate $e_g(t)$ of the labelling efficiency for each 3'UTR g. Some of these estimates will be highly unreliable due to a small number of reads assigned to the 3'UTRs or a slow turnover of the underlying RNA transcripts. Particularly at early time points, the number of newly synthesised transcripts and, correspondingly, the number of labelled reads is very low. Conversions in reads derived from these transcripts will mainly be introduced by sequencing errors, resulting in very small efficiency estimates. Due to statistical noise, the number of observed conversions can even be smaller than the number of conversions expected by sequencing errors, thereby returning negative efficiency estimates.

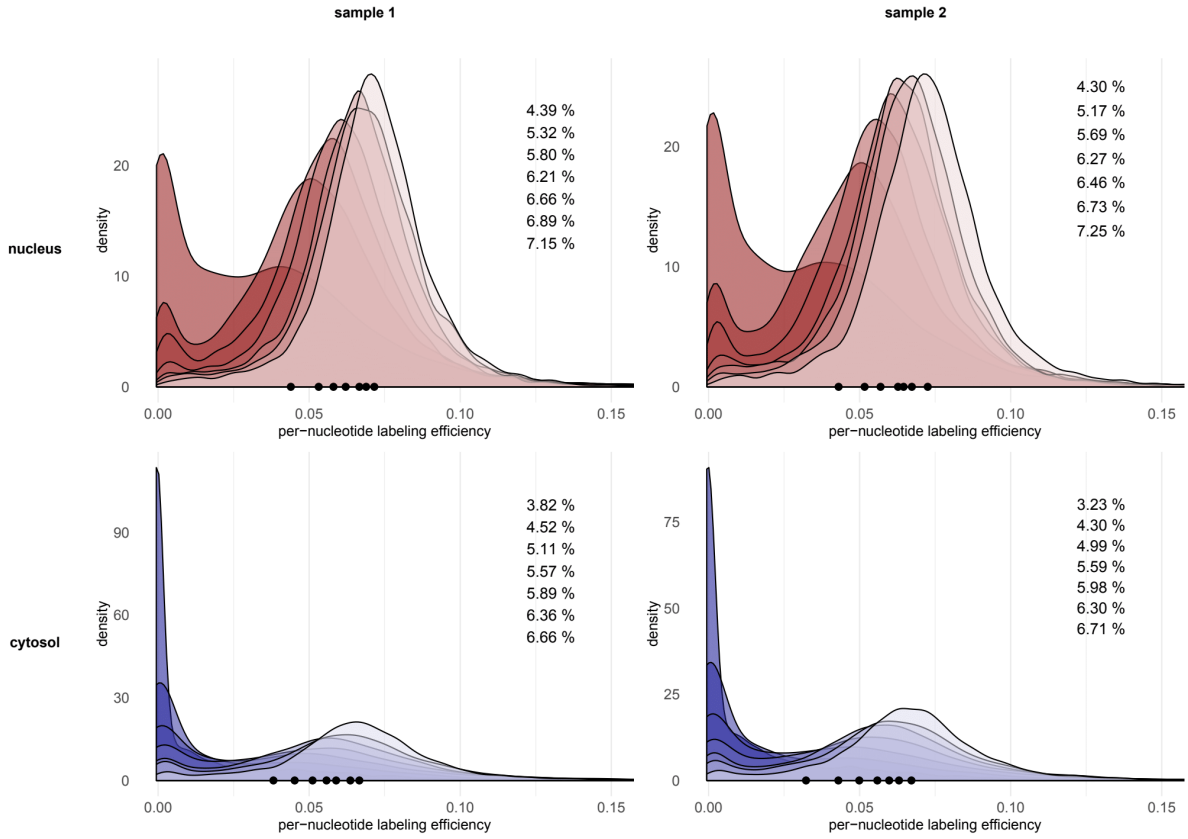
To stabilize the estimation results, those 3'UTRs with less than 100 assigned reads in a measurement t are henceforth excluded. 3'UTRs with slow turnover transcripts and underestimated efficiencies show up as a peak at values close to 0 in the distribution of the $e_g(t)$ values, which is especially visible for early time points (Figure 10). Therefore, all 3'UTRs with $e_g(t)$ values smaller than 0.01 are discarded from the next steps. The global estimate $e(t)$

of the labelling efficiency for measurement t is defined as the median of the remaining values $e_g(t)$ (Figure 10).

The EM algorithm is subsequently repeated for each 3'UTR g and measurement t, however, with the labelling efficiency fixed to $e(t)$ to obtain the transcript-specific estimates of the newly synthesised RNA ratio, $\rho_g(t)$. $\rho_g(t)$ is initialized by the value that has been obtained from the previous application of the EM algorithm.

The labelling efficiency estimates of the 3'UTR regions and non-3'UTR coverage peaks are similar. Yet, the efficiencies of the non-3'UTR peaks are lower in the beginning and higher at the end of the time series compared to the 3'UTR regions. It is hard to find the source of this deviation, as many factors can play a role. These include: the loss of uniquely mapped, labelled reads in the nuclear fraction; unidentified and thereby uncorrected editing sites; gene expression changes during the labelling experiment; potentially different 4sU incorporation efficiencies during RNA synthesis between RNA Polymerases I, II and III. Recapitulate that mRNA and thereby the 3'UTRs are transcribed by RNA Polymerase II, whereas the non-3'UTR peaks can originate from both unannotated mRNAs and other RNA species. Therefore, some of the non-3'UTR peaks may be transcribed by RNA polymerases I or III.

A 3'UTRs



B non-3'UTR peaks

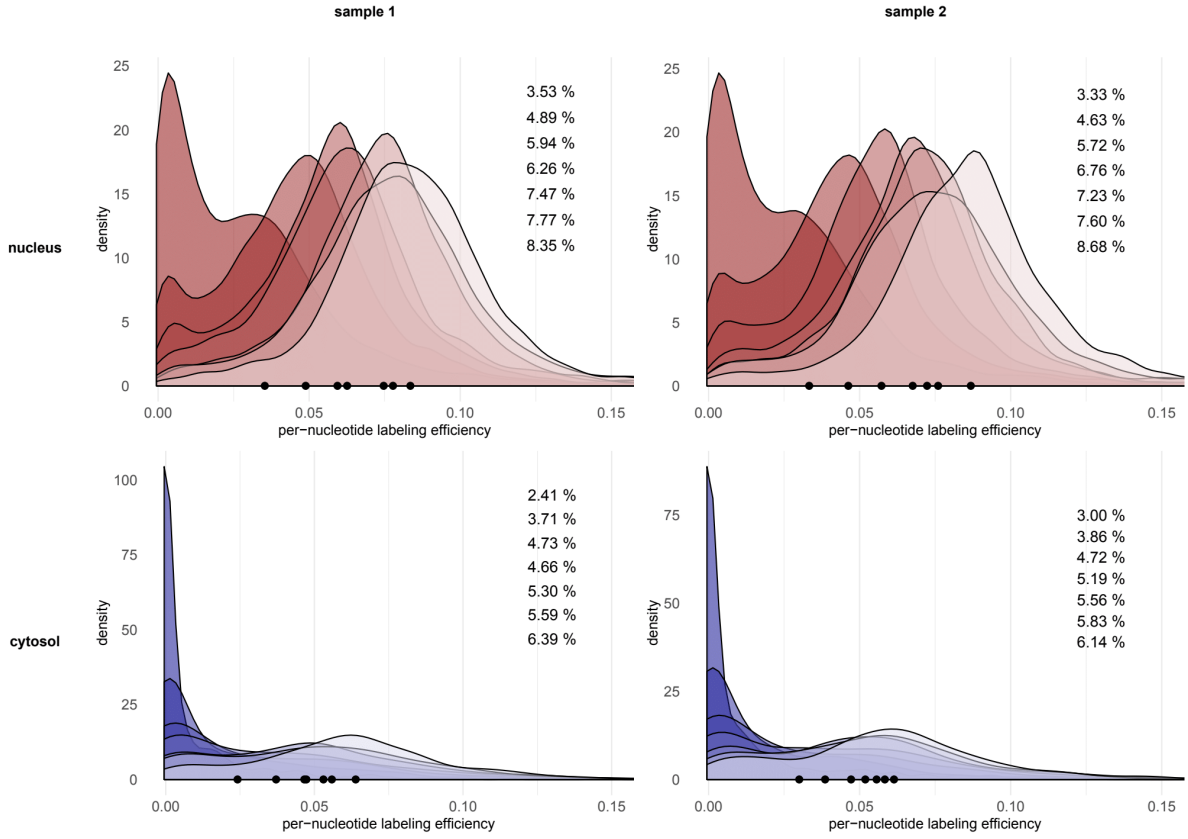


Figure 10 Distribution of labelling efficiency estimates $e_g(t)$. Estimates are shown for annotated 3'UTRs in (A) and non-3'UTR coverage peaks in (B). Labelling efficiencies were estimated per measurement t of one time point in one compartment and sample, and for each 3'UTR or non-3'UTR peak g. Only annotated 3'UTRs or non-3'UTR peaks with at least 100 assigned reads in measurement t were included in the estimation. The distributions of estimates $e_g(t)$ are displayed in the plots. The darkest, background-most distribution refers to the earliest time point (15min) and the brightest, foreground-most distribution to the latest time point (180min), respectively. The median of each distribution was chosen as the final labelling efficiency estimate $e(t)$ for the respective read identity and measurement t. Medians are indicated by the black dots on the x-axis and written at the right of the plot panel, ordered by time points. Estimates $e_g(t)$ smaller than 0.01 were discarded from the median calculations to exclude 3'UTRs or non-3'UTR peaks with highly unreliable estimates due to slow RNA metabolism and thereby small number of newly synthesised, labelled transcripts.

2.10 Variance stabilizing transformation

The variance of the newly synthesised RNA ratios is computed as the variance between experimental sample 1 and 2, calculated for each annotated 3'UTR and measurement time point. It can be observed that the variance of the newly synthesised RNA ratios is dependent on the ratio itself. This scenario leads to an unequal weighting of the ratios measured across the time series during parameter fitting. Yet, all time point measurements are equally informative and should therefore be equally weighted.

Under the assumption that the new/total ratios ρ with total read counts J follow a Binomial distribution $\mathrm{Bin}(J, \rho)$, an $\arcsin(\sqrt{\rho})$ transformation is applied to stabilize variance. It can be shown that the distribution of the $\arcsin(\sqrt{\rho})$ transformed ratios approximates to a normal distribution with variance $\sigma^2 = \frac{1}{4 \cdot J}$ (Bromiley & Thacker, 2002). Thereby, the variance is dependent only on the total read counts, which can be accounted for in the parameter fitting procedure (Figure 11).

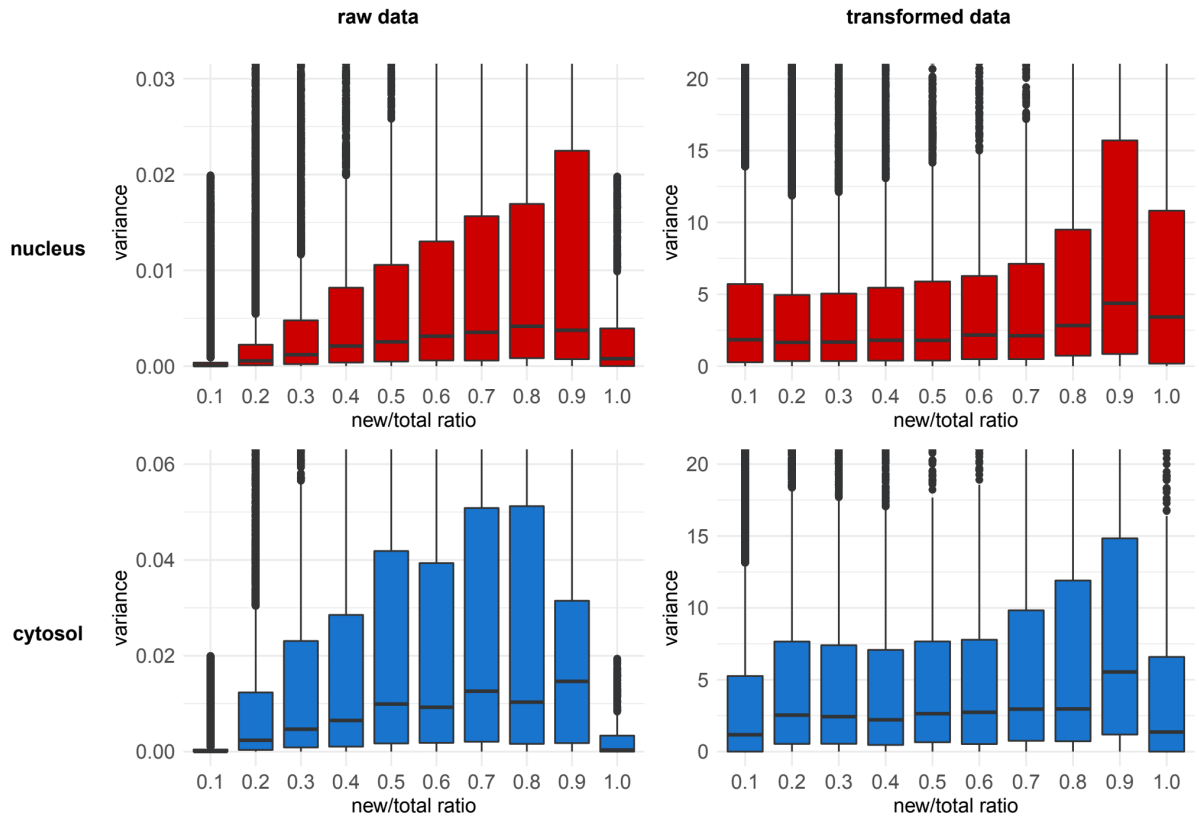


Figure 11 Variance stabilizing transformation. The panel shows the variance of the newly synthesised RNA ratios ρ (new/total ratio) for the untransformed ratios ρ (raw data) and the variance stabilized ratios $arcsin(\sqrt{\rho})$ (transformed data). Data is shown for the nuclear and cytosolic fraction. Variances were calculated for each annotated 3'UTR and time point measurement as the variance between experimental sample 1 and 2. The distribution of the variance dependent on the average ratio between sample 1 and 2 is displayed. Average ratios are binned into intervals of size 0.1. The variance stabilizing transformation of ρ is expected to transform the variance to $\sigma^2 = \frac{1}{4 \cdot J}$ (J: total read counts). To validate if the variance stabilizing transformation was successful, the transformed data's variances were additionally multiplied with $4 \cdot$ the average total read counts between sample 1 and 2.

2.11 Parameter estimation

Parameter fitting was performed on the newly synthesised RNA ratios after variance stabilizing transformation of the ratios (2.10 Variance stabilizing transformation). It can be assumed that the transformed ratios are normally distributed with variance $\sigma^2 = \frac{1}{4 \cdot J}$ (J: total read counts) (Bromiley & Thacker, 2002).

Fix a 3'UTR (or a coverage peak). Given a measurement t of one time point in one compartment, let $q(t)$ be the observed transformed ratio of the 3'UTR in measurement t. Let

$\hat{q}(t,\Theta)$ be the true ratio which is defined by the RNA metabolic parameters $\Theta = (\tau + \nu, \lambda)$. Then, the probability of observing the transformed ratio $q(t)$ can be expressed as

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

or expressed as log probability

$$\log\big(P(q(t) \mid \Theta)\big) = \log\frac{\sqrt{4J}}{\sqrt{2\pi}} - 2J\big(q(t) - \hat{q}(t,\Theta)\big)^2$$

The summed log probability over all measurements t, $\sum_t \log\big(P(q(t) \mid \Theta)\big)$, has to be maximized with respect to parameters Θ. The first term, $\log\frac{\sqrt{4J}}{\sqrt{2\pi}}$, can be omitted since it is independent of Θ.

Instead of fitting the parameters $\Theta = (\tau + \nu, \lambda)$ jointly with the nuclear and cytosolic fraction measurements, parameter estimation was performed sequentially as the nuclear data appeared to be less noisy than the cytosolic data. Possible reasons for this observation are the lower coverage and the slower increase in the ratio of newly synthesised RNA in the cytosolic data. In detail, the nuclear removal rate $\tau + \nu$ was estimated first with the data obtained from the nuclear fraction. The cytosolic degradation rate λ was fit subsequently with the data from the cytosolic fraction, plugging in the nuclear removal rate estimate which is also needed to define the ratio of newly synthesised RNA in the cytosol. Thereby, more nuclear removal rates could be estimated reliably.

Parameter estimation was executed in 3 steps. First, a seed for the subsequent MCMC run was determined. Second, MCMC sampling of the parameter space was performed to derive a credibility interval for the parameter estimate. Third, the final estimates were calculated. The single steps are described in more detail in the following.

We noticed that, without an appropriate seed, the MCMC sampling sometimes got stuck in low-probability regions of the parameter space. Therefore, a global optimization approach (differential evolution algorithm) was applied with a maximum of 500 iterations to determine a seed for the MCMC chain. Global optimization was not used to calculate the final parameter estimates as it is computationally expensive and thus unfeasibly slow. The nuclear removal rate

seed was determined first and plugged into the global optimization run for the cytosolic decay rate seed. Subsequently, the MCMC chain was run to sample the parameter space of the nuclear removal or cytosolic degradation rate, respectively. The credibility intervals of the parameter estimates were defined as the 95% percentile of the MCMC samples. Lastly, Nelder-Mead optimization was applied to obtain the final estimate of the nuclear removal or cytosolic degradation rate. The Nelder-Mead optimization was seeded with the median of the MCMC-sampled parameter space. Parameter estimation was executed separately for sample 1 and 2, and estimates were averaged over both samples afterwards.

2.12 Estimation reliability criteria

Multiple filtering criteria were applied to categorize the parameter estimates as reliable and unreliable. Estimates are considered reliable if they meet the following conditions: (1) The genomic region for which the parameters were estimated is detectable (it has a minimum of 1 read assigned in each measurement, as well as a minimum of 30 reads assigned on average per compartment and sample). (2) The difference between the estimates of sample 1 and 2, divided by the average estimate between sample 1 and 2, is not larger than 0.33. (3) The difference between the estimate and its credibility interval border, divided by the estimate, is not larger than 0.3 (see 2.11 Parameter estimation). This holds for both the upper and lower border of the credibility interval in both samples. (4) The relative expression level does not increase or decrease by more than 0.0025[min^{-1}] or -0.0025[min^{-1}], respectively (for details, see below in this section). This holds for both samples, but specifically for the nuclear expression for the nuclear removal rate estimates, respectively for the cytosolic expression for the cytosolic degradation rate estimates. (5) The R-squared value of the fit is equal to or higher than 0.4. This holds for both samples.

Additionally, the cytosolic degradation rate estimates are considered reliable only if the respective nuclear removal rate estimates are, as the latter are plugged in to fit the cytosolic degradation rates. To check that the model assumption of steady-state metabolism is not violated, the 3'UTRs relative expression levels were computed for each time point measurement. First, a robust read count average was calculated by omitting the bottom and top 25% of 3'UTRs w.r.t total read counts and taking the average read count of the remaining 50% 3'UTRs in the time point measurement. We choose this robust average over the simple read

count average, as the latter is biased if the gene expression of a few highly abundant transcripts changes. These calculations were based on all 3'UTRs with at least 1 read count in each time point measurement of a sample. Second, each 3'UTR's read counts were divided by the robust average read count to obtain the relative expression levels. A simple linear regression was performed on the relative expression levels of each 3'UTR along the time series, separately for sample 1 and 2 as well as the nuclear and cytosolic compartment. The regression level slopes served as indicators of whether expression levels were constant.

2.13 Cyt/nuc ratio estimation

2.13.1 Cyt/nuc ratio estimation based on the nuclear degradation rate ν

The problem when estimating the ratio between the total amount of RNA in the cytosol and the total amount of RNA in the nucleus (the cyt/nuc ratio) is that the measurements in each compartment measure RNA levels only up to a global, multiplicative constant. However, we exploit our knowledge of the RNA dynamics to derive an estimate of the cyt/nuc ratio. Given a set G of annotated 3'UTRs g for which total read counts u_{nuc}^{g}, u_{cyt}^{g} in the nuclear and cytosolic compartment were measured. Let $\hat{u}_{nuc}, \hat{u}_{cyt}$ be the robust averages of the nuclear and cytosolic total counts of set G. Then, normalised expression levels n^{g} and c^{g} in the nucleus and the cytosol can be calculated from the total counts as:

$$n^{g} = \frac{u_{nuc}^{g}}{\hat{u}_{nuc}}, c^{g} = \frac{u_{cyt}^{g}}{\hat{u}_{cyt}}$$

Let N_g^{total}, C_g^{total} be the number of transcripts associated to 3'UTR g in the nucleus and cytosol of a single cell. Furthermore, let $\mu^{g}, \tau^{g}, \nu^{g}, \lambda^{g}$ be the parameters for mRNA synthesis, nuclear export, nuclear degradation and cytosolic degradation of the corresponding mRNA transcripts. Under steady-state conditions, N_g^{total}, C_g^{total} can be expressed as combinations of the mRNA metabolic parameters:

$$N_g^{total} = \frac{\mu^{g}}{\tau^{g}+\nu^{g}}, C_g^{total} = \frac{\mu^{g}}{\tau^{g}+\nu^{g}} \cdot \frac{\tau^{g}}{\lambda^{g}}$$

The numbers of cellular transcripts are proportional to the respective normalised expression levels up to global constants r_{nuc}, r_{cyt}:

$$N_g^{total} = r_{nuc} \cdot n^{g},\ C_g^{total} = r_{cyt} \cdot c^{g}$$

Similarly, the ratio of the numbers of cellular transcripts is proportional to the ratio of the normalised expression levels up to a global constant r:

$$\frac{C_g^{total}}{N_g^{total}} = \frac{r_{cyt}}{r_{nuc}} \cdot \frac{c^g}{n^g} = r \cdot \frac{c^g}{n^g}$$

Again, these ratios can be expressed as a combination of mRNA metabolic parameters:

$$\begin{aligned} \frac{C_g^{total}}{N_g^{total}} &= r \cdot \frac{c^g}{n^g} &&= \frac{\mu^g}{\tau^g + \nu^g} \cdot \frac{\tau^g}{\lambda^g} \cdot \frac{\tau^g + \nu^g}{\mu^g} \\ &= && \frac{\tau^g}{\lambda^g} \end{aligned}$$

Let $H \in G$ be a subset of annotated 3'UTRs h for which $\tau^h + \nu^h$ and λ^h can reliably be estimated. The link between cellular transcript count ratios, normalised expression level ratios and metabolic parameters leaves us with a combination of the measured and/or estimated values $n^h, c^h, \tau^h + \nu^h, \lambda^h$ and unknown parameters ν^h, r:

$$\begin{aligned} r \cdot \frac{c^h}{n^h} &= \frac{\tau^h}{\lambda^h} \\ r \cdot \frac{c^h}{n^h} \cdot \lambda^h + \nu^h &= \tau^h + \nu^h \\ \nu^h &= \tau^h + \nu^h - r \cdot \frac{c^h}{n^h} \cdot \lambda^h \end{aligned}$$

Since the nuclear degradation rate cannot be negative, it holds that $\nu^h \geq 0$. Then, it can be expected that the distribution of ν^h calculated for all $h \in H$ for a specific guess of r is mostly non-negative, unless the guess for r is unreasonably large. Thereby, an upper limit of r can be estimated.

2.13.2 Cyt/nuc ratio estimation based on the spike-in RNAs

For each RNA sample (one time point measurement in one compartment of one sample), a robust read count average was calculated by omitting the bottom and top 25% of 3'UTRs w.r.t total read counts, and taking the average read count of the remaining 50% 3'UTR. These calculations were based on all 3'UTRs which had at least 1 read count in each time point measurement of a sample. Each RNA sample's robust average was then normalised with the spike-in read counts in the same sample (separately for the unlabelled RNA spike-ins and labelled RNA spike-ins). The cyt/nuc ratio estimate of a time point measurement was then

calculated as the normalised read count average of the cytosolic fraction divided by the normalised read count average of the corresponding nuclear fraction.

2.14 Selection of ER-translated transcripts

The mRNA transcripts which encode for endoplasmatic reticulum components are likely translated at the ER itself. Therefore, 3'UTRs of ER-translated transcripts were selected based on the following 6 Gene Ontology terms (Ashburner et al., 2000; The Gene Ontology Consortium, 2021):

GOCC_INTEGRAL_COMPONENT_OF_CYTOPLASMIC_SIDE_OF_ENDOPLASMIC_RETICULUM_MEMBRANE
GOCC_INTRINSIC_COMPONENT_OF_ENDOPLASMIC_RETICULUM_MEMBRANE
GOCC_LUMENAL_SIDE_OF_ENDOPLASMIC_RETICULUM_MEMBRANE
GOCC_PERINUCLEAR_ENDOPLASMIC_RETICULUM
GOCC_ROUGH_ENDOPLASMIC_RETICULUM
GOCC_ROUGH_ENDOPLASMIC_RETICULUM_MEMBRANE

2.15 Software

As described in 2.2 Read alignment, read alignment and alignment quality filtering were performed with slamdunk (version 0.3.0).

The model was implemented in Python (version 3.4.3) (van Rossum & de Boer, 1991) and R (version 3.4.4) (R Core Team, 2018). Additional non-base Python modules used were pysam (version 0.13) (Li et al., 2009) (also see https://github.com/pysam-developers/pysam) for processing SAM/BAM formatted files, numpy (version 1.9.2) (Harris et al., 2020) and scipy (version 0.15.1) (Virtanen et al., 2020) for mathematical operations. Additional non-base R packages used were mcmc (version 0.9-7) (Geyer & Johnson, 2020) for MCMC sampling, DEoptim (version 2.2-5) (Ardia et al., 2020) for global optimization, foreach (version 1.5.1) (Microsoft and Steve Weston, 2020) and doParallel (version 1.0.16) (Microsoft Corporation and Steve Weston, 2020) for parallelization of the parameter estimation process, and stringr (version 1.4.0) (Wickham, 2019) for string manipulation.

Figures were created with the help of R (version 3.4.4) (R Core Team, 2018) and the following non-base R packages: ggplot2 (version 3.3.2) (Wickham, 2016), reshape2 (version 1.4.4) (Wickham, 2007), gridExtra (version 2.3) (Auguie, 2017), lemon (version 0.4.5) (Edwards, 2020), LSD (version 4.1-0) (Schwalb et al., 2020), scales (version 1.1.1) (Wickham & Seidel, 2020), RColorBrewer (version 1.1-2) (Neuwirth, 2014).

3 Results

3.1 Data processing

Reads were mapped to the human genome hg19 and quality filtered using slamdunk (Neumann et al., 2019), yielding between 8.5M and 24.4M uniquely mapped reads per sample (2.2 Read alignment). A decrease in the percentage of uniquely mapped reads could be observed for the nuclear fraction over time, probably due to the introduced labelling conversions. This read loss could not be recovered (2.3 Recovery of multi-mapped and unmapped reads). Next, nucleotide positions listed as known SNP positions (Sherry et al., 2001) or suspicious of post-transcriptional editing were masked (2.7 Mismatch statistics and correction for SNP and editing sites), since these are conversions not introduced by 4sU labelling.

A fraction of 61% to 78% of all uniquely mapped reads per sample could be assigned to 3’UTRs of known genes as annotated in the UCSC Table Browser (Karolchik et al., 2004) (2.4 Assigning reads to annotated 3’UTRs). Overlapping 3’UTRs were merged considering strand orientation, resulting in 61834 3'UTRs. We define a 3’UTR as detectable if it (1) has a minimum of 1 read assigned in each measurement (i.e., each time point measured in each compartment and sample), as well as (2) a minimum of 30 reads assigned on average across a time series (i.e., per compartment and sample). 8119 of the 61834 annotated 3’UTRs were detectable.

The read distribution within a 3’UTR often showed multiple, distinct peaks potentially derived from alternative polyA sites. Further peaks could be observed outside of annotated 3’UTRs, probably originating from alternative priming sites to A-rich sequences. To identify these dense read clusters both within and beyond annotated 3’UTRs, peak calling was performed (2.5 Peak calling). We apply the same criteria used for the 3’UTR regions to define a peak as detectable. This yielded 10,170 detectable peaks within 3’UTR regions (3’UTR peaks), and 1651 detectable peaks outside of annotated 3’UTR regions (non-3’UTR peaks). The high number of undetectable peaks can be attributed to intronic peaks only being found in the nucleus, low expression levels, poor priming strength to internal priming sites, or mapping artefacts.

The subsequent modelling relies on the assumption that all transcripts corresponding to either an aggregated 3’UTR region or an individual peak share the same metabolism.

3.2 A dynamic model of mRNA metabolism

The life cycle of a transcript starts with its synthesis by RNA polymerases. It is followed by either nuclear export into the cytosolic compartment after transcriptional processing or, eventually, nucleoplasmic degradation. Exported transcripts are finally degraded, irrespective of the function they fulfil in the cytosol. Fix a population of RNA transcripts which share the same metabolism, say, all transcripts derived from the same gene. The dynamics of the nuclear and cytosolic subpopulations are modelled quantitatively by introducing 4 metabolic parameters: synthesis rate μ (defined as the production rate of polyadenylated RNA transcripts), nuclear export rate τ, nuclear degradation rate ν and cytosolic degradation rate λ (Figure 12). These 4 rates are assumed to be constant over time, as the experiments were conducted under constant, optimal growth conditions.

Let $N = N(t)$ and $C = C(t)$ denote the time courses of nuclear respectively cytosolic RNA abundances (which are averaged over all cells in the bulk measurement). N and C can be viewed as continuous numbers and modeled by a simple two-compartment ODE system:

$$\dot{N} = \mu - (\tau + \nu)N$$

$$\dot{C} = \tau N - \lambda C$$

Under the assumption of steady-state (all metabolic rates are constant over time), the system has a closed form solution with steady-state abundances $N^{total} = \frac{\mu}{\tau+\nu}$ and $C^{total} = \frac{\mu}{\tau+\nu} \cdot \frac{\tau}{\lambda}$.

In contrast to RNA labelling methods that involve separation of newly synthesised and pre-existing RNA fractions, SLAM seq delivers an unbiased measurement of the relative abundance of newly synthesised RNA ${N^{new}(t)}/{N^{total}}$ and ${C^{new}(t)}/{C^{total}}$. Under steady state conditions, the relative abundances can be expressed as:

$$\frac{N^{new}(t)}{N^{total}} = 1 - e^{-(\tau+\nu)t}$$

$$\frac{C^{new}(t)}{C^{total}} = 1 - \frac{\lambda e^{-(\tau+\nu)t} - (\tau+\nu)e^{-\lambda t}}{\lambda - (\tau+\nu)}$$

The sum of nuclear export and degradation rate, $\tau + \nu$, will henceforth be called nuclear removal rate. $\frac{N^{new}(t)}{N^{total}}$ is defined in terms of the nuclear removal rate, and $\frac{C^{new}(t)}{C^{total}}$ is defined by both the nuclear removal rate and the cytosolic degradation rate λ. Thereby, the relative abundances of newly synthesised RNA observed in the experimental data serves to fit these two parameters.

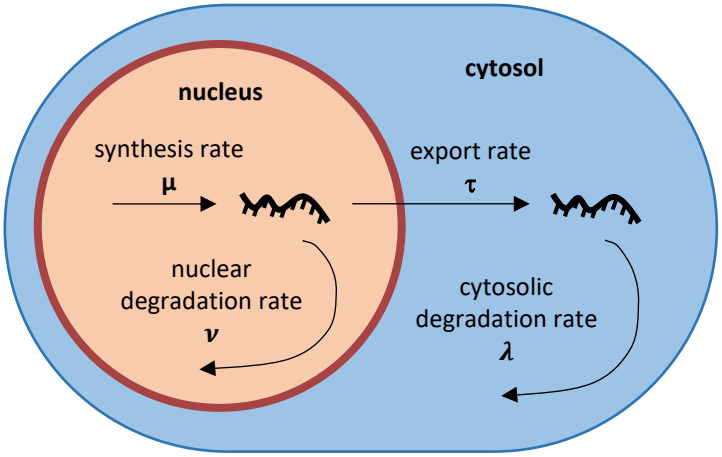


Figure 12 Two-compartment model of mRNA metabolism. mRNA metabolism is described by four processes which are characterized by their respective metabolic parameters: mRNA synthesis in the nucleus with synthesis rate μ, nuclear degradation with rate ν, export from the nucleus to the cytosol with rate τ, and cytosolic degradation with rate λ.

3.3 Estimation of metabolic parameters

The efficiency by which a uridine within a newly synthesised RNA transcript is labelled is in the range of 5% (Herzog et al., 2017). Consequently, not all of the reads that originate from newly synthesised transcripts will display a labelling conversion and will therefore be misclassified as pre-existing. Sequencing errors mimicking or masking labelling conversions can introduce further misclassifications. There have been attempts to account for these errors by Jürges et al. (2018), yet their algorithm relies on a minor fraction of sequencing reads containing a high number of conversions. This read selection may be heavily biased towards high labelling efficiencies by posttranscriptional modifications. Further, they assume that labelling efficiency is constant over time. In order to account for these possible problems, read counts were corrected for nucleotide positions suspected of being editing sites (2.7 Mismatch statistics and correction for SNP and editing sites). Afterwards, labelling efficiencies were estimated alongside the newly synthesised RNA ratios in each time point measurement (2.8 Estimation of labelling efficiency and newly synthesised RNA ratio).

With increasing duration of 4sU labelling, the labelling efficiency increases from around 4.3% to 7.2% (3’UTRs) respectively 3.4% to 8.5% (non-3’UTR peaks) in the nuclear fraction, and 3.5% to 6.7% (3’UTRs) respectively 2.7% to 6.3% (non-3’UTR peaks) in the cytosolic fraction (Figure 10).

Parameters were fit based on the variance stabilized data using a least-squares approach (Figure 13, Figure 14) (2.10 Variance stabilizing transformation, 2.11 Parameter estimation). Fitting was performed using standard numerical optimization. Credibility bounds were obtained from MCMC sampling (Figure 13B) (2.11 Parameter estimation, 2.12 Estimation reliability criteria). Stringent quality criteria were applied to filter for transcripts whose metabolic rates could be fitted reliably (2.12 Estimation reliability criteria). Out of the 8119 detectable 3'UTRs, 1297 had a reliable estimate for the nuclear removal rate (Figure 15). 251 3'UTRs furthermore had reliable estimates for the cytosolic degradation rate (Figure 15).

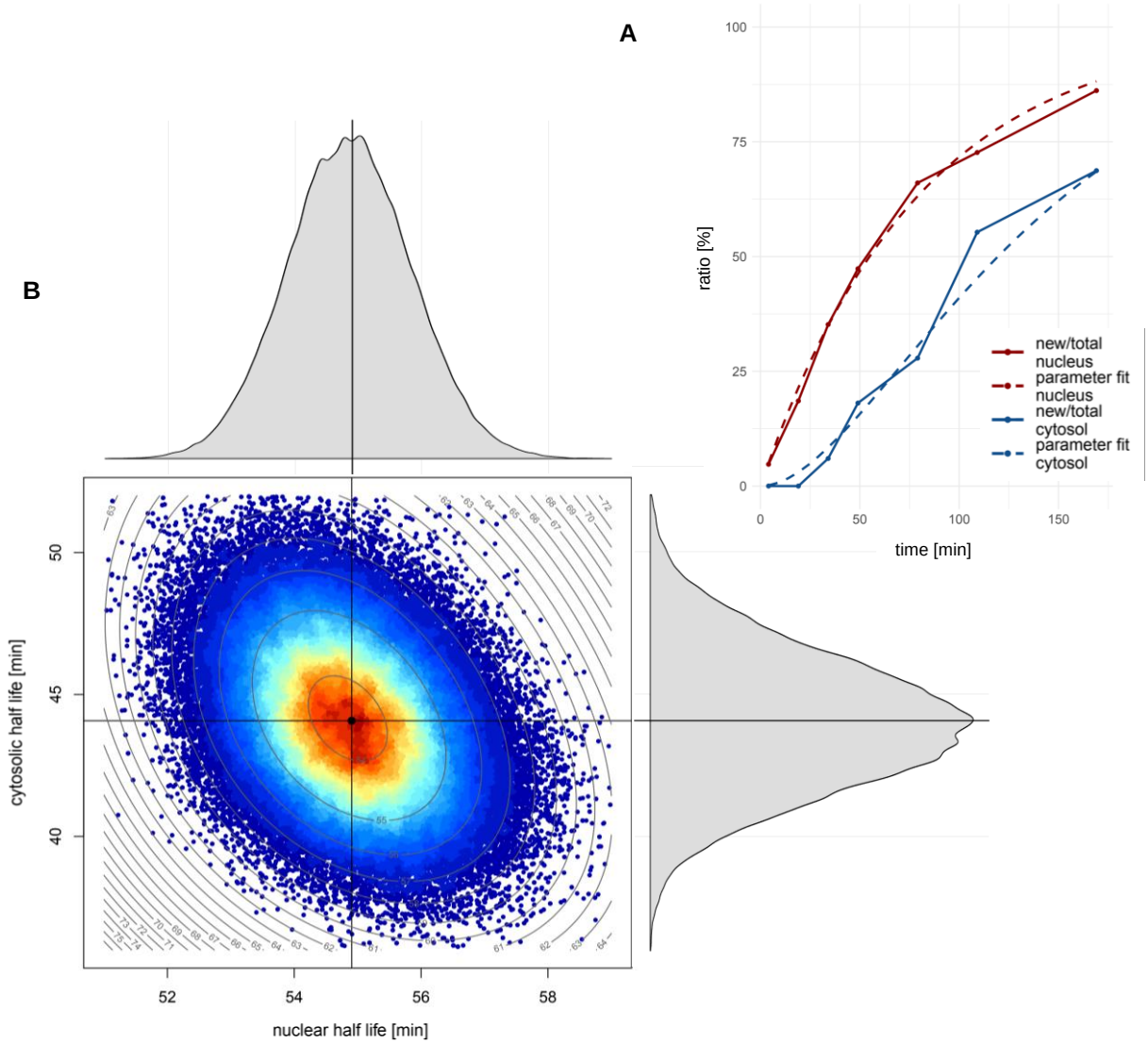


Figure 13 Estimation of RNA metabolic parameters. Results are shown for one exemplary 3'UTR with the ID uc002kcm.3_utr3_0_0_chr17_79876145_r in sample 2. This 3'UTR has reliable parameter estimates. **(A)** Comparison of newly synthesised RNA ratios as inferred from the data (new/total) and as predicted by the parameter estimates for the nuclear removal rate $\tau + \nu$ and cytosolic degradation rate λ (parameter fit). **(B)** MCMC sample of the two-dimensional parameter space of the nuclear half life (defined by the nuclear removal rate $\tau + \nu$) and cytosolic half life (defined by cytosolic degradation rate λ). Note that parameter estimation is not performed in a 2D manner but sequentially for the nuclear removal rate and cytosolic degradation rate; 2D MCMC sampling was performed for visualization purposes only.

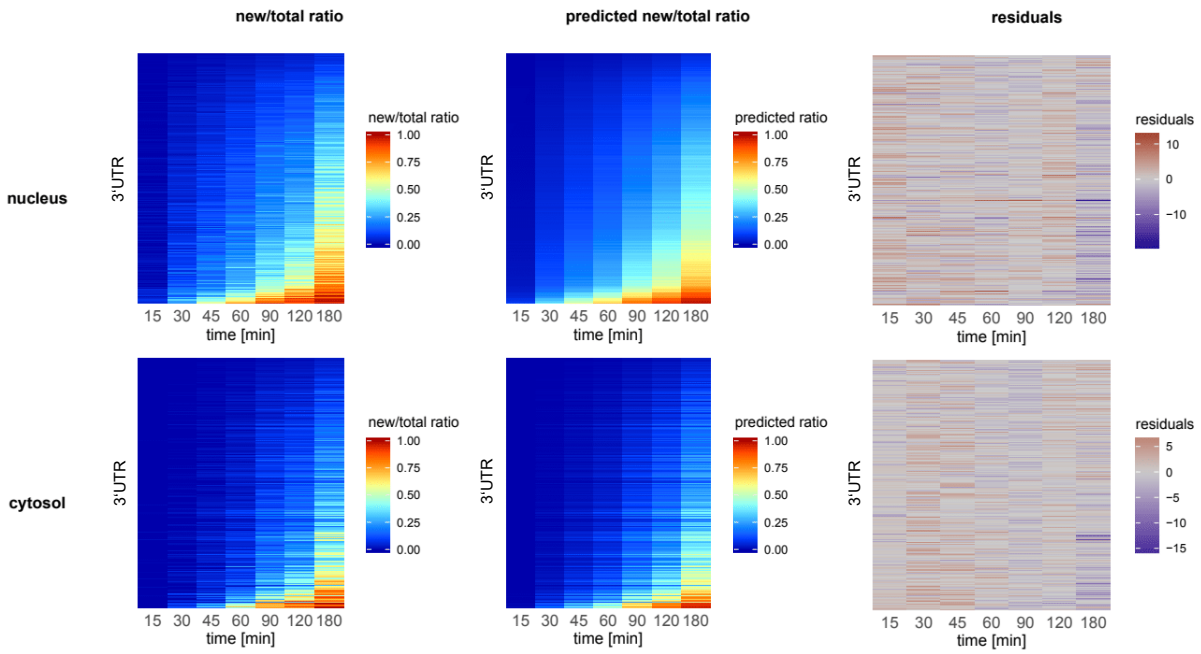


Figure 14 Comparison of the newly synthesised RNA ratios inferred from the data, predicted according to parameter estimates, and respective residuals. Ratios inferred from the data: new/total ratio. Ratios predicted according to parameter estimates: predicted new/total ratio. Data is shown for 3'UTRs with reliable parameter estimates for both the nuclear removal rate $\tau + \nu$ and the cytosolic degradation rate λ, and for the data of sample 2. The panels' rows represent the 3'UTRs and are ordered by decreasing nuclear half life estimate for all panels. Panel columns represent time point measurements. The residuals shown compare to the residuals as calculated during the parameter estimation process. A value above 0 indicates that the predicted ratio exceeds, and a value below 0 that the predicted ratio falls below the inferred ratio. In detail, shown residuals are calculated as $\sqrt{2J} \cdot \left(\hat{q}(t, \Theta) - q(t)\right)$ (see also 2.11 Parameter estimation).

3.4 Nuclear half life is substantially higher than cytosolic half life

From the estimates of RNA metabolic parameters, nuclear half life is calculated as $^{\ln 2}/_{(\tau + \nu)}$ and cytosolic half life as $^{\ln 2}/_{\lambda}$. Based on the 3'UTRs for which parameters were estimated reliably (nucleus: 1297, cytosol: 251), nuclear half life has a median of 291min (IQR 311) and cytosolic half life of 45min (IQR 32) (Figure 15). The ratios between nuclear and cytosolic half life have a median of 5.76 (IQR 6.88) (Figure 16). Half life estimates correlated well between both samples (Supplemental Figure 2, Supplemental Figure 3). These findings unexpectedly imply that nuclear export is much slower than cytosolic degradation. Notably, these

observations don't change if the parameter estimates of all detectable 3'UTRs are considered (Figure 15, Figure 16).

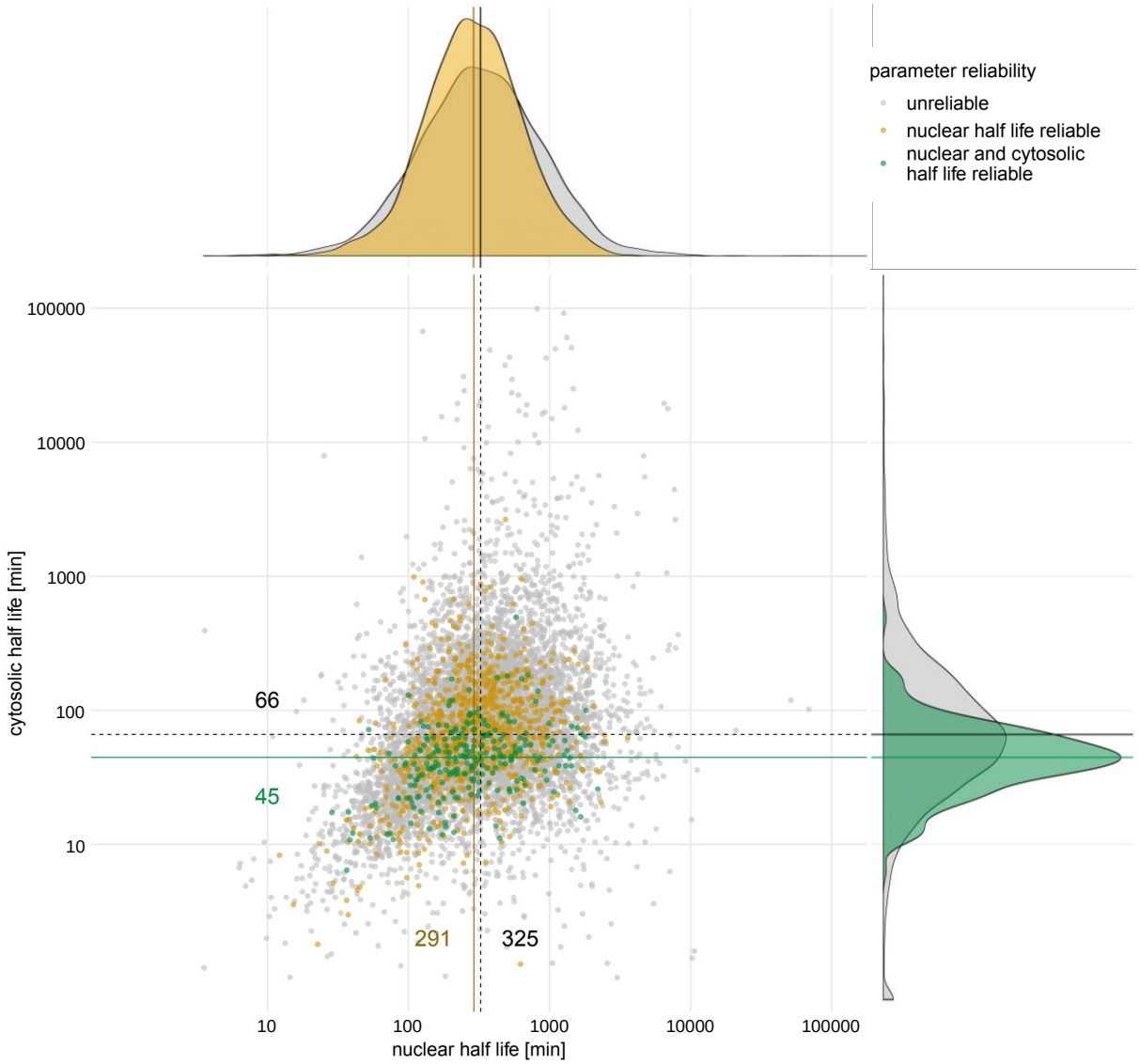


Figure 15 Nuclear and cytosolic half life estimates. Nuclear and cytosolic RNA half life estimates were calculated for detectable 3'UTR regions. Half lifes are derived from the nuclear removal and cytosolic degradation rate estimates as $^{\ln 2}/_{(\tau + \nu)}$ and $^{\ln 2}/_{\lambda}$, respectively. Stringent quality criteria were applied to score the reliability of the estimates (2.12 Estimation reliability criteria). The cytosolic estimates are always unreliable if the nuclear estimates are, as the nuclear removal rates were fit first and then plugged in to fit the cytosolic degradation rate subsequently (2.11 Parameter estimation). The vertical and horizontal lines with (rounded) number labels indicate the respective half life medians (black: of all detectable 3'UTRs, yellow: of 3'UTRs with reliable nuclear half life estimate, green: of 3'UTRs with reliable cytosolic half life estimate). The plot was trimmed, excluding 168 3'UTRs with unreliable estimates and 1 3'UTR with reliable nuclear half life estimate. The full plot can be found in Supplemental Figure 6.

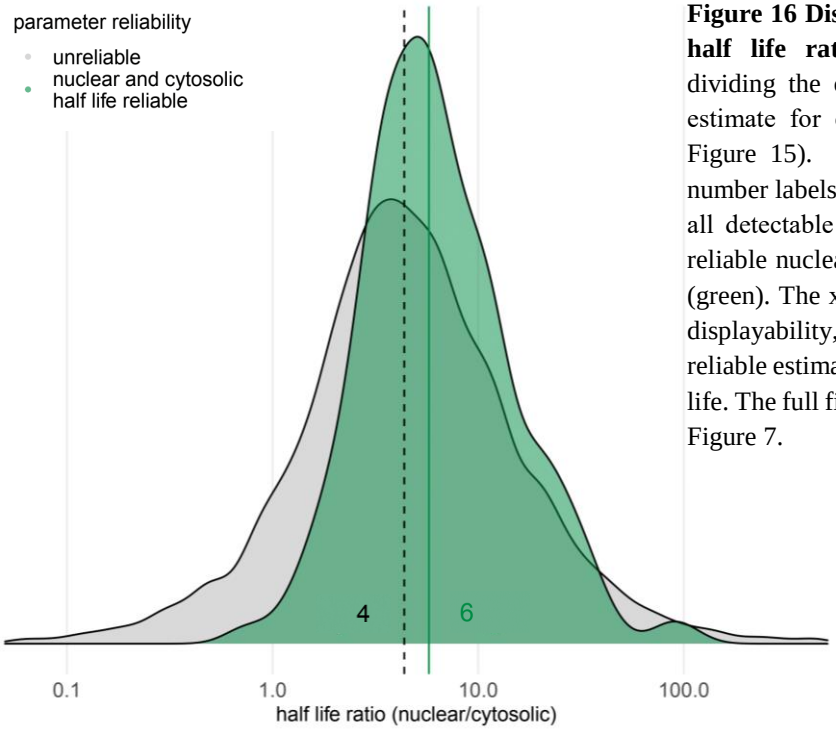


Figure 16 Distribution of nuclear by cytosolic half life ratios. Ratios were calculated by dividing the cytosolic by the nuclear half life estimate for each detectable 3'UTR (see also Figure 15). The vertical lines and (rounded) number labels indicate the median ratio based on all detectable 3'UTRs (black) or 3'UTRs with reliable nuclear and cytosolic half life estimates (green). The x-axis was trimmed for the sake of displayability, removing 189 data points without reliable estimates for the nuclear or cytosolic half life. The full figure can be found in Supplemental Figure 7.

Naively, the metabolism of a transcript's exon should equal the metabolism of its 3'UTR, ignoring transcript isoforms in which the exon is spliced out or different polyA sites are used. Accordingly, a decent correlation can be expected between the half life estimates of non-3'UTR peaks within exonic regions and their corresponding 3'UTRs. For the 62 detectable exonic peaks which also had a detectable 3'UTR for their gene (considering only genes with exactly one 3'UTR annotation), a good correlation can be confirmed (Figure 17A).

In contrast, introns are usually spliced out in the nucleus and not exported to the cytosol. Looking at cellular compartments separately, the relative abundance of intronic peaks which are detectable in the nucleus is higher than it is in the cytosol. Yet there are still 581 (sample 1) and 494 (sample 2) detectable intronic peaks in the cytosol (detectable intronic peaks in the nucleus: 12,453 and 9787, respectively) (Supplemental Table 5). These peaks may arise from intron retention events or unknown alternative exons and non-coding RNA species. For the 103 intronic peaks detectable in both cellular compartments and samples and have a detectable 3'UTR for their gene (considering only genes with exactly one 3'UTR annotation), the half life estimates do not correlate between introns and 3'UTRs (Figure 17B). Most of these intronic peaks probably originate from unknown alternative exons of transcript isoforms that are metabolised differently than most other isoforms or non-coding RNA species which are encoded within the intronic regions.

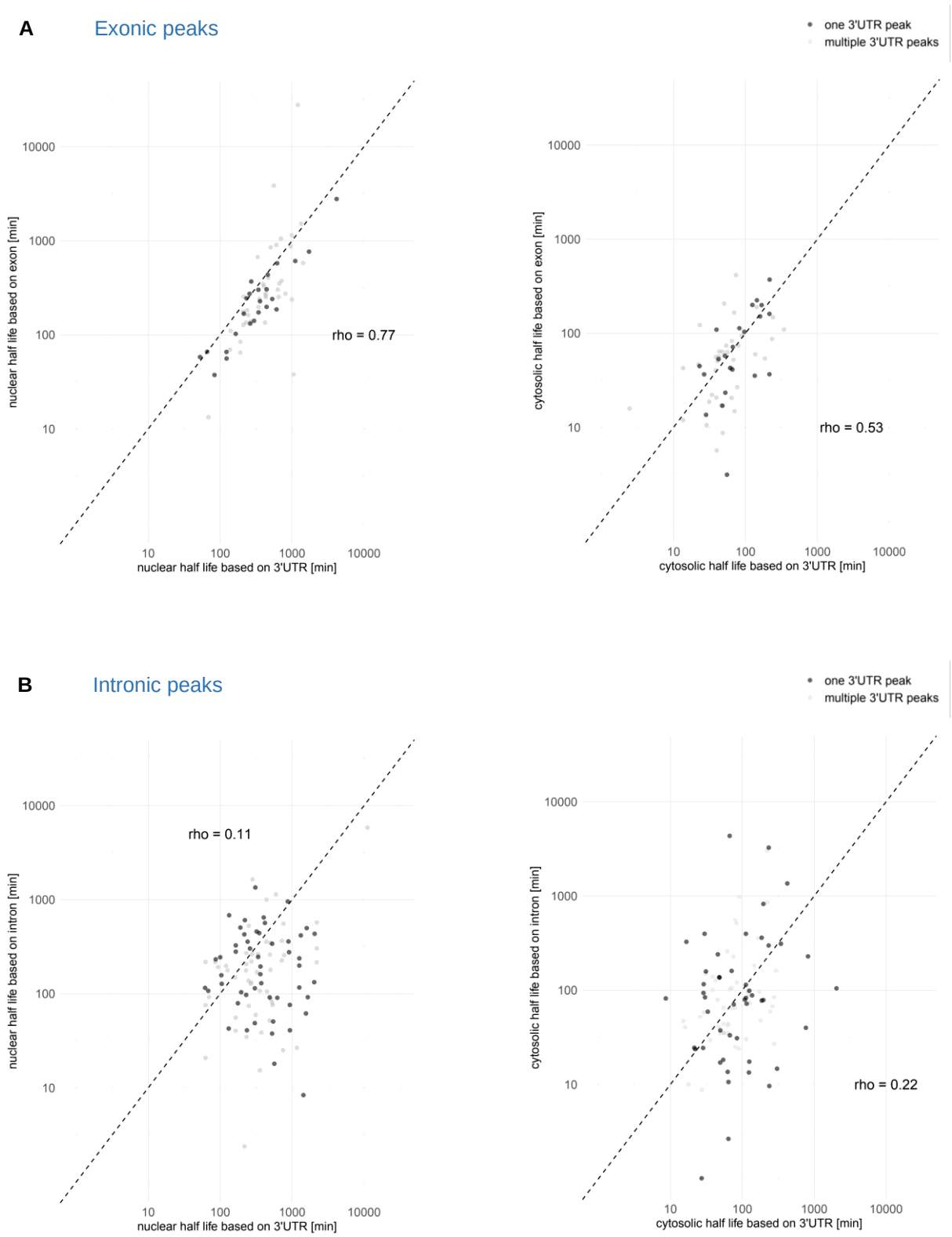

A
Exonic peaks
one 3'UTR peak
multiple 3'UTR peaks
nuclear half life based on exon [min]
nuclear half life based on 3'UTR [min]
rho = 0.77
cytosolic half life based on exon [min]
cytosolic half life based on 3'UTR [min]
rho = 0.53
B
Intronic peaks
one 3'UTR peak
multiple 3'UTR peaks
nuclear half life based on intron [min]
nuclear half life based on 3'UTR [min]
rho = 0.11
cytosolic half life based on intron [min]
cytosolic half life based on 3'UTR [min]
rho = 0.22
10
100
1000
10000

Figure 17 Correlation of half lifes estimates between 3'UTRs and exonic or intronic regions of the same gene. Correlations between 3'UTRs and exons are shown in (A), correlations between 3'UTRs and introns are shown in (B), separately for nuclear half lives at the left and cytosolic half lifes at the right. Spearman's rho correlation coefficients are labelled in each plot. The dashed lines indicate the diagonal. Only genes with exactly one 3'UTR annotation, and only detectable 3'UTRs and non-3'UTR peaks were considered. 62 pairs of 3'UTRs and non-3'UTR peaks of the same gene were found for exonic peaks (58 genes with 1 exonic peak, 2 genes with 2 exonic peaks) and 102 for intronic peaks (86 genes with 1 peak, 5 genes with 2 peaks, 2 genes with 3 peaks). Data points are coloured by whether the 3'UTR harbours one or multiple 3'UTR peaks (3'UTR peaks with a minimum of 1 read count in each time point measurement of each compartment and sample, but irrespective of the peaks' mean coverage), suspicious of alterative polyA site usage. The plots were trimmed, excluding the following data points: cytosolic half lifes of 3'UTRs vs exons: 2 data points with one 3'UTR peak and 1 data point with multiple 3'UTR peaks; cytosolic half lifes of 3'UTRs vs introns: 3 data points with one 3'UTR peak and 3 data points with multiple 3'UTR peaks.

3.5 mRNA is more abundant in the nucleus than in the cytosol

As the nuclear half life estimates are much higher than the cytosolic half life estimates, it can be expected that most RNA transcripts are more abundant in the nucleus than in the cytosol. To verify this hypothesis, the ratio of total RNA molecules in the cytosol versus the nucleus (cyt/nuc ratio) was estimated using two different approaches: (1) calculation of the nuclear degradation rate ν in dependence of the cyt/nuc ratio, (2) comparing the spike-in read counts in the nuclear and cytosolic fraction (2.13 Cyt/nuc ratio estimation).

For all 3'UTRs with reliable half life estimates, the nuclear degradation rate estimates were computed as a function of the cyt/nuc ratio. As nuclear degradation rates are by definition equal to or larger than 0, so should be the majority of their estimates. Cyt/nuc ratios for which more than 75% of the nuclear degradation rate estimates were negative were therefore considered unreasonable, which yields an upper bound for the cyt/nuc ratio of < 0.225 (Figure 18). When the cyt/nuc ratio was calculated based on the spike-in reads in the nuclear and cytosolic fraction, the estimates were consistently smaller than 1 (0.276 - 0.697) for each time point measurement in both samples (Figure 19).

The estimates obtained with approach (2) are higher than obtained with approach (1). One potential source of bias can be the loss of uniquely mapped, labelled reads in the nuclear fraction (3.1 Data processing, Figure 3). This may lead to an under-estimation of the nuclear removal rate and thereby an under-estimation of the cyt/nuc ratio calculated with approach (1). Simulations on the impact of a labelled read loss in the nuclear fraction showed that this effect

cannot completely explain the observed deviations in the cyt/nuc ratio estimates (data not shown). Another source of bias could be a difference in the efficiency of RNA extraction for the nuclear and cytosolic compartments. As the cyt/nuc ratio estimates are still smaller than 1 throughout the two different approaches, it can be concluded that the transcripts of the analysed 3'UTRs are more abundant in the nucleus than in the cytosol and spend most of their lifetime in the nucleus.

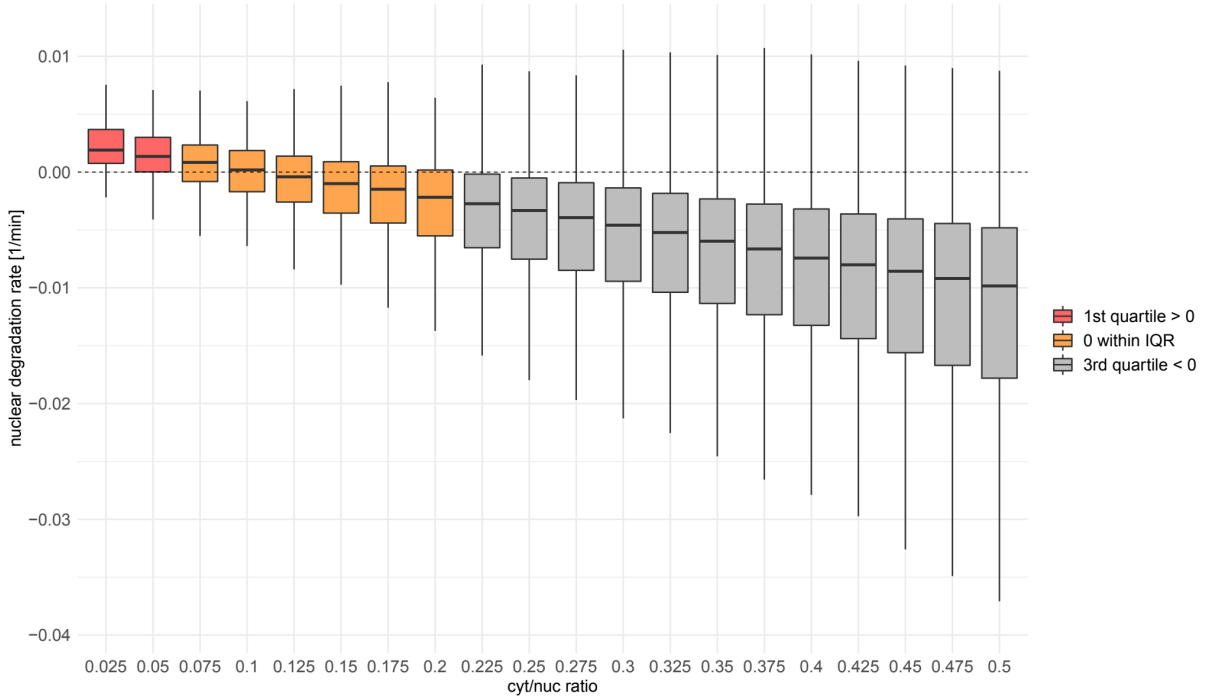


Figure 18 Nuclear degradation rate ν in dependence of the cyt/nuc ratio. The calculation of nuclear degradation rates ν in dependence of the cyt/nuc ratio r is described in 2.13 Cyt/nuc ratio estimation. Only 3'UTRs with reliable estimates for both the nuclear removal rate and the cytosolic degradation rate were evaluated. Ratios are considered unreasonable if more than 75% of the corresponding nuclear degradation rates are negative (3rd quartile < 0).

3.6 3'UTR isoforms differ in their metabolism

Peak calling identified 10170 peaks within annotated 3'UTRs which were detectable in both samples. These peaks may either correspond to 3'UTR isoforms or internal priming sites. To distinguish between potential isoforms safely, these peaks were additionally filtered. Only peaks assigned to exactly one annotated 3'UTR and with a minimum distance of 100

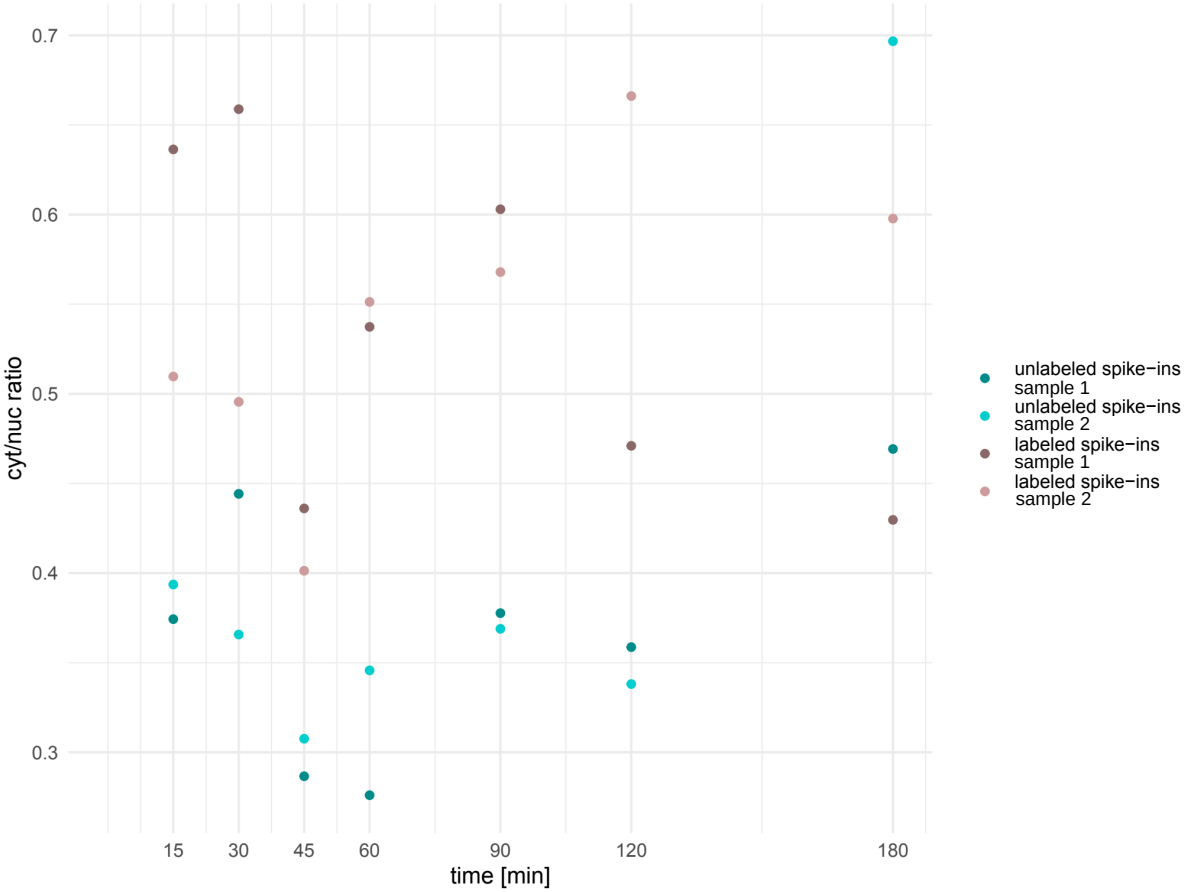


Figure 19 Cyt/nuc ratios calculated from the spike-in RNAs. Each RNA sample was added a fixed amount of spike-in RNA directly after cellular fractionation. Equal amounts of both unlabelled and labelled spike-ins were added to each sample. Cyt/nuc ratios were calculated separately with the unlabelled and labelled spike-in RNAs, in each time point measurement and sample. The cyt/nuc ratio calculation based on the spike-in RNAs is described in 2.13 Cyt/nuc ratio estimation.

nucleotides to their direct neighbours were kept, resulting in a collection of 2987 peaks. Half life estimates correlated well between sample 1 and 2 (Supplemental Figure 4, Supplemental Figure 5).

118 annotated 3'UTRs harboured more than 1 of these filtered peaks. To investigate whether these peaks generally derived from internal priming sites or distinct 3'UTR isoforms, the log2 fold changes of half life estimates were calculated between (1) all pairs of peaks within the same 3'UTR, and (2) sample 1 and 2. Since internal priming sites can be considered technical replicates, the log2 fold changes in (1) can be expected to resemble the log2 fold changes in (2) if most of the peaks are derived from internal priming sites, but higher than in (2) if the peaks measure distinct 3'UTR isoforms which adds biological variation. In this calculation, it was not distinguished between peaks with reliable or unreliable estimates as only 2 out of the 118 3'UTRs had a peak pair with two reliable nuclear estimates.

The median log2 fold change of nuclear half lives was much larger for the comparison of potential isoforms than for the comparison of technical replicates (Figure 20). It can be concluded from the results that a large fraction of these peaks measure 3'UTR isoforms that differ in their RNA metabolism. In contrast, the median log2 fold changes of cytosolic half lives only differed slightly (Figure 20). Since the noise is much higher for the cytosolic than for the nuclear fraction (compare Supplemental Figure 2 and Supplemental Figure 3), biological effects may be masked here.

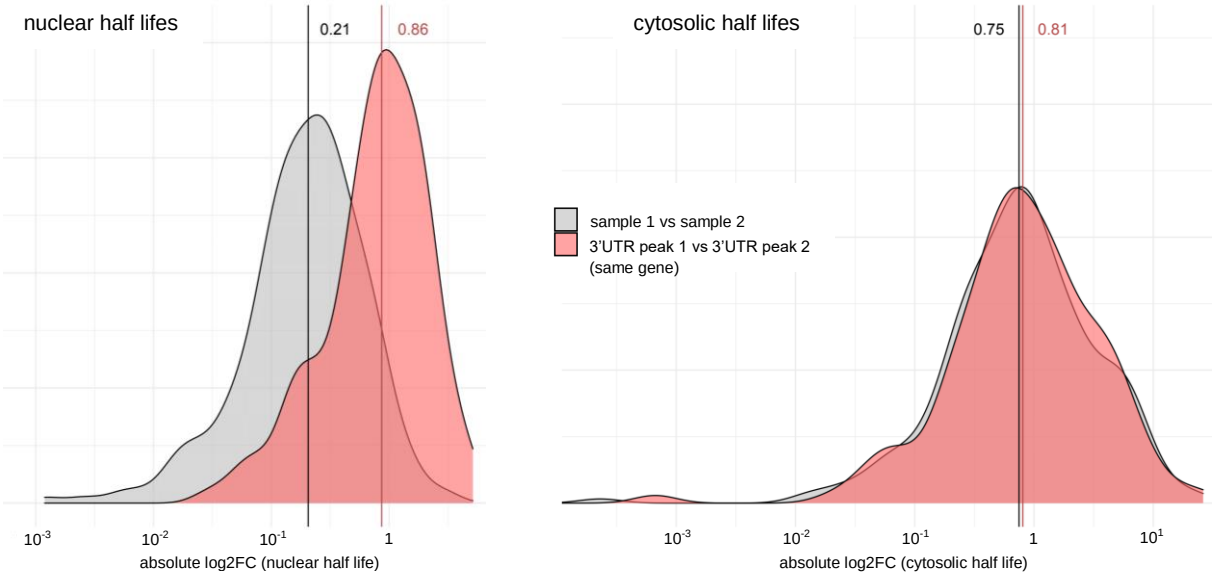


Figure 20 Differences in half life estimates of potential 3'UTR isoforms. Some 3'UTRs harboured at least two detectable 3'UTR peaks with a minimum distance of 100 nucleotides, representing potential 3'UTR isoforms (242 peaks from 118 3'UTRs). The absolute log2 fold change of nuclear and respectively cytosolic half lives was computed for all pairwise comparisons of potential isoform irrespective of the half life estimate reliability (red distributions, at the left for nuclear and at the right for cytosolic half lives). For comparison, absolute log2 fold changes were calculated for an isoform's estimates between sample 1 and 2 (grey distributions). The red and black lines with corresponding text labels indicate the medians. For the comparison of cytosolic half lives between sample 1 and 2, 5 data points dropped out since their value was 0 and could not be displayed with the log-scaled x-axis.

3.7 Export of labelling-induced genes is substantially faster than that of most other genes

The model of RNA metabolism assumes steady-state conditions, i.e. metabolic parameters and expression levels are constant over time. Linear regression was applied on expression levels to identify transcripts that strongly deviate from this assumption (2.12 Estimation reliability criteria). Among the transcripts violating this assumption, one outstanding group was a collection of transcripts that showed low expression at the beginning and end of the measured time series but very high expression shortly after labelling onset (Figure 21). This group of transcripts is therefore referred to as 'Supernova' genes. The nuclear expression decreases from and the cytosolic expression level increases to its maximum within 120min after start of the labelling. This hints to a very fast export of these transcripts, potentially via an export mechanism distinct from those of the majority of all other transcripts. According to UniProt (The UniProt Consortium, 2021), many of the Supernova genes are annotated as involved in stress response, growth or cell cycle.

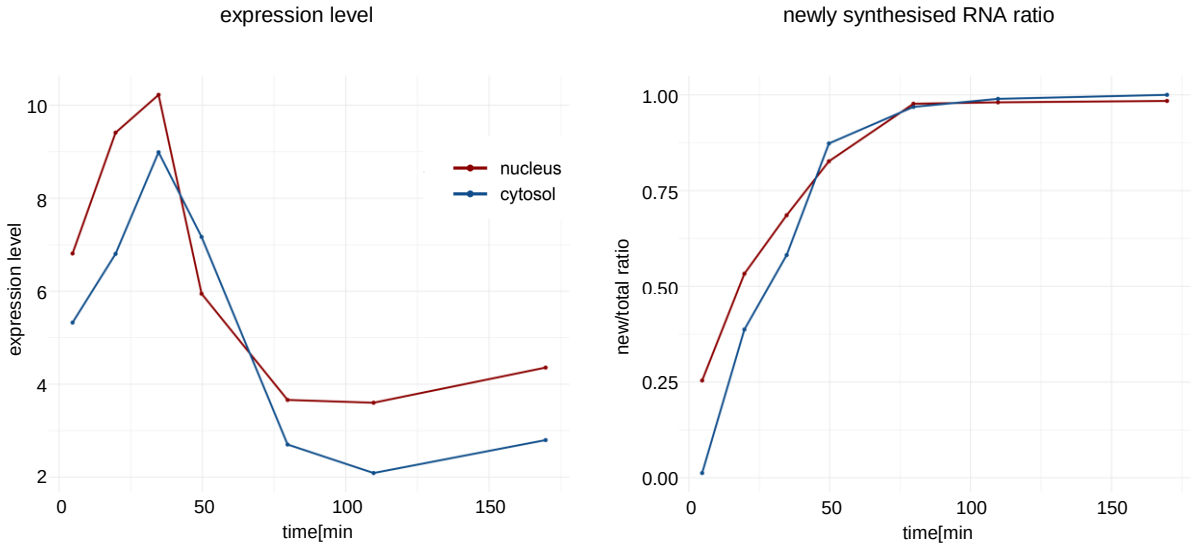


Figure 21 Gene expression and newly synthesised RNA ratio of a 'Supernova' gene. Gene expression and newly synthesised RNA ratio measured over time are shown for the exemplary 'Supernova' gene SGK1 (3'UTR id uc011ect.2_utr3_0_0_chr6_134490384_r) in sample 1. Gene expression was calculated relative to a robust average across all 3'UTRs' total read counts (see 2.12 Estimation reliability criteria).

4 Conclusion

4.1 Wrap-up

RNA metabolic labelling is a powerful technique to measure newly synthesised (labelled) and pre-existing (unlabelled) RNA. Nucleoside analogues which deviate from the cells' native nucleosides, but are similar enough to be incorporated into nascent RNA, are added to the cells. This way, newly synthesised RNA can be tagged. The most commonly used nucleoside analogue is 4-thiouridine (4sU). Previously, 4sU-labelled RNA was separated from unlabelled RNA by the biotinylation of the 4sU residues and subsequent pulldown with streptavidin beads. The recently developed method SLAM seq instead alkylates the 4sU residues, which then causes T>C conversions at the alkylated 4sU positions during reverse transcription. Newly synthesised and pre-existing RNA can then be distinguished by the observed T>C conversions in the read alignment. SLAM seq outperforms former methods as the data acquired is less noisy and the experimental approach is less laborious.

The analysis of RNA metabolic labelling data produced with SLAM seq remains challenging. First, an appropriate model of RNA metabolism is needed to estimate RNA metabolic rates. Second, uncertainties and biases introduced by experimental or biological effects need to be considered. Especially the efficiency by which a 4sU residue is incorporated into a nascent RNA transcript is usually smaller than 10%. Therefore, only a fraction of the newly synthesised transcripts will actually be labelled.

Here, I presented a novel, dynamic model of mRNA metabolism which discriminates between nuclear and cytosolic mRNA. The two-compartment modelling allows to investigate mRNA export in addition to cytosolic degradation. This represents a major step forward in the research on mRNA metabolism. RNA editing events can distort the labelling conversion counts; therefore, potential RNA editing sites are defined as nucleotide positions with high conversion rates in the control samples and masked in the subsequent analysis steps. The 4sU incorporation efficiency (labelling efficiency) is estimated together with the ratio of newly synthesised RNA using an EM algorithm. The algorithm models the observed T>C conversions with a binomial distribution and is parametrised with the ratio of newly synthesised RNA, the labelling efficiency, and the sequencing errors. Importantly, labelling efficiencies were calculated for

each time point separately, which revealed that the efficiency increases over time as opposed to the assumption of constant efficiency in Jürges et al. (2018). The parameter fitting procedure of RNA metabolic rates is applied to a variance-stabilizing transform of the newly synthesised RNA ratios and it assumes an approximate normal distribution of the transformed ratios. For each genomic region analysed, the fitting procedure returns one estimate of the nuclear removal rate (the sum of nuclear degradation and nuclear export rate) and one of the cytosolic degradation rate. Nuclear and cytosolic half lives can directly be calculated from the metabolic rates.

We conducted a SLAM seq metabolic labelling experiment combined with cellular fractionation and 3'-sequencing, and applied the model to estimate the nuclear and cytosolic half lives of the measured 3'UTRs. Half lives had a high correlation between both samples measured, which shows that the estimation procedure is robust. The median nuclear half life was considerably higher than the median cytosolic half life. As nuclear degradation is expected to be negligibly small, we conclude from these results that nuclear export is substantially slower than cytosolic degradation. Consequently, the amount of mRNA transcripts in the nucleus has to be much higher than in the cytosol. This hypothesis was confirmed by two approaches to estimate the ratio of cytosolic by nuclear total mRNA (cyt/nuc ratio). All cyt/nuc ratios calculated (1) based on the metabolic rate estimates and (2) based on the observed spike-in abundances were smaller than 1, i.e. predicting a lower RNA abundance in the cytosol.

4.2 Potential biases and comparison with literature

Intuitively, mRNA export is expected to be much faster than cytosolic degradation, as the cytosolic RNA is needed to produce the proteins. Yet, our results imply that export is much slower than cytosolic decay. There are multiple sources of bias that could contribute to an underestimation of nuclear export rates: Loss of reads with many labelling conversions during the alignment; pulldown of the endoplasmatic reticulum together with the nucleus during cell fractionation and, thereby, also mRNA translated at the rough ER; disassembly of the nucleus during mitosis, which leads to leakage of nuclear RNA to the cytoplasmic fraction, and reassembly of the nucleus after mitosis, which leads to the inclusion of cytosolic RNA into the nucleus.

The loss of reads during the alignment process was observed only in the nucleus, where the fraction of uniquely mapped reads decreased from around 80% to 70% with increasing labelling time. This read loss is likely caused by the labelling conversions in the newly synthesised RNA transcripts that complicate the mapping. The mapping quality in the cytosol was stable, which is explained by the lower newly synthesised RNA ratio in this fraction.

Attempts to recover the reads lost during the alignment step failed. The proportion of reads mapping to multiple genomic loci (multimappers) and the proportion of unmapped reads did not show any systematic trend over labelling time. In contrast to that, the fraction of ambiguously mapped reads steadily increased. Ambiguously aligned reads include both the multimappers and reads with poor alignment quality. It can be concluded that reads are lost because they harbour too many conversions to pass the aligner's quality filtering criteria. We decided to keep the alignment results as they were, as a relaxation of the quality filtering thresholds can lead to a higher number of misaligned reads.

If labelled reads are lost, the newly synthesised RNA ratio will be underestimated. Accordingly, the nuclear vanishing rate will be underestimated as well. Still, the read loss is expected to only have a minor impact on the estimates for the following two reasons: (1) All time points of the time series are equally weighted during parameter fitting; the less-biased measurements of the early time points stabilize the more-biased measurements of the later time points. (2) The labelling efficiency is also underestimated since especially the highly labelled reads are lost; this leads to an overestimation of the newly synthesised RNA ratio, which buffers the effect of the read loss.

The mapping process may be improved by masking SNP and potential editing positions within the reference genome. Paired-end sequencing can help in many ways: As the two read mates cover more nucleotide positions than a single read, it is easier to find their correct position in the genome. Conversions observed at a position covered by both mates can much better be classified: As the probability of a sequencing error at the exact same position in both mates is very low, a conversion is very likely an RNA editing event (for the control samples) or a labelling event (for the metabolic labelling samples) if it occurs in both mates.

The endoplasmatic reticulum could, at least partially, be pulled down together with the nucleus during the cell fractionation. mRNAs translated at the rough ER will therefore be measured in

the nuclear and not in the cytosolic fraction. As these mRNAs will mostly be pre-existing, unlabelled transcripts, the nuclear fraction will be biased towards a higher amount of unlabelled transcripts and the cytosolic fraction towards a higher amount of labelled transcripts. Consequently, the nuclear vanishing rate will be underestimated and the cytosolic decay rate will be overestimated for the affected mRNAs. To check whether ER pulldown affects our data, the half lives of 3'UTRs whose transcripts are probably translated at the ER were compared with the half lives of all detectable 3'UTRs (Figure 22) (2.14 Selection of ER-translated transcripts). Indeed, the ER-translated transcripts' nuclear half life median was larger (i.e., nuclear vanishing rate lower), and the cytosolic half life median was shorter (i.e. cytosolic degradation rate higher) compared to all detectable 3'UTRs. Whether this observation is due to ER pulldown or has a biological reason may be worth studying in future experiments.

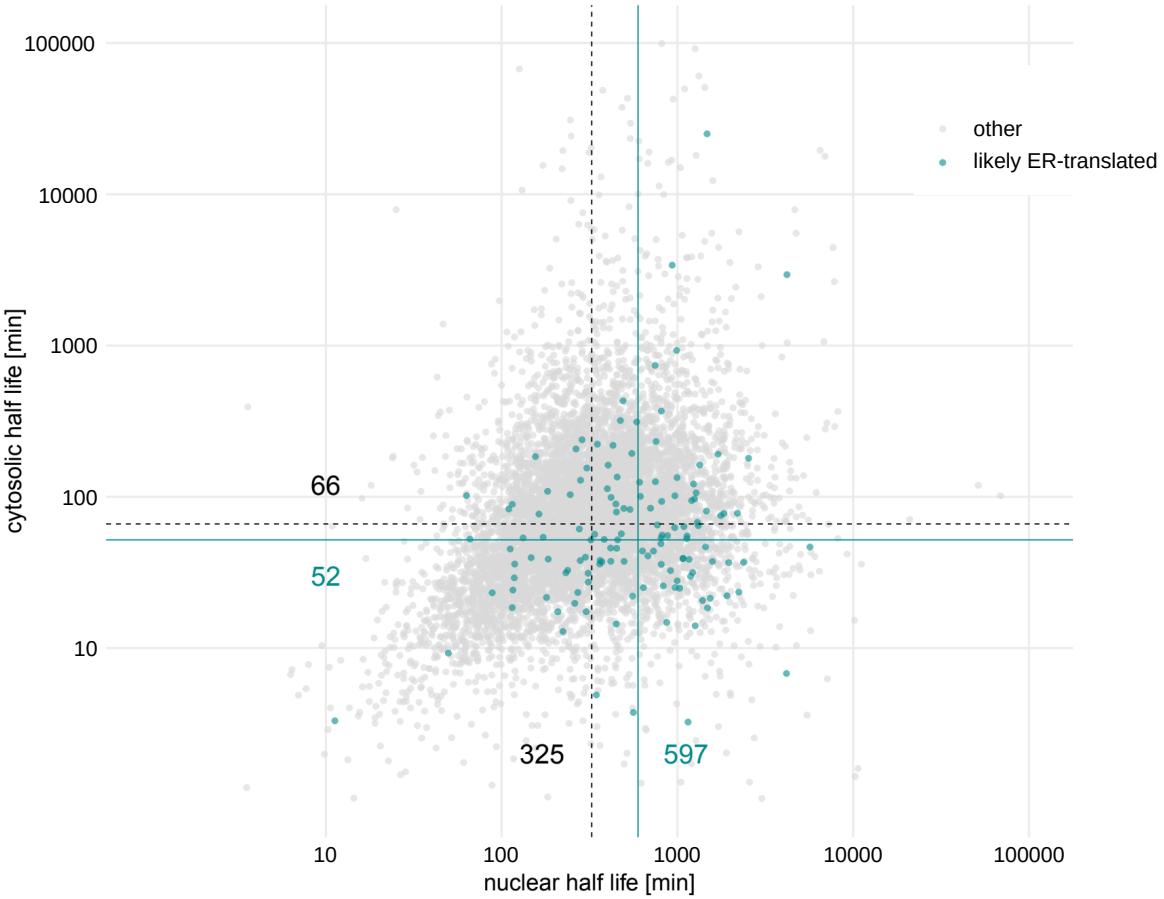


Figure 22 Half lifes of transcripts likely translated at the ER. A subset of detectable 3'UTRs was selected whose transcripts are likely translated at the ER (data points in blue) (2.14 Selection of ER-translated transcripts). All other 3'UTRs are coloured in grey. The blue and black lines with corresponding text labels indicate the median half lives of the subset and all detectable 3'UTRs, respectively. The plot was trimmed, excluding 169 non-subset data points. A plot of all detectable 3'UTRs half lives can be found in Supplemental Figure 6.

The dis- and reassembly of the nucleus during mitosis can affect the measurements in two ways. Keep in mind that the newly synthesised RNA ratio is higher in the nucleus than in the cytosol as the RNA is transcribed in the nucleus. Disassembly of the nucleus causes leakage of nuclear RNA to the cytosol, which increases the newly synthesised RNA ratio in the cytosol. Theoretically, this can lead to an overestimation of both the nuclear vanishing and cytosolic degradation rate. Practically, as the parameter fitting of the nuclear vanishing rate is based on the nuclear measurements only, it only affects the cytosolic decay estimates. Reassembly causes inclusion of cytosolic RNA into the nucleus and lowers the newly synthesised RNA ratio of the nucleus. Thereby, the nuclear vanishing rate is underestimated. Disentangling the effects of mitosis on the metabolic rate estimates is very challenging with our data. The RNA of different cell cycle phases is mixed up in the measurements, as we measure bulk RNA and the cell cultures are not synchronized in their cell cycle. At the same time, it is unclear which RNA transcripts are expressed around mitosis and how abundant they are compared to the other cell cycle phases. We don't expect strong expression of most mRNAs immediately during cell division and consider the introduced biases rather low. The differences between dividing and non-dividing cells can be analysed in future experiments with synchronised cells.

Overall, we can expect that the potential biases introduced by read loss and mitosis are small. We cannot exclude pulldown of ER, and thereby ER-translated mRNA, together with the nuclear fraction. Still, ER pulldown only affects a subset of all transcripts and cannot fully explain the high median nuclear half life. To further validate our results, Jason Müller compared our half life estimates with the literature (Schueler et al., 2014) (Figure 23). This study performed 4sU labelling with subsequent biotinylation and streptavidin-pulldown of the labelled RNA. They calculated RNA half lives based on whole-cell extracts (WCE) and a simple exponential decay model. Note that we have to sum our nuclear and cytosolic half life estimates to compare with their whole-cell extract RNA half lives. Although Schueler et al. used a different experimental protocol and different human cell lines (MCF7 and HEK293), our and their half life estimates correlate well (Spearman correlation coefficient = 0.69 for MCF7 and 0.78 for HEK293). The HeLa half lives are substantially higher than the HEK293 half lives, but very comparable to the MCF7 cells, showing cell-type specific variation. HeLa and MCF7 cells might be more similar to each other as they are both derived from cancer tissue.

Based on the stable reproducibility of our estimates between both experimental samples and the good agreement with literature, we conclude that the method is robust and that my dynamic, two-compartment model of mRNA metabolism delivers reliable metabolic rate estimates.

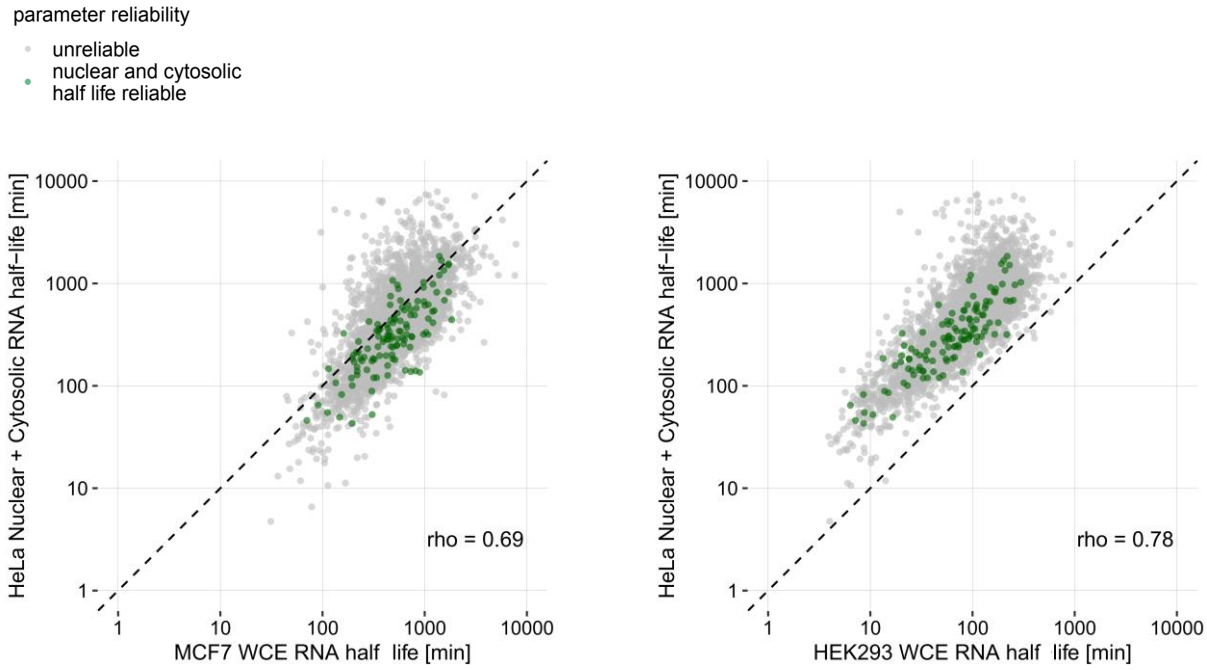


Figure 23 Comparison of our half life estimates with the literature (Schueler et al., 2014). Schueler et al. performed 4sU labelling with biotinylation and streptavidin-pulldown of labelled RNA. They calculated RNA half lives based on whole-cell extracts (WCE) and a simple decay model, for MCF7 cells (left) and HEK293 cells (right). Our nuclear and cytosolic half life estimates were summed to generate pseudo whole-cell estimates. Schueler et al.'s half lives are shown on the x-axis, ours on the y-axis. Green dots represent reliable and grey dots unreliable estimates w.r.t. our estimates and reliability criteria.

The comparisons were performed by Jason Müller. Figure adapted from Jason Müller (personal communication).

4.3 Biological interpretation and outlook

It remains to be discussed why nuclear export is much slower than cytosolic degradation or why, consequently, more mRNA resides in the nucleus than in the cytosol.

One very trivial reason for low nuclear export rates could be that pre-mRNA processing, especially transcript quality control, requires a lot of time. This hypothesis is strengthened by

an observation made in yeast of Zander et al. (2016). They found that transcripts of heat-stress associated proteins skip the nuclear quality control mechanisms when synthesised during heat shock. As it is crucial for the cells to react to extreme environmental stress such as heat shock, fast export of the required mRNA transcripts to the cytosol is essential.

A high amount of mRNA transcripts stored in the nucleus and a low amount of mRNA transcripts available in the cytosol can also be beneficial for gene expression regulation. A small amount of cytosolic mRNA can quickly be degraded to repress protein synthesis. A stock of export-ready mRNA in the nucleus can directly be delivered to the cytosol in demand of increased protein production. Together, this enables fast and dynamic changes in gene expression. A large nuclear mRNA pool could also buffer stochastic noise in mRNA synthesis introduced by transcriptional bursting.

Despite most measured 3'UTRs showing a relatively low nuclear vanishing rate, there was one outstanding group of 3'UTRs which were upregulated upon metabolic labelling and exported to the cytosol very fast. Both the nuclear and cytosolic expression levels reached their maximum within the first 60 minutes and drastically dropped again within the first 120 minutes of the labelling experiment. Recapitulate that the median nuclear half life of 3'UTRs with reliable estimates is 291min. As nuclear degradation is expected to be very low, the expression level drop in the nucleus in combination with the expression level increase in the cytosol can be attributed to a fast export mechanism of the transcripts. Potentially, these transcripts follow a similar export mechanism as described by Zander et al. (2016) in yeast, skipping nuclear quality control steps. The Supernova genes are mainly involved in cell cycle regulation, cellular growth, but also stress response, matching the role of the transcripts analysed by Zander et al. (2016). Our findings show that nuclear export happens on different time scales and emphasise the importance of mRNA export in gene expression regulation. In future studies, my model can be applied to analyse mRNA export and cytosolic degradation under various experimental conditions. The role of specific proteins in mRNA export can be investigated by combining targeted protein knock-down and whole-transcriptome export rate estimation. This can especially deliver new insights into the distinct nuclear export pathways and which subsets of mRNA transcripts they regulate.

The model is also useful to analyse how mRNA transcript isoforms differ in their metabolism. We found that distinct 3'UTR coverage peaks of the same gene show variation in their nuclear

half lives exceeding technical noise. This variation can be explained by differentially metabolised 3'UTR isoforms. Whole-transcript sequencing can be performed to measure further mRNA isoforms. In this case, the nuclear vanishing rate estimates have to be normalised for the transcript length: RNA fragments close to the promoter are synthesised first and detected earlier than fragments close to the 3'UTR; this yields seemingly longer nuclear vanishing rates for promoter-proximal fragments if not corrected for.

Our work has established a method to quantify the step of RNA export on a genome-wide scale. This will allow us to answer questions like: Which RNA-binding proteins are involved in RNA export? Which of those are general export factors that are involved in every export pathway, which proteins bind to specific subclasses of RNA? Is RNA export regulated upon response to specific stimuli and stresses? Does RNA export act as a quality control mechanism? We are looking forward to apply our technique to these questions.

5 References

Adamia, S., Reiman, T., Crainie, M., Mant, M. J., Belch, A. R., & Pilarski, L. M. (2005). Intronic splicing of hyaluronan synthase 1 (HAS1): a biologically relevant indicator of poor outcome in multiple myeloma. *Blood, 105*(12), 4836-4844.

Amorim, M. J., Cotobal, C., Duncan, C., & Mata, J. (2010). Global coordination of transcriptional control and mRNA decay during cellular differentiation. *Molecular systems biology, 6*(1), 380.

An, J. J., Gharami, K., Liao, G. Y., Woo, N. H., Lau, A. G., Vanevski, F., . . . & Xu, B. (2008). Distinct role of long 3′ UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. *Cell, 134*(1), 175-187.

Ardia, D., Mullen, K. M., Peterson, B. G., & Ulrich, J. (2020). 'DEoptim': Differential Evolution in 'R'. version 2.2-5.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature genetics, 25*(1), 25-29.

Bentley, D. L. (2014). Coupling mRNA processing with transcription in time and space. *Nature Reviews Genetics, 15*(3), 163-175.

Boutet, S. C., Cheung, T. H., Quach, N. L., Liu, L., Prescott, S. L., Edalati, A., . . . & Rando, T. A. (2012). Alternative polyadenylation mediates microRNA regulation of muscle stem cell function. *Cell stem cell, 10*(3), 327-336.

Boutz, P. L., Stoilov, P., Li, Q., Lin, C. H., Chawla, G., Ostrow, K., . . . Black, D. L. (2007). A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes & development, 21*(13), 1636-1652.

Brennan, C. M., Gallouzi, I. E., & Steitz, J. A. (2000). Protein ligands to HuR modulate its interaction with target mRNAs in vivo. *The Journal of cell biology, 151*(1), 1-14.

Bretes, H., Rouviere, J. O., Leger, T., Oeffinger, M., Devaux, F., Doye, V., & Palancade, B. (2014). Sumoylation of the THO complex regulates the biogenesis of a subset of mRNPs. *Nucleic acids research, 42*(8), 5043-5058.

Bromiley, P. A., & Thacker, N. A. (2002). The effects of an arcsin square root transform on a binomial distributed quantity. *TINA memo, 2007*, 157-165.

Bunch, H., Lawney, B. P., Burkholder, A., Ma, D., Zheng, X., Motola, S., . . . & Hu, G. (2016). RNA polymerase II promoter-proximal pausing in mammalian long non-coding genes. *Genomics, 108*(2), 64-77.

Cabili, M. N., Dunagin, M. C., McClanahan, P. D., Biaesch, A., Padovan-Merhar, O., Regev, A., . . . & Raj, A. (2015). Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome biology, 16*(1), 1-16.

Carney, L., Pierce, A., Rijnen, M., Sanchez, M. B., Hamzah, H. G., Zhang, L., . . . Whetton, A. D. (2009). THOC5 couples M-CSF receptor signaling to transcription factor expression. *Cellular signalling, 21*(2), 309-316.

Chakraborty, P., Wang, Y., Wei, J. H., Van Deursen, J., Yu, H., Malureanu, L., . . . Fontoura, B. M. (2008). Nucleoporin levels regulate cell cycle progression and phase-specific gene expression. *Developmental cell, 15*(5), 657-667.

Chen, C. Y., Zheng, D., Xia, Z., & Shyu, A. B. (2009). Ago–TNRC6 triggers microRNA-mediated decay by promoting two deadenylation steps. *Nature structural & molecular biology, 16*(11), 1160-1166.

Chen, L. L., Sabripour, M., Wu, E. F., Prieto, V. G., Fuller, G. N., & Frazier, M. L. (2005). A mutation-created novel intra-exonic pre-mRNA splice site causes constitutive activation of KIT in human gastrointestinal stromal tumors. *Oncogene, 24*(26), 4271-4280.

Chendrimada, T. P., Gregory, R. I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., & Shiekhattar, R. (2005). TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature, 436*(7051), 740-744.

Cleary, M. D., Meiering, C. D., Jan, E., Guymon, R., & Boothroyd, J. C. (2005). Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nature biotechnology, 23*(2), 232-237.

Coyle, J. H., Bor, Y. C., Rekosh, D., & Hammarskjold, M. L. (2011). The Tpr protein regulates export of mRNAs with retained introns that traffic through the Nxf1 pathway. *Rna, 17*(7), 1344-1356.

Cramer, P., Pesce, C. G., Baralle, F. E., & Kornblihtt, A. R. (1997). Functional association between promoter structure and transcript alternative splicing. *Proceedings of the National Academy of Sciences, 94*(21), 11456-11460.

Crick, F. (1970). Central dogma of molecular biology. *Nature, 227*(5258), 561-563.

Danckwardt, S., Kaufmann, I., Gentzel, M., Foerstner, K. U., Gantzert, A. S., Gehring, N. H., … & Kulozik, A. E. (2007). Splicing factors stimulate polyadenylation via USEs at non-canonical 3′ end formation signals. *The EMBO journal, 26*(11), 2658-2669.

De Pretis, S., Kress, T., Morelli, M. J., Melloni, G. E., Riva, L., Amati, B., & Pelizzola, M. (2015). INSPEcT: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA-and 4sU-seq time course experiments. *Bioinformatics, 31*(17), 2829-2835.

Decker, C. J., & Parker, R. (1993). A turnover pathway for both stable and unstable mRNAs in yeast: evidence for a requirement for deadenylation. *Genes & development, 7*(8), 1632-1643.

Delaleau, M., & Borden, K. L. (2015). Multiple export mechanisms for mRNAs. *Cells*, *4*(3), 452-473.

Dieci, G., Fiorino, G., Castelnuovo, M., Teichmann, M., & Pagano, A. (2007). The expanding RNA polymerase III transcriptome. *TRENDS in Genetics*, *23*(12), 614-622.

Dölken, L., Ruzsics, Z., Rädle, B., Friedel, C. C., Zimmer, R., Mages, J., . . . & Koszinowski, U. H. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *Rna*, *14*(9), 1959-1972.

Domínguez-Sánchez, M. S., Sáez, C., Japón, M. A., Aguilera, A., & Luna, R. (2011). Differential expression of THOC1 and ALY mRNP biogenesis/export factors in human cancers. *BMC cancer*, *11*(1), 1-11.

Eddy, S. R. (1999). Noncoding RNA genes. *Current opinion in genetics & development*, *9*(6), 695-699.

Edwards, S. M. (2020). lemon: Freshing Up your 'ggplot2' Plots. R package version 0.4.5. https://CRAN.R-project.org/package=lemon.

Eser, P., Demel, C., Maier, K. C., Schwalb, B., Pirkl, N., Martin, D. E., . . . & Tresch, A. (2014). Periodic mRNA synthesis and degradation co-operate during cell cycle gene expression. *Molecular systems biology*, *10*(1), 717.

Eulalio, A., Huntzinger, E., Nishihara, T., Rehwinkel, J., Fauser, M., & Izaurralde, E. (2009). Deadenylation is a widespread effect of miRNA regulation. *Rna*, *15*(1), 21-32.

Farago, M., Nahari, T., Hammel, C., Cole, C. N., & Choder, M. (2003). Rpb4p, a subunit of RNA polymerase II, mediates mRNA export during stress. *Molecular biology of the cell*, *14*(7), 2744-2755.

Faria, A. M., Levay, A., Wang, Y., Kamphorst, A. O., Rosa, M. L., Nussenzveig, D. R., . . . & Fontoura, B. M. (2006). The nucleoporin Nup96 is required for proper expression of interferon-regulated proteins and functions. *Immunity*, *24*(3), 295-304.

Folkmann, A. W., Collier, S. E., Zhan, X., Ohi, M. D., & Wente, S. R. (2013). Gle1 functions during mRNA export in an oligomeric complex that is altered in human disease. *Cell*, *155*(3), 582-593.

Galy, V., Gadal, O., Fromont-Racine, M., Romano, A., Jacquier, A., & Nehrbass, U. (2004). Nuclear retention of unspliced mRNAs in yeast is mediated by perinuclear Mlp1. *Cell*, *116*(1), 63-73.

Girard, C., Will, C. L., Peng, J., Makarov, E. M., Kastner, B., Lemm, I., . . . Lührmann, R. (2012). Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. *Nature communications*, *3*(1), 1-12.

Guo, S., Hakimi, M. A., Baillat, D., Chen, X., Farber, M. J., Klein-Szanto, A. J., . . . & Shiekhattar, R. (2005). Linking transcriptional elongation and messenger RNA export to metastatic breast cancers. *Cancer research*, *65*(8), 3011-3016.

Hackmann, A., Wu, H., Schneider, U. M., Meyer, K., Jung, K., & Krebber, H. (2014). Quality control of spliced mRNAs requires the shuttling SR proteins Gbp2 and Hrb1. *Nature communications, 5*(1), 1-14.

Haimovich, G., Medina, D. A., Causse, S. Z., Garber, M., Millán-Zambrano, G., Barkai, O., . . . & Choder, M. (2013). Gene expression is circular: factors for mRNA degradation also foster mRNA synthesis. *Cell, 153*(5), 1000-1011.

Herold, A., Klymenko, T., & Izaurralde, E. (2001). NXF1/p15 heterodimers are essential for mRNA nuclear export in Drosophila. *Rna, 7*(12), 1768-1780.

Herzog, V. A., Reichholf, B., Neumann, T., Rescheneder, P., Bhat, P., Burkard, T. R., . . . & Ameres, S. L. (2017). Thiol-linked alkylation of RNA to assess expression dynamics. *Nature methods, 14*(12), 1198-1204.

Hirose, Y., & Manley, J. L. (1998). RNA polymerase II is an essential mRNA polyadenylation factor. *Nature, 395*(6697), 93-96.

Ho, C. K., & Shuman, S. (1999). Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme. *Molecular cell, 3*(3), 405-411.

Holt, I., Mittal, S., Furling, D., Butler-Browne, G. S., David Brook, J., & Morris, G. E. (2007). Defective mRNA in myotonic dystrophy accumulates at the periphery of nuclear splicing speckles. *Genes to Cells, 12*(9), 1035-1048.

Ji, Z., & Tian, B. (2009). Reprogramming of 3′ untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PloS one, 4*(12), e8419.

Jiao, X., Chang, J. H., Kilic, T., Tong, L., & Kiledjian, M. (2013). A mammalian pre-mRNA 5′ end capping quality control mechanism and an unexpected link of capping to pre-mRNA processing. *Molecular cell, 50*(1), 104-115.

Jing, Q., Huang, S., Guth, S., Zarubin, T., Motoyama, A., Chen, J., . . . & Han, J. (2005). Involvement of microRNA in AU-rich element-mediated mRNA instability. *Cell, 120*(5), 623-634.

Johnson, C. P. (2000). Tracking Col1a1 RNA in Osteogenesis ImperfectaSplice-Defective Transcripts Initiate Transport from the Gene but Are Retained within the Sc35 Domain. *Journal of Cell Biology, 150*(3), 417-432.

Jürges, C., Dölken, L., & Erhard, F. (2018). Dissecting newly transcribed and old RNA using GRAND-SLAM. *Bioinformatics, 34*(13), i218-i226.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic acids research, 32*(suppl_1), D493-D496.

Katahira, J., Dimitrova, L., Imai, Y., & Hurt, E. (2015). NTF2-like domain of Tap plays a critical role in cargo mRNA recognition and export. *Nucleic acids research, 43*(3), 1894-1904.

Katahira, J., Sträßer, K., Podtelejnikov, A., Mann, M., Jung, J. U., & Hurt, E. (1999). The Mex67p-mediated nuclear mRNA export pathway is conserved from yeast to human. *The EMBO journal, 18*(9), 2593-2609.

Kedde, M., Van Kouwenhove, M., Zwart, W., Vrielink, J. A., Elkon, R., & Agami, R. (2010). APumilio-induced RNA structure switch in p27-3′ UTR controls miR-221 and miR-222 accessibility. *Nature cell biology, 12*(10), 1014-1020.

Khodor, Y. L., Rodriguez, J., Abruzzi, K. C., Tang, C. H., Marr, M. T., & Rosbash, M. (2011). Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes & development, 25*(23), 2502-2512.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature methods, 12*(4), 357-360.

Kong, J., & Liebhaber, S. A. (2007). A cell type–restricted mRNA surveillance pathway triggered by ribosome extension into the 3′ untranslated region. *Nature structural & molecular biology, 14*(7), 670-676.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods, 9*(4), 357-359.

Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S., & Mayr, C. (2013). Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & development, 27*(21), 2380-2396.

Lin, X., Miller, J. W., Mankodi, A., Kanadia, R. N., Yuan, Y., Moxley, R. T., . . . & Thornton, C. A. (2006). Failure of MBNL1-dependent post-natal splicing transitions in myotonic dystrophy. *Human molecular genetics, 15*(13), 2087-2097.

Lotan, R., Bar-On, V. G., Harel-Sharvit, L., Duek, L., Melamed, D., & Choder, M. (2005). The RNA polymerase II subunit Rpb4p mediates decay of a specific class of mRNAs. *Genes & development, 19*(24), 3004-3016.

Lu, F., Gladden, A. B., & Diehl, J. A. (2003). An alternatively spliced cyclin D1 isoform, cyclin D1b, is a nuclear oncogene. *Cancer research, 63*(21), 7056-7061.

Maillet, I., Buhler, J. M., Sentenac, A., & Labarre, J. (1999). Rpb4p is necessary for RNA polymerase II activity at high temperature. *Journal of Biological Chemistry, 274*(32), 22586-22590.

Masuda, S., Das, R., Cheng, H., Hurt, E., Dorman, N., & Reed, R. (2005). Recruitment of the human TREX complex to mRNA during splicing. *Genes & development, 19*(13), 1512-1517.

Mayr, C., & Bartel, D. P. (2009). Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell, 138*(4), 673-684.

McCracken, S., Fong, N., Rosonina, E., Yankulov, K., Brothers, G., Siderovski, D., . . . Bentley, D. L. (1997). 5′-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Genes & development, 11*(24), 3306-3318.

Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., & Tuschl, T. (2004). Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Molecular cell*, *15*(2), 185-197.

Microsoft and Steve Weston. (2020). foreach: Provides Foreach Looping Construct. R package version 1.5.1. https://CRAN.R-project.org/package=foreach.

Microsoft Corporation and Steve Weston. (2020). doParallel: Foreach Parallel Adaptor for the 'parallel' Package. R package version 1.0.16. https://CRAN.R-project.org/package=doParallel.

Miller, C., Schwalb, B., Maier, K., Schulz, D., Dümcke, S., Zacher, B., . . . & Cramer, P. (2011). Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Molecular systems biology*, *7*(1), 458.

Mor, A., Suliman, S., Ben-Yishay, R., Yunger, S., Brody, Y., & Shav-Tal, Y. (2010). Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells. *Nature cell biology*, *12*(6), 543-552.

Muriel, J. M., Dong, C., Hutter, H., & Vogel, B. E. (2005). Fibulin-1C and Fibulin-1D splice variants have distinct functions and assemble in a hemicentin-dependent manner. *Development*, *132*(19), 4223-4234.

Nagaike, T., Logan, C., Hotta, I., Rozenblatt-Rosen, O., Meyerson, M., & Manley, J. L. (2011). Transcriptional activators enhance polyadenylation of mRNA precursors. *Molecular cell*, *41*(4), 409-418.

Neumann, T., Herzog, V. A., Muhar, M., Haeseler, v. A., Zuber, J., Ameres, S. L., & Rescheneder, P. (2019). Quantification of experimentally induced nucleotide conversions in high-throughput sequencing datasets. *BMC Bioinformatics*, *20*(1), 258.

Neuwirth, E. (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer.

Nishikura, K. (2016). A-to-I editing of coding and non-coding RNAs by ADARs. *Nature reviews Molecular cell biology*, *17*(2), 83-96.

Nousiainen, H. O., Kestilä, M., Pakkasjärvi, N., Honkala, H., Kuure, S., Tallila, J., . . . & Peltonen, L. (2008). Mutations in mRNA export mediator GLE1 result in a fetal motoneuron disease. *Nature genetics*, *40*(2), 155-157.

Oesterreich, F. C., Preibisch, S., & Neugebauer, K. M. (2010). Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Molecular cell*, *40*(4), 571-581.

Okamura, M., Inose, H., & Masuda, S. (2015). RNA Export through the NPC in Eukaryotes. *Genes*, *6*(1), 124-149.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., . . . & Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, *44*(D1), D733-D745.

Pelechano, V., & Pérez-Ortín, J. E. (2008). The transcriptional inhibitor thiolutin blocks mRNA degradation in yeast. *Yeast*, *25*(2), 85-92.

Persson, H., Kvist, A., Vallon-Christersson, J., Medstrand, P., Borg, Å., & Rovira, C. (2009). The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. *Nature cell biology, 11*(10), 1268-1271.

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., . . . & Regev, A. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature biotechnology, 29*(5), 436-442.

Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D. J., Pauli, A., . . . & Regev, A. (2014). High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell, 159*(7), 1698-1710.

Raghavan, A., Ogilvie, R. L., Reilly, C., Abelson, M. L., Raghavan, S., Vasdewani, J., . . . Bohjanen, P. R. (2002). Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic acids research, 30*(24), 5529-5538.

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., & Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biol, 4*(10), e309.

Rosonina, E., Bakowski, M. A., McCracken, S., & Blencowe, B. J. (2003). Transcriptional activators control splicing and 3′-end cleavage levels. *Journal of Biological Chemistry, 278*(44), 43034-43040.

Saguez, C., Schmid, M., Olesen, J. R., Ghazy, M. A., Qu, X., Poulsen, M. B., . . . & Jensen, T. H. (2008). Nuclear mRNA surveillance in THO/sub2 mutants is triggered by inefficient polyadenylation. *Molecular cell, 31*(1), 91-103.

Saito, Y., Kasamatsu, A., Yamamoto, A., Shimizu, T., Yokoe, H., Sakamoto, Y., . . . & Uzawa, K. (2013). ALY as a potential contributor to metastasis in human oral squamous cell carcinoma. *Journal of cancer research and clinical oncology, 139*(4), 585-594.

Schueler, M., Munschauer, M., Gregersen, L. H., Finzel, A., Loewer, A., Chen, W., . . . & Dieterich, C. (2014). Differential protein occupancy profiling of the mRNA transcriptome. *Genome biology, 15*(1), 1-17.

Schwalb, B., Tresch, A., Torkler, P., Duemcke, S., Demel, C., Ripley, B., & Venables, B. (2020). LSD: Lots of Superior Depictions. R package version 4.1-0. https://CRAN.R-project.org/package=LSD.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., . . . & Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature, 473*(7347), 337-342.

Shalem, O., Dahan, O., Levo, M., Martinez, M. R., Furman, I., Segal, E., & Pilpel, Y. (2008). Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Molecular systems biology, 4*(1), 4.

Shalem, O., Groisman, B., Choder, M., Dahan, O., & Pilpel, Y. (2011). Transcriptome kinetics is governed by a genome-wide coupling of mRNA production and degradation: a role for RNA Pol II. *PLoS Genet, 7*(9), e1002273.

Shav-Tal, Y., Darzacq, X., Shenoy, S. M., Fusco, D., Janicki, S. M., Spector, D. L., & Singer, R. H. (2004). Dynamics of single mRNPs in nuclei of living cells. *science, 304*(5678), 1797-1800.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research, 29*(1), 308-311.

Shiga, A., Ishihara, T., Miyashita, A., Kuwabara, M., Kato, T., Watanabe, N., . . . & Onodera, O. (2012). Alteration of POLDIP3 splicing associated with loss of function of TDP-43 in tissues affected with ALS. *PloS one, 7*(8), e43120.

Slobodin, B., Bahat, A., Sehrawat, U., Becker-Herman, S., Zuckerman, B., Weiss, A. N., . . . Dikstein, R. (2020). Transcription dynamics regulate poly (A) tails and expression of the RNA degradation machinery to balance mRNA levels. *Molecular cell, 78*(3), 434-444.

Smith, K. P., Byron, M., Johnson, C., Xing, Y., & Lawrence, J. B. (2007). Defining early steps in mRNA transport: mutant mRNA in myotonic dystrophy type I is blocked at entry into SC-35 domains. *Journal of Cell Biology, 178*(6), 951-964.

Soheilypour, M., & Mofrad, M. R. (2018). Quality control of mRNAs at the entry of the nuclear pore: Cooperation in a complex molecular system. *Nucleus, 9*(1), 202-211.

Solomon, D. A., Wang, Y., Fox, S. R., Lambeck, T. C., Giesting, S., Lan, Z., . . . & Knudsen, E. S. (2003). Cyclin D1 splice variants: differential effects on localization, RB phosphorylation, and cellular transformation. *Journal of Biological Chemistry, 278*(32), 30339-30347.

Sträßer, K., Masuda, S., Mason, P., Pfannstiel, J., Oppizzi, M., Rodriguez-Navarro, S., . . . & Hurt, E. (2002). TREX is a conserved complex coupling transcription with messenger RNA export. *Nature, 417*(6886), 304-308.

Sun, M., Schwalb, B., Schulz, D., Pirkl, N., Etzold, S., Larivière, L., . . . & Cramer, P. (2012). Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome research, 22*(7), 1350-1359.

Sun, S., Ling, S. C., Qiu, J., Albuquerque, C. P., Zhou, Y., Tokunaga, S., . . . & Cleveland, D. W. (2015). ALS-causative mutations in FUS/TLS confer gain and loss of function by altered association with SMN and U1-snRNP. *Nature communications, 6*(1), 1-14.

Takagaki, Y., & Manley, J. L. (1998). Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation. *Molecular cell, 2*(6), 761-771.

Takagaki, Y., Seipelt, R. L., Peterson, M. L., & Manley, J. L. (1996). The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell, 87*(5), 941-952.

Topisirovic, I., Siddiqui, N., Lapointe, V. L., Trost, M., Thibault, P., Bangeranye, C., . . . & Borden, K. L. (2009). Molecular dissection of the eukaryotic initiation factor 4E (eIF4E) export-competent RNP. *The EMBO journal, 28*(8), 1087-1098.

Viphakone, N., Hautbergue, G. M., Walsh, M., Chang, C. T., Holland, A., Folco, E. G., . . . & Wilson, S. A. (2012). TREX exposes the RNA-binding domain of Nxf1 to enable mRNA export. *Nature communications, 3*(1), 1-14.

Wang, L., Miao, Y. L., Zheng, X., Lackford, B., Zhou, B., Han, L., . . . Hu, G. (2013). The THO complex regulates pluripotency gene mRNA export and controls embryonic stem cell self-renewal and somatic cell reprogramming. *Cell stem cell, 13*(6), 676-690.

Wang, X. Q., Luk, J. M., Leung, P. P., Wong, B. W., Stanbridge, E. J., & Fan, S. T. (2005). Alternative mRNA splicing of liver intestine-cadherin in hepatocellular carcinoma. *Clinical Cancer Research, 11*(2), 483-489.

Watanabe, M., Fukuda, M., Yoshida, M., Yanagida, M., & Nishida, E. (1999). Involvement of CRM1, a nuclear export receptor, in mRNA export in mammalian cells and fission yeast. *Genes to Cells, 4*(5), 291-297.

Werner, M., Thuriaux, P., & Soutourina, J. (2009). Structure–function analysis of RNA polymerases I and III. *Current opinion in structural biology, 19*(6), 740-745.

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software, 21*(12), pp. 1-20. http://www.jstatsoft.org/v21/i12/.

Wickham, H. (2019). stringr: Simple, Consistent Wrappers for Common String.

Wickham, H., & Seidel, D. (2020). scales: Scale Functions for Visualization. R package version 1.1.1. https://CRAN.R-project.org/package=scales.

Wickramasinghe, V. O., Savill, J. M., Chavali, S., Jonsdottir, A. B., Rajendra, E., Grüner, T., . . . Venkitaraman, A. R. (2013). Human inositol polyphosphate multikinase regulates transcript-selective nuclear mRNA export to preserve genome integrity. *Molecular cell, 51*(6), 737-750.

Wu, L. F., & Belasco, J. G. (2006). MicroRNAs direct rapid deadenylation of mRNA. *Proceedings of the National Academy of Sciences, 103*(11), 4034-4039.

Wu, X., & Brewer, G. (2012). The regulation of mRNA stability in mammalian cells: 2.0. *Gene, 500*(1), 10-21.

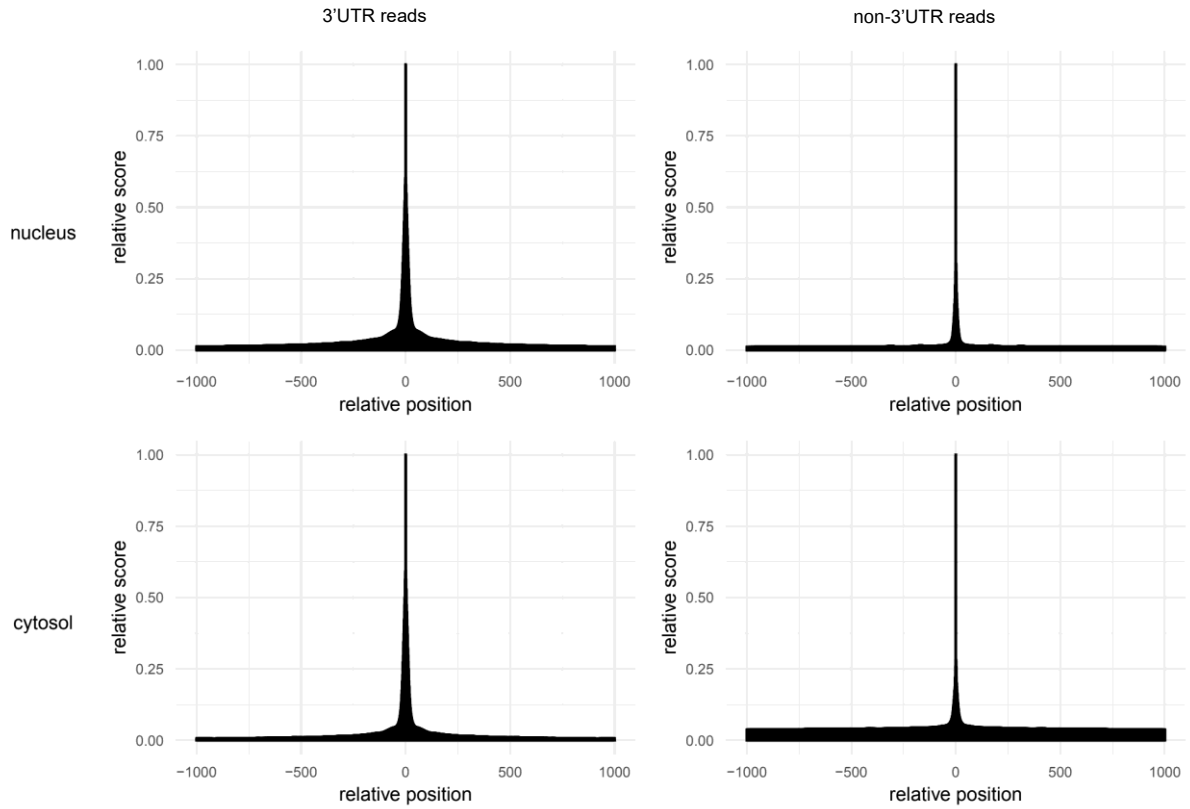
Yamashita, A., Chang, T. C., Yamashita, Y., Zhu, W., Zhong, Z., Chen, C. Y., & Shyu, A. B. (2005). Concerted action of poly (A) nucleases and decapping enzyme in mammalian mRNA turnover. *Nature structural & molecular biology, 12*(12), 1054-1063.

Zander, G., Hackmann, A., Bender, L., Becker, D., Lingner, T., Salinas, G., & Krebber, H. (2016). mRNA quality control is bypassed for immediate export of stress-responsive transcripts. *Nature, 540*(7634), 593-596.

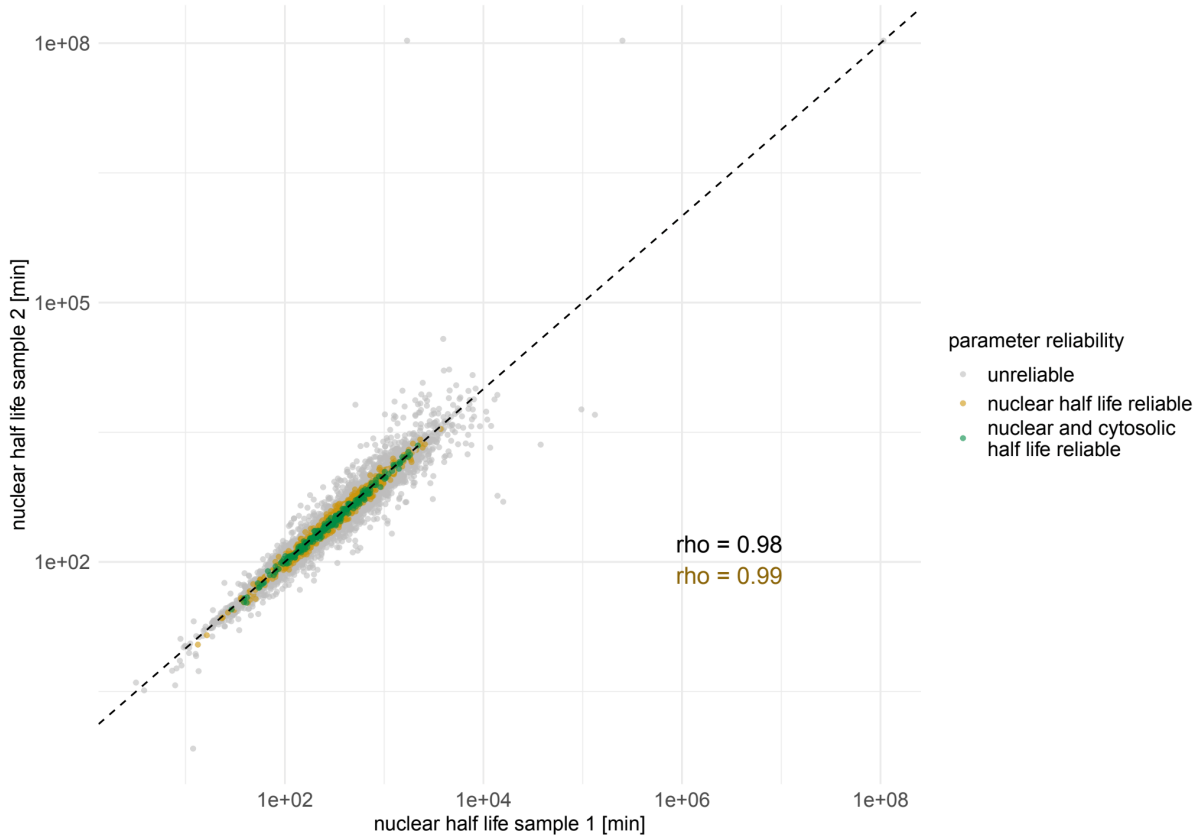
Zeisel, A. K.-H. (2011). Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Molecular systems biology, 7*(1), 529.

6 Supplemental Material

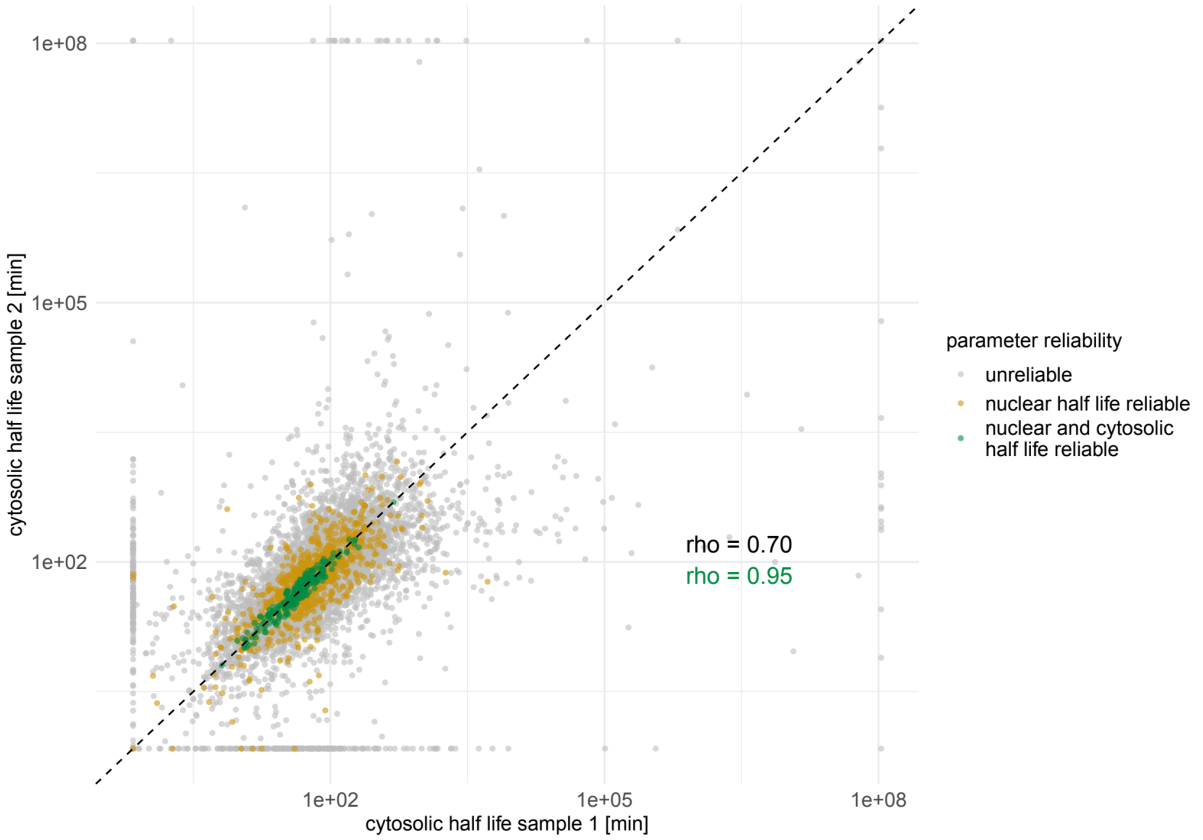
6.1 Supplemental Figures



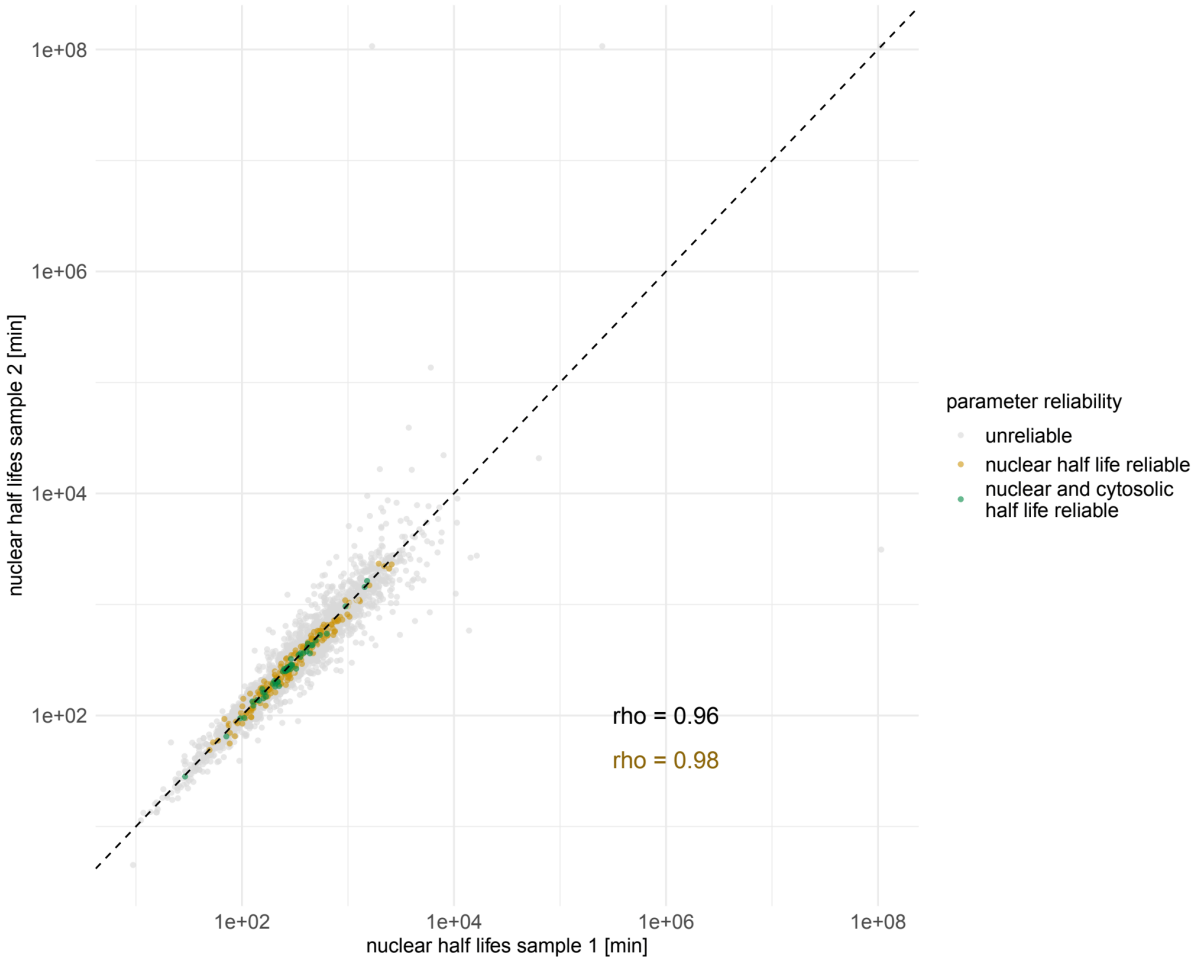
Supplemental Figure 1 Relative coverage profiles. Coverage profiles as shown in Figure 6, but displaying the untrimmed profiles.



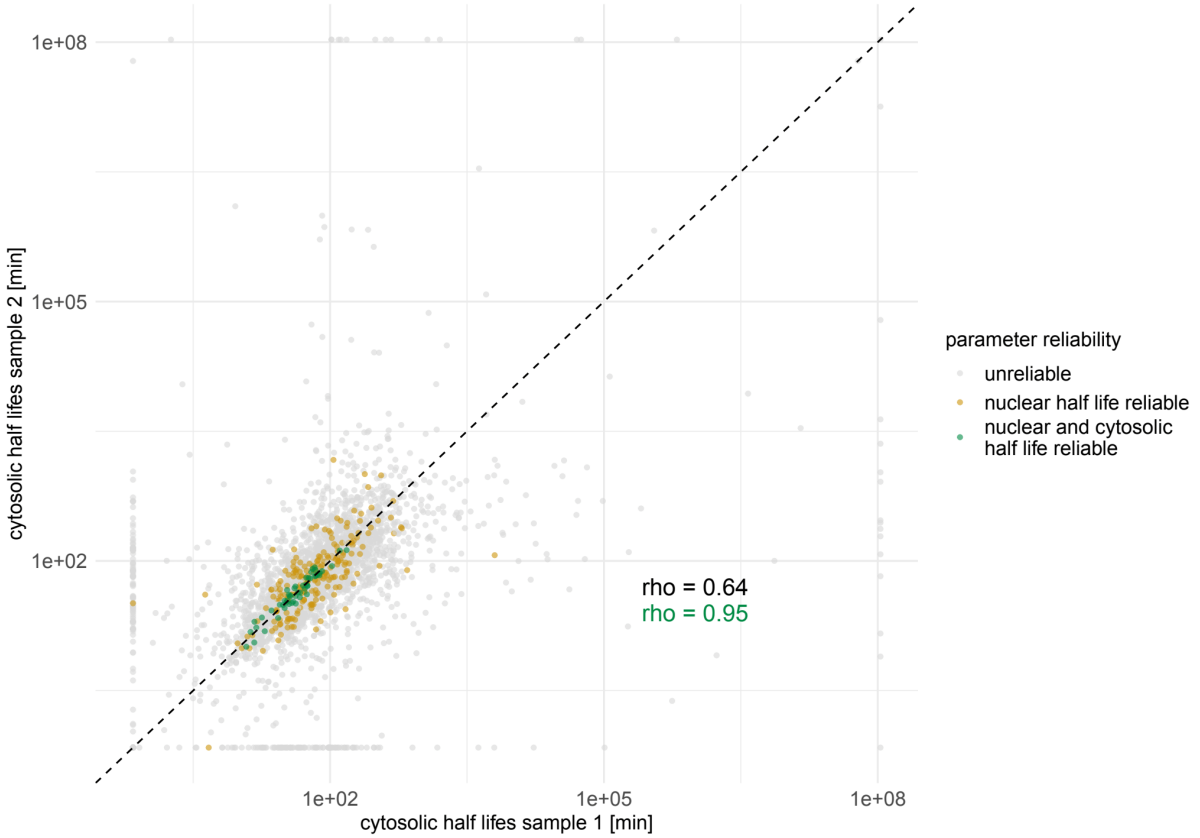
Supplemental Figure 2 Nuclear half life estimates of 3'UTRs compared between sample 1 and 2. Estimates are shown for all detectable 3'UTRs. The dashed line indicates the diagonal. Spearman's rho is indicated for all detectable 3'UTRs (8119 UTRs) (black) and 3'UTRs with reliable estimates (1297 UTRs) (yellow).



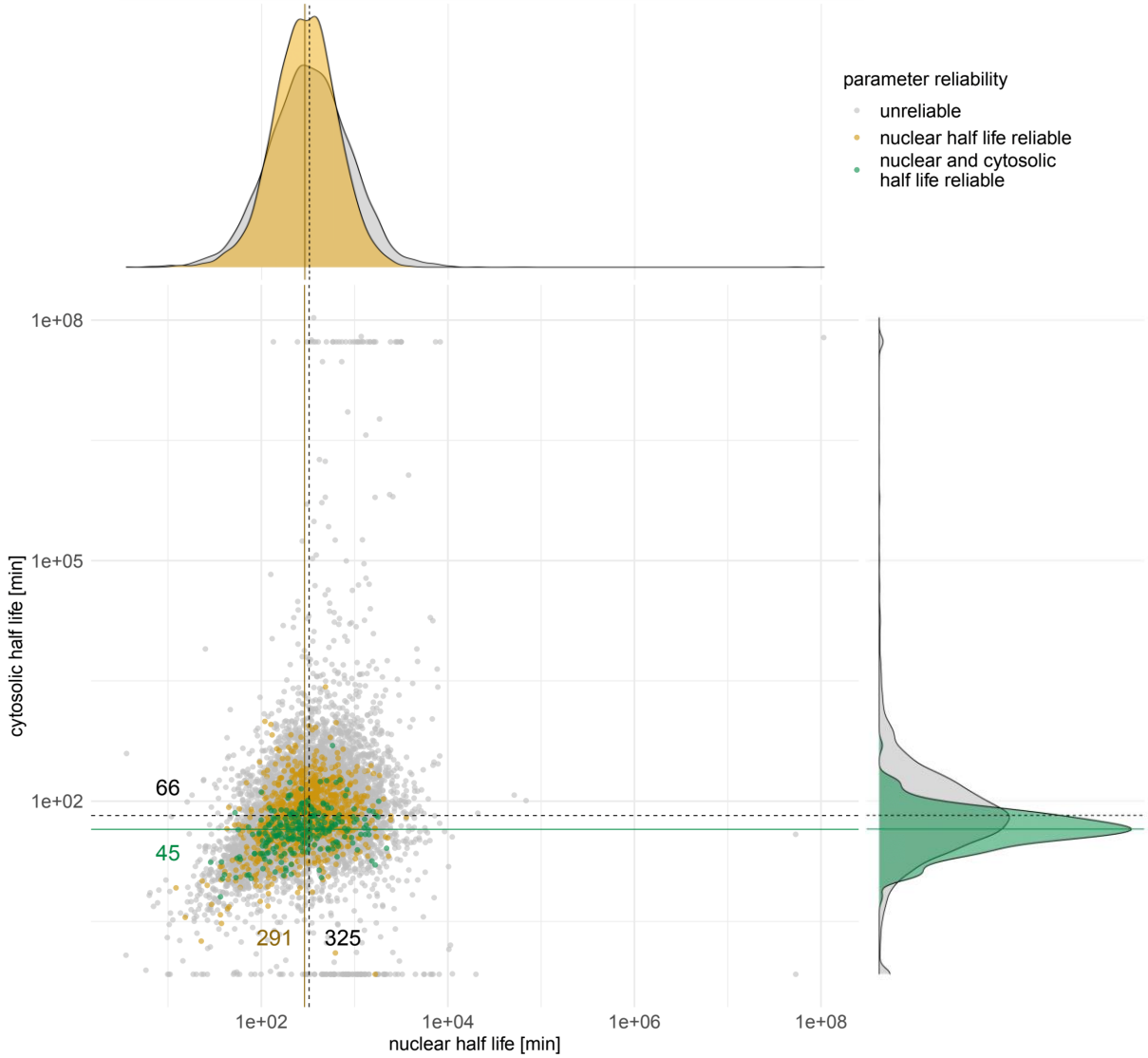
Supplemental Figure 3 Cytosolic half life estimates of 3'UTRs compared between sample 1 and 2. Estimates are shown for all detectable 3'UTRs. The dashed line indicates the diagonal. Spearman's rho is indicated for all detectable 3'UTRs (8119 UTRs) (black) and 3'UTRs with reliable estimates (251 UTRs) (green).



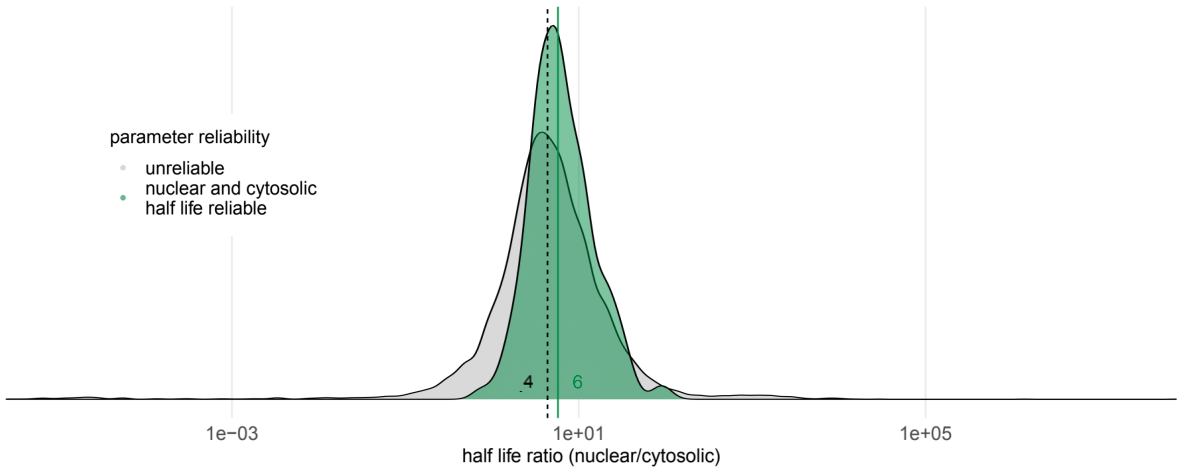
Supplemental Figure 4 Nuclear half life estimates of 3'UTR peaks compared between sample 1 and 2. Estimates are shown for all detectable 3'UTR peaks. The dashed line indicates the diagonal. Spearman's rho is indicated for all detectable peaks (2987 peaks) (black) and peaks with reliable estimates (261) (yellow).



Supplemental Figure 5 Cytosolic half life estimates of 3'UTR peaks compared between sample 1 and 2. Estimates are shown for all detectable 3'UTR peaks. The dashed line indicates the diagonal. Spearman's rho is indicated for all detectable peaks (2987 peaks) (black) and peaks with reliable estimates (51 peaks) (green).



Supplemental Figure 6 Nuclear and cytosolic half life estimates. Nuclear and cytosolic RNA half life estimates as shown in Figure 15, but displaying the untrimmed plot.



Supplemental Figure 7 Distribution of nuclear by cytosolic half life ratios. Ratios as shown in Figure 16, but displaying the untrimmed plot.

6.2 Supplemental Tables

3'UTR reads

time	inferred A>G		inferred T>C	
	nucleus, (-) strand	cytosol, (-) strand	nucleus, (+) strand	cytosol, (+) strand
15min	0.0527	0.0605	0.0517	0.0681
30min	0.0511	0.0627	0.0497	0.0587
45min	0.0504	0.0625	0.0493	0.0589
60min	0.0547	0.0626	0.0540	0.0587
90min	0.0510	0.0597	0.0499	0.0560
120min	0.0529	0.0606	0.0517	0.0568
180min	0.0541	0.0570	0.0524	0.0538
Average	**0.0524**	**0.0608**	**0.0512**	**0.0587**

False positive error nucleus: **0.0518%**
False positive error cytosol: **0.0598%**

non-3'UTR reads

time	inferred A>G		inferred T>C	
	nucleus, (-) strand	cytosol, (-) strand	nucleus, (+) strand	cytosol, (+) strand
15min	0.0585	0.0536	0.0593	0.0707
30min	0.0541	0.0552	0.0547	0.0527
45min	0.0524	0.0549	0.0527	0.0524
60min	0.0562	0.0552	0.0575	0.0526
90min	0.0535	0.0510	0.0543	0.0501
120min	0.0527	0.0507	0.0540	0.0503
180min	0.0544	0.0466	0.0551	0.0470
Average	**0.0545**	**0.0524**	**0.0554**	**0.0537**

False positive error nucleus: **0.0550%**
False positive error cytosol: **0.0531%**

Supplemental Table 1 Estimation of false-positive sequencing error rates. A>G sequencing errors on the (-) strand or T>C sequencing errors on the (+) strand, respectively, are misinterpreted as labeling conversions (false-positive conversions). False-positive sequencing error rates were inferred using the control samples as described in 2.8 Estimation of sequencing errors and labelling time shift. The tables summarize the estimation results for 3'UTR reads (top table) and non-3'UTR reads, considering only non-3'UTR coverage peaks with at least 5 reads (bottom table). Results are shown for the single measurement time points of the nuclear and cytosolic compartment, both for the (-) and the (+) strand. The final false positive rates for the nuclear and cytosolic compartment are obtained by averaging over measurement time points and the (-) and (+) strand.

3'UTR reads

time	G>non-G		C>non-C	
	nucleus, (-) strand	cytosol, (-) strand	nucleus, (+) strand	cytosol, (+) strand
15min	0.2897	0.3467	0.2754	0.3231
30min	0.2817	0.3672	0.2676	0.3413
45min	0.2775	0.3613	0.2640	0.3362
60min	0.3141	0.3664	0.2968	0.3417
90min	0.2812	0.3457	0.2675	0.3236
120min	0.3007	0.3574	0.2825	0.3326
180min	0.3076	0.3306	0.2860	0.3109
Average	**0.2932**	**0.3536**	**0.2771**	**0.3299**

False negative error nucleus: **0.2852%**
False negative error cytosol: **0.3418%**

non-3'UTR reads

time	G>non-G		C>non-C	
	nucleus, (-) strand	cytosol, (-) strand	nucleus, (+) strand	cytosol, (+) strand
15min	0.2541	0.2649	0.2561	0.2522
30min	0.2385	0.2821	0.2357	0.2668
45min	0.2311	0.2780	0.2235	0.2575
60min	0.2565	0.2791	0.2505	0.2634
90min	0.2374	0.2568	0.2339	0.2535
120min	0.2369	0.2634	0.2317	0.2549
180min	0.2438	0.2371	0.2372	0.2372
Average	**0.2426**	**0.2659**	**0.2384**	**0.2551**

False negative error nucleus: **0.2405%**
False negative error cytosol: **0.2605%**

Supplemental Table 2 Estimation of false-negative sequencing error rates. G>non-G sequencing errors on the (-) strand or C>non-C sequencing errors on the (+) strand, respectively, are masking labeling conversions (false-negative conversions). False-negative sequencing error rates were calculated from the final SLAM seq experiment data as described in 2.8 Estimation of sequencing errors and labelling time shift. The tables summarize the estimation results for 3'UTR reads (top table) and non-3'UTR reads, considering only non-3'UTR coverage peaks with at least 5 reads (bottom table). Results are shown for the single measurement time points of the nuclear and cytosolic compartment, both for the (-) and the (+) strand. The final false negative rates for the nuclear and cytosolic compartment are obtained by averaging over measurement time points and the (-) and (+) strand.

3'UTR reads

labeling conversion rates

time	nucleus, (-) strand inferred A>G		nucleus, (+) strand inferred T>C	
	sample 1	sample 2	sample 1	sample 2
0min	0.0518	0.0518	0.0518	0.0518
15min	0.1122	0.1104	0.1283	0.1243
30min	0.3148	0.3347	0.3593	0.3797

differences in labeling conversion rates

difference	nucleus, (-) strand		nucleus, (+) strand	
	sample 1	sample 2	sample 1	sample 2
15min-0min	0.0604	0.0586	0.0764	0.0725
30min-15min	0.2026	0.2243	0.2310	0.2553

fraction of first difference from second difference

sample	nucleus, (-) strand			nucleus, (+) strand		
	fraction	effective labeling [min]	labeling shift [min]	fraction	effective labeling [min]	labeling shift [min]
sample 1	0.2980	4.4700	10.5300	0.3308	4.9616	10.0384
sample 2	0.2613	3.9190	11.0810	0.2840	4.2596	10.7404

average time shift: **sample 1: 10.28min**
sample 2: 10.91min

Supplemental Table 3 Estimation of the labeling time shift for the 3'UTR reads. The labeling time shift was calculated by comparing the increase in the labeling conversion rate between 0min to 15min and 15min to 30min of the nuclear fraction, as described in 2.8 Estimation of sequencing errors and labelling time shift. The results of the single computational steps are shown in the distinct tables. The final labeling time shifts for the experimental samples 1 and 2 are obtained by averaging the time shift estimates of the (-) and (+) strand.

non-3'UTR reads

labeling conversions

time	nucleus, (-) strand inferred A>G		nucleus, (+) strand inferred T>C	
	sample 1	sample 2	sample 1	sample 2
0min	0.0550	0.0550	0.0550	0.0550
15min	0.6105	0.5979	0.6583	0.6539
30min	1.6913	1.7588	1.8647	1.9244

labeling conversion differences

difference	nucleus, (-) strand		nucleus, (+) strand	
	sample 1	sample 2	sample 1	sample 2
15min-0min	0.5555	0.5429	0.6033	0.5989
30min-15min	1.0808	1.1609	1.2064	1.2705

fraction of first difference from second difference

sample	nucleus, (-) strand			nucleus, (+) strand		
	fraction	effective labeling [min]	labeling shift [min]	fraction	effective labeling [min]	labeling shift [min]
sample 1	0.5140	7.7101	7.2899	0.5001	7.5015	7.4985
sample 2	0.4677	7.0151	7.9849	0.4714	7.0709	7.9291

average time shift: **sample 1:** **7.39min**
sample 2: **7.96min**

Supplemental Table 4 Estimation of the labeling time shift for the non-3'UTR reads. Only non-3'UTR coverage peaks with at least 5 reads were considered. The labeling time shift was calculated by comparing the increase in the labeling conversion rate between 0min to 15min and 15min to 30min of the nuclear fraction, as described in 2.8 Estimation of sequencing errors and labelling time shift. The results of the single computational steps are shown in the distinct tables. The final labeling time shifts for the experimental samples 1 and 2 are obtained by averaging the time shift estimates of the (-) and (+) strand.

type of peak	nucleus				cytosol			
	sample 1		sample 2		sample 1		sample 2	
	abs	rel [%]	abs	rel [%]	abs	rel [%]	abs	rel [%]
exonic	872	4.61	733	4.90	233	10.61	211	10.90
intronic	12453	65.82	9787	65.46	581	26.45	494	25.53
other	5594	29.57	4432	29.64	1383	62.95	1230	63.57
total	18919	100.00	14952	100.00	2197	100.00	1935	100.00

Supplemental Table 5 Number of detectable exonic, intronic, and other non-3'UTR coverage peaks. A peak is detectable in a specific compartment and sample, if it has a minimum of 1 read assigned in each time point measurement, and an average of minimum 30 reads assigned across the time series. Other peaks comprise peaks which are assigned to untranslated regions, to a mixture of exonic, intronic, or untranslated regions, or without annotation.

Erklärung zur Dissertation

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde.

Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Achim Tresch betreut worden.

Freiburg, 14th of February 2022

K. Moos

(Katharina Moos)