UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Dynamic Quality Control by RNA 3’ End Modifying Enzymes

Permalink

https://escholarship.org/uc/item/9fk90675

Author

Nicholson-Shaw, Timothy Luis Cabrera

Publication Date

2022

Supplemental Material

https://escholarship.org/uc/item/9fk90675#supplemental

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Dynamic Quality Control by RNA 3' End Modifying Enzymes

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Biology

by

Timothy Luis Cabrera Nicholson-Shaw

Committee in charge:

Professor Jens Lykke-Andersen, Chair
Professor Eric Bennett
Professor Amy Pasquinelli
Professor Miles Wilkinson
Professor Brian Zid

2022

©

Timothy Luis Cabrera Nicholson-Shaw, 2022

All rights reserved

The dissertation of Timothy Luis Cabrera Nicholson-Shaw is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my favorite Angel, AngelA.

EPIGRAPH

Deeds will not be less valiant because they are unpraised.

J.R.R. Tolkien

TABLE OF CONTENTS

LIST OF SUPPLEMENTAL FILES

Nicholson-Shaw_Labno_Candidates_Supplementary_Table_2.1.csv
Nicholson-Shaw_Son_Candidates_Supplementary_Table_2.2.csv
Nicholson-Shaw_Angel1_Raw_Peptides_Supplementary_Table_3.1.xlsx

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my doctoral advisor, Jens Lykke-Andersen, whose kindness and dedication to mentorship have cultivated an amazing environment to learn, grow, and contribute to the field. Jens' training and support for my own independent thinking have helped me become something resembling a scientist and for that I will be forever grateful.

Thank you to all the past and present members of the Lykke-Andersen lab who have overlapped with me. Together, they have created a supportive and collaborative lab filled with mutual respect. Thank you especially to Rea who helped me get my feet underneath me and has always been amazingly supportive.

Thank you to Berto who joined the lab with me at the same time. He's been an encouraging lab-mate that has always given me perspective that keeps my two feet firmly on the ground.

To all my friends that I've gained over my years in grad school, thank you. Grad school would not have been anywhere near as pleasant without your love and support.

I also want to extend my thanks to the entire division of biological sciences. I appreciate all the wonderful advice, conversations, and reagents shared. A special thank you to the Pasquinelli lab who always had more of that reagent I ran out of on 8pm on a Friday without which, this dissertation would have taken twice as long.

Finally, I would like to thank my wife Angela Nicholson-Shaw. Her love and support mean more to me than I could ever possibly express to her. Without her, this dissertation would not have happened.

Chapter 2, in full, is a reprint of material that is in press at *RNA*. “Tailer: A Pipeline for Sequencing-Based Analysis of Non-Polyadenylated RNA 3’ End Processing,” Nicholson-Shaw, T., Lykke-Andersen, J., 2022. I was the primary author.

Chapter 3, sections 1-6, are a reprint of material that is being prepared for submission. “The 2’,3’ cyclic phosphatase Angel1 facilitates mRNA degradation during human ribosome-associated quality control,” Nicholson-Shaw, T., Ajaj, Y., Perelis, M., Fulzele, A., Yeo, G., Bennett, E., Lykke-Andersen, J., 2022. I was the primary author.

Chapter 4 contains unpublished material and data collected in collaboration with Isabel Breuer, Mark Perelis, Amit Fulzele, Gene Yeo, Eric Bennett, and Jens Lykke-Andersen. I was the primary author of this material.

VITA

2006 – 2010	Bachelor of Science in Biochemistry/Cell Biology University of California, San Diego, La Jolla, CA
2008 – 2012	Research Assistant UC San Diego Medical Center, San Diego, CA
2012 – 2015	Product Development Associate Biolegend Inc., Sorrento Valley, CA
2015 – 2022	Doctor of Philosophy in Biology University of California, San Diego

PUBLICATIONS

Parsons CL, Proctor J, Teichman JS, Nickel JC, Davis E, Evans R, Zupkas P, Phillips C, Shaw T, Naidu N, et al. 2011. A multi-site study confirms abnormal glycosylation in the Tamm-Horsfall protein of patients with interstitial cystitis. J Urol 186: 112–116.

Argade S, Shaw T, Su Y, Parsons CL. 2013. Tamm-Horsfall protein-associated nucleotides in patients with interstitial cystitis. BJU Int 111: 811–819.

Parsons CL, Shaw T, Zupkas P, Argade S. 2013. The role of urinary cations in interstitial cystitis. J Urol 189.

Argade S, Shaw T, Chen T, Zupkas P, Lagares E, Parsons CL, Sur R. 2013. The role of Tamm-Horsfall protein in urinary stone disease. J Urol 189.

Parsons CL, Shaw T, Berecz Z, Su Y, Zupkas P, Argade S. 2014. Role of urinary cations in the etiology of bladder symptoms and interstitial cystitis. BJU Int 114: 286–293.

Argade S, Chen T, Shaw T, Berecz Z, Shi W, Choudhury B, Lowell Parsons C, Sur RL. 2015. An evaluation of Tamm–Horsfall protein glycans in kidney stone formers using novel techniques. Urolithiasis 43: 303–312.

Lardelli RM, Schaffer AE, Eggens VRC, Zaki MS, Grainger S, Sathe S, Van Nostrand EL, Schlachetzki Z, Rosti B, Akizu N, Scott E, Silhavy JL, Heckman LD, Rosti RO, Dikoglu E, Gregor A, Guemez-Gamboa A, Musaev D, Mande R, Widjaja A, Shaw TL, Markmiller S, Marin-Valencia I, Davies JH, De Meirleir L, Kayserili H, Altunoglu U, Freckmann ML, Warwick L, Chitayat D, Blaser S, Ça Layan AO, Bilguvar K, Per H, Fagerberg C, Christesen HT, Kibaek M, Aldinger KA, Manchester D, Matsumoto N, Muramatsu K, Saitsu H, Shiina M, Ogata K, Foulds N, Dobyns WB, Chi NC, Traver

D, Spaccini L, Bova SM, Gabriel SB, Gunel M, Valente EM, Nassogne MC, Bennett EJ, Yeo GW, Baas F, Lykke-Andersen J, Gleeson JG. 2017. Biallelic mutations in the 3’ exonuclease TOE1 cause pontocerebellar hypoplasia and uncover a role in snRNA processing. Nat Genet 49: 457–464.

Nicholson-Shaw T, Lykke-Andersen J. 2021. Tailer: A Pipeline for Sequencing-Based Analysis of Non-Polyadenylated RNA 3’ End Processing. In Press at RNA.

Nicholson-Shaw T, Ajaj Y, Perelis M, Fulzele A, Yeo GW, Bennett EJ, Lykke-Andersen JL. 2022. The cyclic phosphatase Angel1 facilitates RNA degradation during human ribosome-associated quality control. In Preparation.

ABSTRACT OF THE DISSERTATION

Dynamic Quality Control by RNA 3' End Modifying Enzymes

by

Timothy Luis Cabrera Nicholson-Shaw

Doctor of Philosophy in Biology

University of California San Diego, 2022

Professor Jens Lykke-Andersen, Chair

All RNAs in the cell are monitored by robust molecular machinery that ensures its fidelity and quality. Maintenance of the integrity of the transcriptome is vital to the proper functioning of the cell. Dysfunctions in the machinery responsible for this quality control can lead to physiological problems for the cell and wide-ranging diseases in the context of a whole organism. A deeper understanding of these

systems is critically important to illuminate the workings of the cell, multi-cellular organisms, and to provide new therapeutic handles to treat human disease.

The cell contains multiple overlapping quality control pathways to monitor RNAs that are tailored to the type and purpose of a particular RNA. RNAs that do not code for proteins, non-coding RNAs (ncRNAs), are increasingly being understood for the dynamic trimming and tailing that occurs in their 3' end to control their maturation and quality. Advances in next generation sequencing have allowed researchers to examine the changing 3' ends of ncRNAs at nucleotide resolution, but robust, open source, and publicly available tools to analyze this data have been missing from the field. In chapter 2 of this dissertation, I outline a bioinformatic tool dubbed "Tailer" that allows for visualization and quantification of this kind of data. This tool is thoroughly validated on public datasets and faithfully recapitulates the findings of the original studies from which they were drawn.

Another key hub for RNA quality control, specifically for mRNAs, is translation. When translating a dysfunctional mRNA or an mRNA in a difficult context, a ribosome can run into regions that cause a local ribosome stall. This allows a trailing ribosome to collide, signaling to the cell that the nascent protein and mRNA template need to be degraded. In chapter 3, I present evidence that Angel1, a 2',3' cyclic phosphatase, is a novel rate-limiting factor for this ribosome-associated quality control (RQC) pathway. Angel1 associates with proteins important for this process and nucleotide sequences that have been implicated in ribosome stalling. Angel1 depletion also stabilizes reporter substrates that are targets of this pathway. I also show evidence

that N4BP2 is the human ortholog of the RQC endonuclease identified in *Saccharomyces cerevisiae* and *Caenorhabditis elegans*.

Lastly, in chapter 4, I examine Angel2, a protein homologous to Angel1, that was recently identified as a 2',3' cyclic phosphatase. Using discovery-based techniques, I look for clues that could suggest Angel2's function. With IP-MS/MS I find that Angel2 associates with components of the tRNA splicing ligase. Using eCLIP, I find evidence that Angel2 directly associates with tRNAs but is not necessarily enriched for the subset of tRNAs that undergo splicing. Finally, with RNA-seq, I find that Angel2 depletion upregulates mRNAs with a low fraction of rare codons and codons decoded by intron-containing tRNAs.

These studies provide new insights into RNA quality control and provide a leaping off point for future studies into the modification and processing of RNA 3' ends.

Chapter 1

Introduction

1.1 Quality control is critical to gene regulation

RNA is a key mediator between the cell's genetic material, DNA, and how the cell responds to its context and environment. A flurry of experimentation in the late 1950's and early 1960's led to the conclusion that several different types of RNA were essential for the production of proteins (Watson 1963). This concept has been pithily encapsulated as the central dogma; DNA codes for RNA and RNA in turn codes for proteins that perform the functional business of the cell. However, with the sequencing of the human genome (Lander et al. 2001; Venter et al. 2001), we learned that most of our DNA does not code for protein. Work done before this milestone and after has elucidated the wide variety of RNA molecules that exist in the cell and their broad range of functions beyond coding for proteins. This body of work underscores the versatility and fundamental importance of RNA to the cell and our framework for how life exists.

One of the fundamental challenges of the cell is to properly control and modulate this vast and complex pool of RNA known collectively as the "transcriptome." The cell needs to not only maintain the transcriptome but also remain

flexible enough to make sweeping changes in its composition depending on the context of the cell. Importantly, levels of RNA transcripts are tightly controlled by cellular processes and dysregulation of these systems can have detrimental effects on an organism.

Beyond maintaining correct levels of RNA transcripts, cells also need to ensure the quality and fidelity of transcribed RNAs. Cellular machinery, like any other machine, can make mistakes and these mistakes can potentially create dysfunctional transcripts. Furthermore, RNAs are susceptible to damage and being rendered non-productive through normal cellular wear-and-tear, such as through self-cleavage or damage by environmental agents like UV light or chemicals. Recognition and disposal of these problematic transcripts is necessary to prevent trapping cellular machinery in unproductive events and the accumulation of toxic products. Problems with the machineries necessary for this function are associated with many disease states (Nguyen et al. 2014; Frischmeyer and Dietz 1999; Terrey et al. 2020; Lardelli et al. 2017).

1.2 Non-coding RNA quality control by 3' modifying enzymes

Rather than code for protein, the majority of RNA in the cell participates in critical cellular functions including splicing by small nuclear RNAs (snRNAs), direction of RNA modifications by small nucleolar RNAs (snoRNAs), translation using

ribosomal RNAs (rRNAs), and gene silencing by microRNAs (miRNAs) (Eddy 2001). Given their importance to fundamental cellular processes, a robust system to ensure proper protein loading, modification, and localization is necessary to protect their function. An important way that cells regulate quality for non-coding RNAs is through modification of their 3' ends. The 3' ends of ncRNAs are dynamic and will often change many times over the life of the transcript. Modifications include addition of nucleotides post-transcriptionally through the action of terminal end nucleotidyl transferases (TENTs) (Scott and Norbury 2013; Aphasizhev et al. 2016; Heo et al. 2012; Berndt et al. 2012), trimming of genomically encoded nucleotides or nucleotides added post-transcriptionally (Lardelli et al. 2017; Son et al. 2018; Shukla and Parker 2017), and through chemical modification of the terminal nucleotide of RNA (Gu et al. 1997).

Depending on the type of modification, timing, location, and context, 3' end modification can have drastically different functional outcomes. Trimming of ncRNAs has been demonstrated to promote maturation as is the case with snRNAs that are trimmed after cleavage by the mediator complex (Lardelli et al. 2017) or trimming can trigger decay by more processive 3'–5' exonucleases (Łabno et al. 2016). Post-transcriptional addition of nucleotides has been shown to promote maturation (Berndt et al. 2012), promote decay (Scott and Norbury 2013; Aphasizhev et al. 2016), be necessary for proper function as is the case with tRNAs (Deutscher 1973), or have yet to be determined consequences as is the case for the single A addition given to 7SL RNA (Sinha et al. 1998).

One prevailing view of ncRNA maturation is that it is a balance or competition between factors that promote maturation and those that promote degradation (Lardelli and Lykke-Andersen 2020). ncRNAs are thought to progress through maturation at a steady rate. However, if for some reason maturation slows or stops—potentially because of a dysfunctional ncRNA or improper loading of proteins—it can allow factors that promote degradation to “win out” in the balance and push the scales toward decay (Figure 1.1). The factors that are important for this dance between maturation and degradation are poorly described. Understanding the factors involved in this competition and the consequences of their actions is vital to understanding the life cycle of ncRNAs.

The eukaryotic proteome contains many enzymes that have been shown to or are predicted to alter the 3’ end of ncRNAs. Which ncRNA targets are regulated by a particular 3’ modifying enzyme and how the enzyme shapes that 3' end is a fundamental question in the field of ncRNA quality control and maturation. A typical paradigm for investigating this question is to deplete or remove a potential processing enzyme and then use sequencing techniques to identify changes in 3’ ends (Łabno et al. 2016; Son et al. 2018). Problems of this approach include that it is often extremely targeted and therefore low throughput and that the tools used to analyze these datasets are often lab, protocol, and machine specific. This requires labs working in this field to create their own unique scripts and bioinformatic workflows potentially leading to errors and variability between groups. Chapter two of this dissertation outlines and makes available to the public an open-source tool, Tailer, that standardizes this type of workflow and allows for rapid visual analysis.

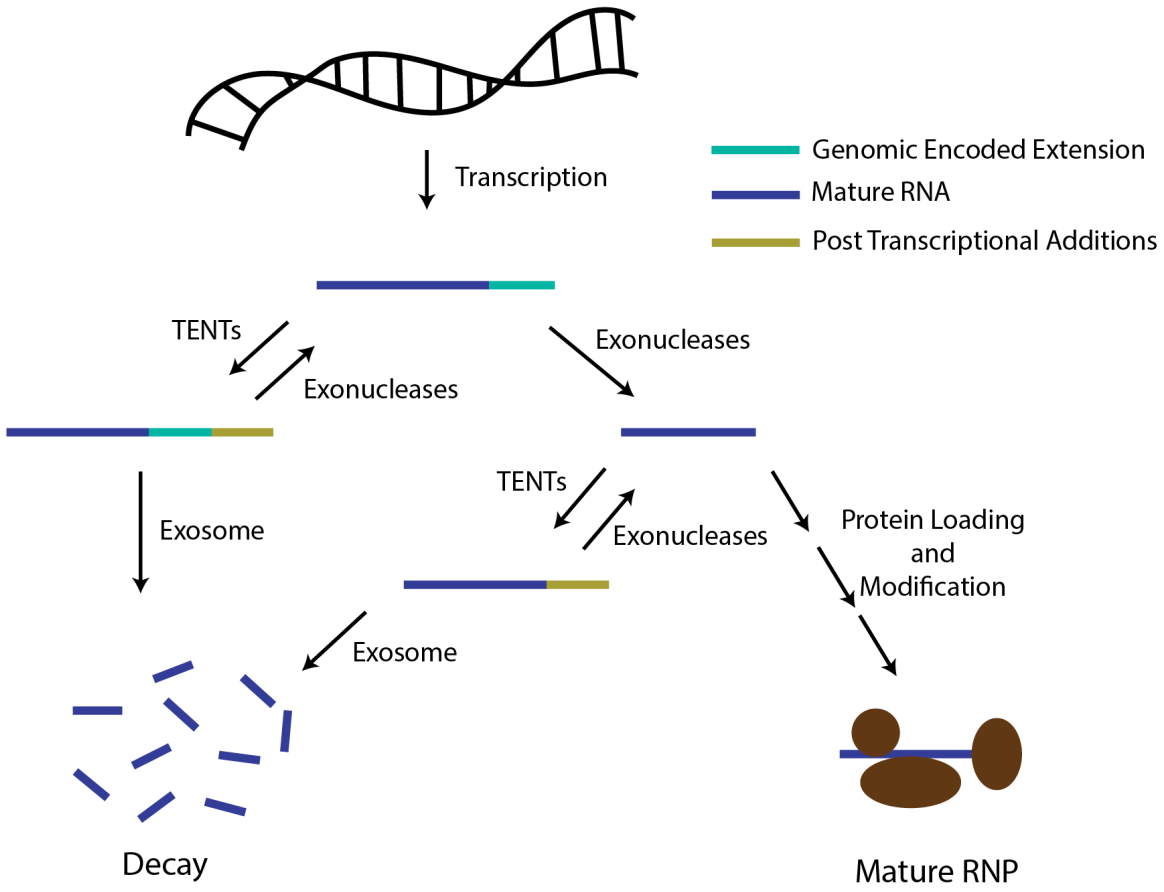


Figure 1.1: Examples of ways that 3' end tailing is thought to control ncRNA quality control and maturation. Non-coding RNAs (ncRNAs) are often transcribed with genomic encoded 3' end extensions. This 3' end is thought to be competed over by the action of exonucleases and Terminal End Nucleotidyl Transferases (TENTs) which can promote decay or maturation. At many steps of maturation, problems are thought to allow the action of subsequent TENTs which promote decay.

1.3 mRNA surveillance by the ribosome

While ncRNAs are often surveilled for quality by modifications to their 3' ends, a major route of quality control for RNAs made from protein coding genes (mRNAs) is

through the ribosome. During translation, the cell can sense irregularities in the function of the ribosome that imply problems with the mRNA. This often triggers disposal of the potentially problematic nascent protein and the troublesome mRNA.

The best understood example of this kind of surveillance is the Nonsense-Mediated Decay pathway (NMD), responsible for degrading mRNAs containing premature termination codons (PTCs) (Maquat 2004). PTCs are often created by mistakes made during splicing and are sensed by a ribosome that terminates translation prematurely. In human cells, this is often while an exon junction complex is still on the mRNA, which can be detected by the NMD machinery (Lykke-Andersen et al. 2000; Kishor et al. 2019). PTCs can also be introduced through genomic mutations that create a stop codon in the coding region of a gene and are the main cause of many congenital diseases, such as cystic fibrosis (Nguyen et al. 2014).

Another mechanism of ribosomal surveillance, which is less well understood, is the Ribosome-associated Quality Control (RQC) pathway. RQC begins when a ribosome stalls during translation. This allows a trailing ribosome to catch up and collide with it, presenting a unique surface that is recognized by the ubiquitin ligase ZNF598 (Sundaramoorthy et al. 2017; Juszkiewicz et al. 2018). Collided ribosomes indicate a problem to cellular machinery, triggering ubiquitylation of the nascent polypeptide chain by another ubiquitin ligase Ltn1 and decay by the proteasome (Bengtson and Joazeiro 2010). In certain contexts, this can also trigger decay of the mRNA, though much is still unknown about the factors that promote this.

The types of issues that can stall ribosomes are broadly grouped into two pathways based on historical rather than functional reasons: No-Go Decay and Non-

Stop Decay (Doma and Parker 2006; Frischmeyer et al. 2002; Van Hoof et al. 2002) (Figure 1.2). No-Go Decay typically refers to mRNA degradation that occurs upon stalling of ribosomes in the coding region of an mRNA. Stall-inducing contexts include strong secondary structural elements (Bao et al. 2020; Endoh and Sugimoto 2016), certain amino acid motifs (Huter et al. 2017), rare codons, lack of decoding tRNAs (Yang et al. 2019), and oxidative or UV damage to an mRNA (Simms et al. 2014). Non-stop decay refers to mRNA decay triggered by translation events that stall because the ribosome never encounters a stop codon; this can be caused by truncation in the coding sequence (CDS) or an improper polyadenylation event (Frischmeyer et al. 2002; Van Hoof et al. 2002). This latter event leads to translation into the poly-A tail and stalling of the ribosome through a recently elucidated mechanism (Chandrasekaran et al. 2019). In either case, the signal is fundamentally the same, a stalled ribosome that undergoes collision by a trailing ribosome.

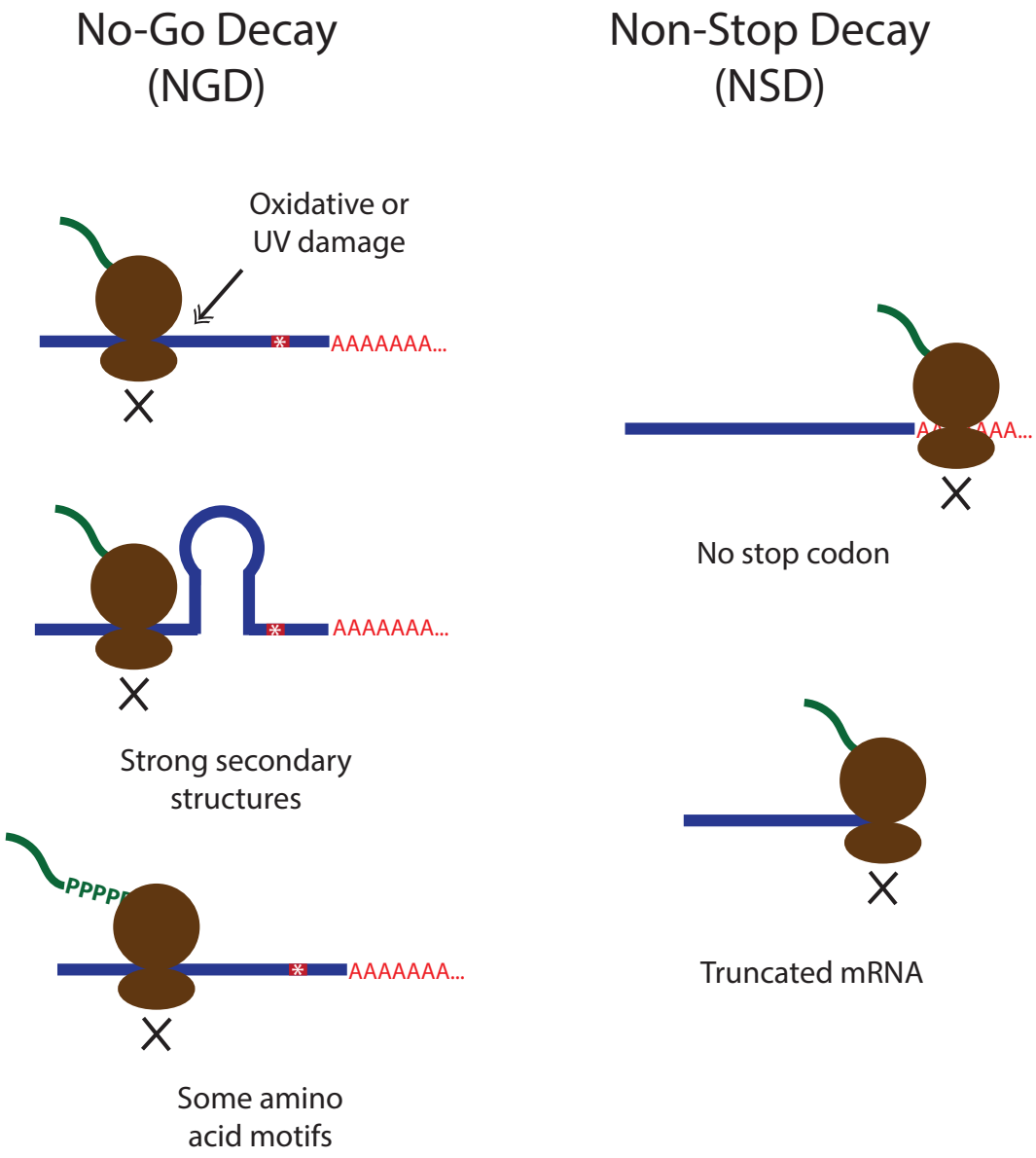


Figure 1.2: Features of transcripts that can cause ribosome stalling. A wide variety of features can cause the ribosome to stall depending on environment and context. NGD typically refers to ribosomes that becomes stalled in the coding region, while NSD typically refers to ribosomes that never encounter a stop codon and cannot be released from the mRNA. Either way, the functional output is the same, a collided ribosome that triggers downstream decay of the nascent polypeptide and mRNA.

While mechanisms of protein degradation during RQC are well studied, the mechanisms concerning disposal of the mRNA are poorly described in humans. We have the most insight into this process through work done in the budding yeast *Saccharomyces cerevisiae* (Goldman et al. 2021). Decay of the mRNA substrate is often initiated by cleavage at the site of the stall by the recently identified endonuclease, Cue2 in budding yeast (D'Orazio et al. 2019), and Nonu-1 in *Caenorhabditis elegans* (Glover et al. 2020). The 3' fragment is then degraded in the 5'–3' direction in a process dependent on the exonuclease Xrn1 (Tsuboi et al. 2012). The 5' fragment is ratcheted out of the ribosome by the activity of the Ski helicase complex and is degraded in the 3'–5' direction by the cytoplasmic exosome (Zinoviev et al. 2020) (Figure 1.3).

Collisions in the coding transcriptome appear to happen often and at repeatable motifs (Chyzyńska et al. 2021). However, it seems that these collisions are not always monitored by RQC and do not necessarily lead to decay of the mRNA. On the other hand, some evidence appears to indicate that endonucleolytic cleavage is widespread in the transcriptome and may be the outcome of stalling events (Ibrahim et al. 2018). The factors that determine the fate of these problematic mRNAs, the enzymatic machinery, and the mechanisms of mRNA decay are poorly understood, especially in humans. Chapter 3 of this dissertation shows the use of a reporter substrate to interrogate this system in human cells and the identification of a novel factor, Angel1, that is involved in the disposal of NSD substrates.

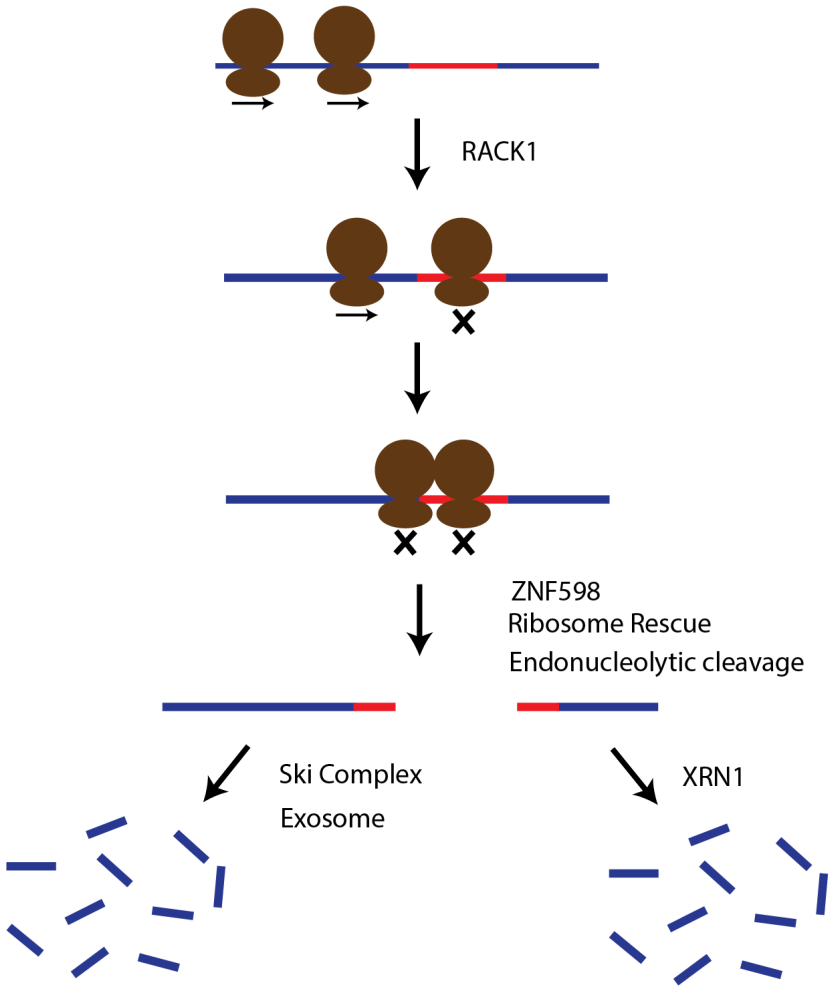


Figure 1.3 Mechanism of RNA decay by RQC. While translating an mRNA, a ribosome can reach a problematic region, causing the ribosome to stall. This allows a trailing ribosome to collide with it, presenting a unique surface that is recognized and ubiquitylated by ZNF598. This triggers ribosome rescue and cleavage of the RNA substrate by an endonuclease (Cue2 in budding yeast and Nonu-1 in *C. elegans*). Decay of the 5' fragment is facilitated by the Ski complex and the exosome. Decay of the 3' fragment is facilitated by XRN1.

1.4 Cyclic phosphates are a poorly understood 3' terminal modification

The last nucleotide on an RNA molecule typically terminates in a 3'OH. However, certain endonucleases, 3'–5' exonucleases, and random self-cleavage of RNAs can leave behind a 2',3' cyclic phosphate (Shigematsu et al. 2018) (Figure 1.4). The best known ncRNA with this modification is U6 snRNA whose 3' end is trimmed to a correct number of uracil nucleotides by the action of USB1, leaving behind a 2',3' cyclic phosphate (Mroczek and Dziembowski 2013). The function of the terminating cyclic phosphate is poorly understood but may help to stabilize U6 snRNA and protect its 3' end from the action of exonucleases.

2',3' cyclic phosphates are difficult to study in part because RNA species terminating in them are essentially invisible to the most common types of sequencing protocols (Shigematsu et al. 2018). Most library preparation methods that capture non-polyadenylated RNAs begin with a ligation step. Because 2',3' cyclic phosphate terminating species are not capable of being ligated by typical enzymes, they are lost early in library preparation. Capture of RNAs with cyclic phosphates requires a mild acid hydrolysis or treatment with poly-nucleotide kinase (PNK) (Honda et al. 2016) which will remove the cyclic phosphate and leave behind a 3'OH. However, information about which species originally terminated in a cyclic phosphate will be lost. A protocol called cP-seq was recently developed, allowing for specific capture and sequencing of cyclic phosphate-containing species. cP-seq leverages a periodate treatment that destroys the ends of any RNA that does not contain a cyclic phosphate (Honda et al. 2016).

While the existence of cyclic phosphates have been known for some time, identification of eukaryotic enzymes capable of removing a cyclic phosphate have not been described until recently (Pinto et al. 2020). Using biochemical fractionation and an *in vitro* cyclic phosphatase assay, Angel2 and, to a lesser extent, Angel1 were identified as enzymes in human cells that are capable of this function. Further investigation of the functions of Angel2 and Angel1 require assays that will allow us to measure proportions of cyclic phosphate-containing RNA species in the cell. Chapter four of this dissertation explores potential roles for Angel2 and outlines a protocol that may be used to measure changes in cyclic phosphorylation in the cell.

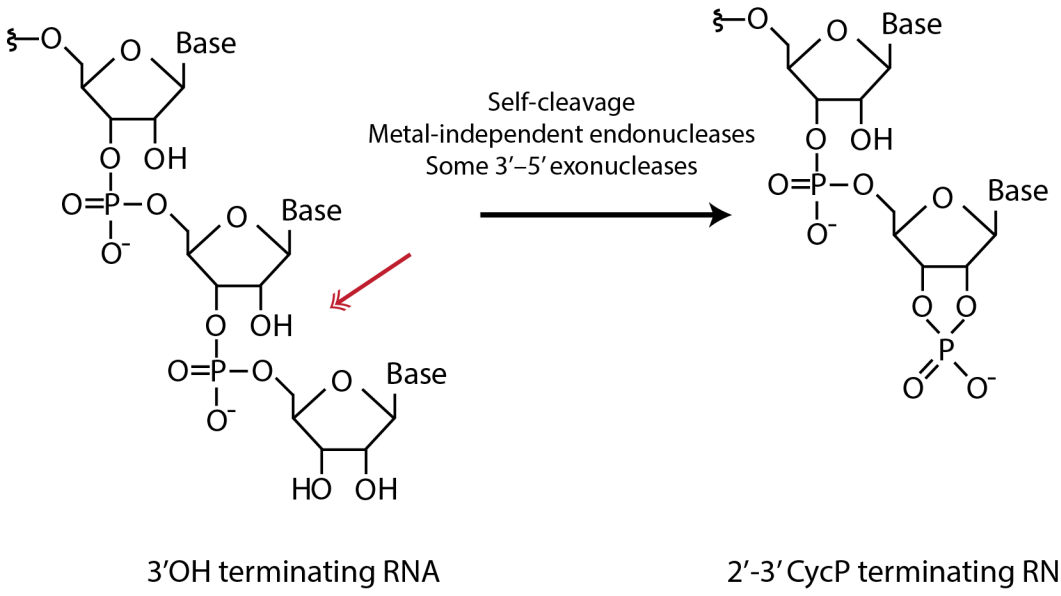


Figure 1.4: Structure of 3'OH and 2',3' Cyclic Phosphate terminating RNA species. RNA typically terminates in a 3'OH. Certain cleavage reactions will leave behind a cyclic phosphate.

Chapter 2

Tailer: A Pipeline for Sequencing-Based Analysis of Non-Polyadenylated RNA 3’ End Processing

2.1 Abstract

Post-transcriptional trimming and tailing of RNA 3’ ends play key roles in the processing and quality control of non-coding RNAs (ncRNAs). However, bioinformatic tools to examine changes in the RNA 3’ “tailome” are sparse and not standardized. Here we present Tailer, a bioinformatic pipeline in two parts that allows for robust quantification and analysis of tail information from next generation sequencing experiments that preserve RNA 3’ end information. The first part of Tailer, Tailer-Processing, uses genome annotation or reference FASTA gene sequences to quantify RNA 3’ ends from SAM-formatted alignment files or FASTQ sequence read files produced from sequencing experiments. The second part, Tailer-Analysis, uses the output of Tailer-Processing to identify statistically significant RNA targets of trimming and tailing and create graphs for data exploration. We apply Tailer to RNA 3’ end sequencing experiments from three published studies and find that it accurately and reproducibly recapitulates key findings. Thus, Tailer should be a

useful and easily accessible tool to globally investigate tailing dynamics of non-polyadenylated RNAs and conditions that perturb them.

2.2 Introduction

Dynamic post-transcriptional addition and removal of nucleotides from the 3' ends of RNAs is a key hub for RNA maturation and regulation. While these dynamics are perhaps best understood for eukaryotic mRNAs that undergo polyadenylation (Lee et al. 1971; Darnell et al. 1971; Edmonds et al. 1971) and deadenylation (Goldstrohm and Wickens 2008) to regulate translation and stability (Nicholson and Pasquinelli 2019), non-coding RNAs also experience a wide variety of 3' end modifications. These events, which include 3' end trimming, tailing, or chemical modification (Yu and Kim 2020; Liudkovska and Dziembowski 2021; Perumal and Reddy 2002), can have different functional consequences depending on the RNA, the modification, and the cellular context. Some 3' end modification events, exemplified by CCA addition to tRNAs (Deutscher 1973), play key roles in RNA maturation and serve to produce mature 3' ends that promote RNA stability and/or function (Dupasquier et al. 2008; Katoh et al. 2009; Nguyen et al. 2015; Shukla and Parker 2017). Other modifications promote rapid degradation, for example as part of quality control pathways that detect and degrade aberrant or damaged transcripts (LaCava et al. 2005; Shcherbik et al. 2010; Liu et al. 2014; Lardelli and Lykke-Andersen 2020). These processes are essential to life and their dysfunction can lead to human disease (Wolin and Maquat 2019); yet, how enzymes acting at RNA 3'

ends cooperate and compete to dictate RNA function and stability remains poorly defined for the majority of RNAs.

Early characterizations of non-coding (nc)RNAs and their 3' end sequences focused on single RNA species, initially using radioisotope labeling and enzymatic digestions, and later, RNA 3' end amplification methods coupled with cloning and sequencing (Rinke and Steitz 1982; Frohman et al. 1988; Lund and Dahlberg 1992). More recent advances in sequencing technology have allowed for examination of ncRNA ends on a transcriptome-wide level, and for monitoring how those ends change globally in response to perturbations. Techniques such as ligation-based 3' Rapid Amplification of cDNA Ends (3'RACE) coupled with high-throughput sequencing (Lee et al. 2014; Shukla and Parker 2017) can provide a snapshot at nucleotide level resolution of RNA 3' ends globally. A typical reverse genetics approach to understanding RNA 3' end dynamics involves identifying enzymes capable of modulating RNA tails, depleting them from cells, and monitoring changes in RNA 3' ends, thereby identifying potential direct targets of those enzymes (Allmang et al. 1999; Berndt et al. 2012; Łabno et al. 2016; Lardelli et al. 2017; Son et al. 2018; Lardelli and Lykke-Andersen 2020). During data analysis, changes to RNA 3' ends are generally quantified with scripts and pipelines individual to each lab. While some of these scripts have been made publicly available (for example (Welch et al. 2015; Pirouz et al. 2019)), easy-to-use and generalizable tools have been missing to make these types of analyses accessible to the broader research community.

Here we present Tailer, an easy to use and open-source pipeline that can analyze the status and perturbations of non-coding RNA 3' ends from sequencing

datasets for which RNA 3’ ends have been preserved. Tailer is fully featured, easily installable, and allows for analysis of new and previously published datasets. This pipeline takes mapped SAM or BAM files from 3’ end sequencing experiments, globally identifies positions and compositions of RNA 3’ ends, including their post-transcriptional tails, and outputs the data into a human readable CSV format. This output CSV file can then be uploaded to a web server, which provides utilities to discover RNAs undergoing statistically significant changes at their 3’ ends and to visualize RNA tail dynamics. The pipeline also allows for analysis of individual RNAs of interest from global or gene-specific sequencing experiments using local alignment.

To validate Tailer, we reanalyzed publicly available global and gene-specific 3’ end sequencing datasets from three studies focused on the exonucleases DIS3L2, TOE1 and PARN in human cells (Łabno et al. 2016; Son et al. 2018; Lardelli and Lykke-Andersen 2020). In all cases, Tailer identified target RNAs highlighted in the studies and faithfully reproduced observed effects on RNA 3’ ends. This validates the utility of Tailer as a tool to monitor global and gene-specific 3’ end processing of non-coding RNAs. While applied here to human RNA sequencing datasets, the pipeline is compatible with datasets from any organism of interest with reliable annotation information.

2.3 Results and Discussion

2.3.1 Pipeline Overview

Tailer is comprised of two arms (Fig. 1.1), Tailer-Processing, which identifies and quantifies 3' end compositions of non-polyadenylated RNAs from 3' end sequencing data, and Tailer-Analysis, an R-based Shiny app for candidate discovery and data visualization. Tailer is written in Python 3 (Van Rossum G and Drake FL. 2019), can be installed using the Package Installer for Python (PIP) accessed from the PyPi index (detailed installation instructions can be found on the readme page), and can be run from the command-line. The output of Tailer-Processing is a comma separated values (CSV) file, hereafter referred to as a Tail CSV file, which lists the identity and quantity of all 3' ends of RNAs observed in the analyzed 3' sequencing experiment that match a given annotation file, or a given list of genes. The Tail CSV file can then be fed into the Shiny-based (Chang et al. 2021) Tailer-Analysis web application for further analysis.

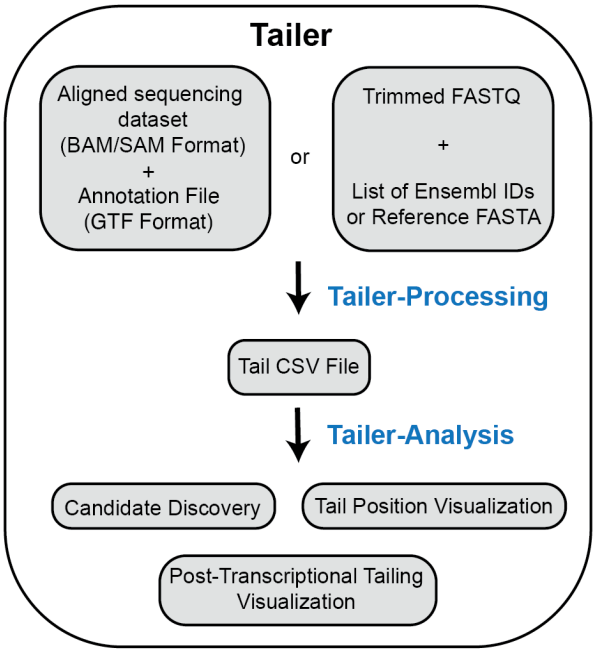


Figure 2.1: A general overview of Tailer's workflow. Tailer is split into two major parts, a processing function, and an analysis webserver. Tailer-Processing infers RNA 3' ends using a BAM/SAM alignment file and a GTF formatted annotation file, or a FASTQ sequence read file and a reference FASTA gene file (which can be generated from Ensembl IDs). For either method, the output is a standardized Tail CSV file, which can be analyzed directly, or fed into the Tailer-Analysis web server for discovery of candidate tailing changes in comparison between datasets as well as visualization of tails with a variety of graphing tools.

2.3.2 Tailer-Processing in global mode annotates SAM/BAM files and calculates RNA 3' end information

Tailer-Processing can be run from the command-line and can be used in a global mode to identify all RNA 3' ends matching a genome annotation, or in local mode to identify 3' ends of specific RNAs of interest (Fig. 2.2A). When used in global

mode (Fig. 2.2B, left), Tailer-Processing requires two inputs, a SAM or BAM formatted alignment file and a GTF formatted annotation database. Experimentally, the sequencing data entered into Tailer-Processing should be generated using a library preparation method that preserves the 3’ end information of RNAs, such as a 3’ RACE (Frohman et al. 1988) experiment, and should be performed on RNA that is not poly(A)-selected. For small RNA 3’ end analyses, the RNA can be size selected prior to sequencing, but analyses can be performed on any sequencing experiment that preserves RNA 3’ ends. For longer RNAs, some platforms, such as current Illumina platforms, require nucleotide input below a certain size range, which would require a method that made use of 5' truncation prior to sequencing such as internal upstream priming or limited RNase-treatment after ligation of a 3’ adapter. Sequencing can be performed either as single end reads from the 3’ end, or as paired end reads for improved alignment accuracy. Sequencing outputs need to be pre-processed by trimming of any adapters and linkers and removal of PCR duplicates, and subsequently aligned to a reference genome using any aligner that supports soft-clipping and produces a SAM/BAM-formatted alignment file output. Care should be taken to ensure that pre-processing does not introduce any artifacts such as improper trimming, which would lead to incorrect 3’ end calls. It is important that the aligner supports soft-clipping as Tailer uses this feature to determine post-transcriptional tailing (for options that bypass soft-clipping, see below). Typically, we use STAR aligner (Dobin et al. 2013) with the following settings from Son, et al 2018 when interested in small non-coding RNAs (--alignIntronMin 9999999 --outFilterMultimapNmax 1000), which allows alignment to multicopy genes and

disallows unannotated introns. The GTF file can be provided to Tailer as a full genome annotation or filtered to contain only genes of interest to create smaller sized output files. In case of paired-end sequencing, the specific read that corresponds to the RNA 3' end needs to be specified with the "--read" flag. This pipeline has been most rigorously tested with annotations provided by the Ensembl database (Howe et al. 2021).

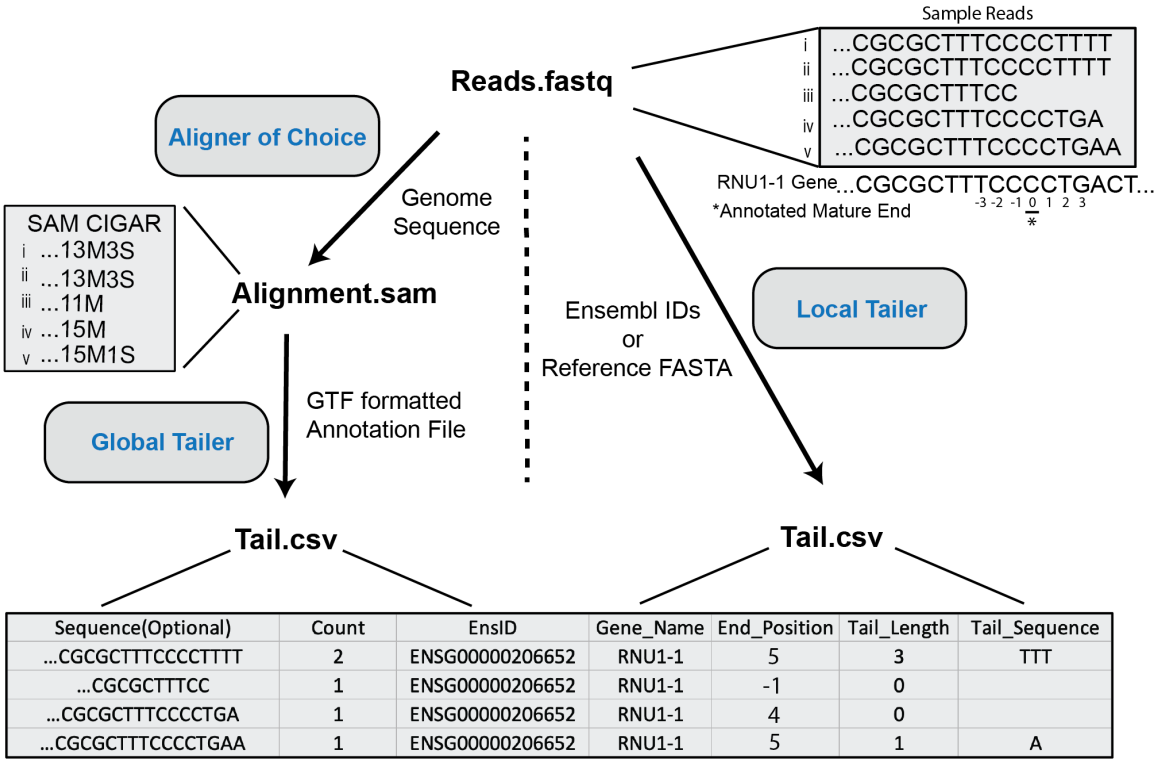
A

```
~ % Tailer -a annotation.gtf [SAM/BAM Files]

~ % Tailer -e [Comma separated Ensembl IDs] [Trimmed FASTQs]

~ % Tailer -f [Reference FASTA] [Trimmed FASTQs]
```

B



Sequence(Optional)	Count	EnsID	Gene_Name	End_Position	Tail_Length	Tail_Sequence
...CGCGCTTTCCCCTTTT	2	ENSG00000206652	RNU1-1	5	3	TTT
...CGCGCTTTCC	1	ENSG00000206652	RNU1-1	-1	0	
...CGCGCTTTCCCCTGA	1	ENSG00000206652	RNU1-1	4	0	
...CGCGCTTTCCCCTGAA	1	ENSG00000206652	RNU1-1	5	1	A

Figure 2.2: Tailer-Processing commands and examples of tail inference. (A) Example commands to run Tailer-Processing. After installation with PIP, Tailer-Processing can be invoked by typing “Tailer” into the command line. An “-h” flag will provide usage information that is also available in a readme.txt file. Examples of running Tailer-Processing in global mode, in local mode using Ensembl IDs, and local mode with a reference FASTA gene file are shown. (B) Global Tailer-Processing (left) uses SAM/BAM-formatted alignment files to infer RNA 3’ ends and post-transcriptional tails based on the SAM CIGAR and GTF annotation files. Local Tailer (right) does not require pre-alignment of the sequencing data and uses BLAST to align to a user provided FASTA gene database or one generated from provided Ensembl IDs. Local Tailer makes use of reported BLAST metrics of last query mapping position and last reference mapping position to infer tail information. Both modes produce a Tail CSV file with the same columns (bottom). An example output from five reads aligning to the RNU1-1 gene with corresponding SAM-file CIGAR strings is shown.

The output tail CSV file produced by Tailer-Processing reports the number of occurrences of each type of RNA 3' end that is detected in the sequencing data (Fig. 2.2B, bottom). For each alignment reported in the input SAM alignment file, Tailer-Processing identifies the corresponding gene from the input annotation GTF file, identifies the 3' end position of the read relative to the annotated gene 3' end, and predicts any post-transcriptionally added tail. The gene, which is reported in the 'Gene' column, is identified as the gene in the same orientation as the aligned read that has the closest annotated 3' end to the 3'-most aligned nucleotide of the read, with a requirement that the 3' ends are within a default of 100 nucleotides of one another, an option that can be modified with the "-t" flag. To identify the read 3' end position and predict any post-transcriptional tail, Tailer examines the CIGAR string reported in the SAM alignment file and searches for soft-clipping at the 3' end of the read, i.e. the 3' terminal nucleotides of the read that did not align to the genome (Fig. 2.2B, left). The length and composition of the soft-clipped nucleotides are reported as the post-transcriptional tail of the read in the output tail file ("Tail_Length" and "Tail_Sequence" columns). The position, including any post-transcriptional tail, of the last nucleotide of the read relative to the annotated 3' end of the gene is reported as the 3' end position of the read ("End_Position" column). For multi-mapped reads, all annotated genes aligning to the read are reported in the output file, which allows for more accurate downstream analyses of RNAs that are produced from multiple loci or have closely related pseudogenes (see below). In cases where reads align to multiple genes that are annotated with different 3' ends, Tailer reports only the gene

whose annotated 3’ end is closest to the calculated genome-encoded 3’ end of the read. Reads corresponding to identical RNA 3’ end sequences are finally combined and the number of reads for each are reported in the “Counts” column. This output format greatly reduces the size of the file to focus on information that is pertinent to tail analysis and can be reasonably uploaded to a web server. An optional column with 3’ end read sequences that can be useful for verification and troubleshooting purposes can be included in the tail file by implementing an “-s” flag.

2.3.3 Running Tailer-Processing in local mode allows for rapid analysis of specific RNAs without the necessity for previous alignment or reliance on soft-clipping

Analysis by Tailer also lends itself to a gene-specific approach for greater depth on specific genes of interest using local alignments (Fig. 2.2B, right). This mode requires the user to have command line BLAST installed and a reference to it stored in the PATH variable on their workstation (https://www.ncbi.nlm.nih.gov/books/NBK279690/). The required inputs are a FASTQ file containing the called bases from the sequencing experiment, trimmed of any linkers and PCR duplicates, and one or more genes, either identified by their Ensembl IDs or provided in a FASTA file. For paired-end sequencing, the FASTQ file used should be the read file that corresponds to the 3’ end (typically read 1 for Illumina sequencing).

This mode is most useful for analyses of gene-specific 3' end sequencing data (Lardelli et al. 2017; Lardelli and Lykke-Andersen 2020). It is also useful in cases where soft-clipping is problematic for correct alignments (Suzuki et al. 2011), such as for genes that have closely related variants. In this case, initial global sequence alignments can be performed in the absence of soft-clipping and reads aligning to specific genes can subsequently be extracted from the SAM/BAM alignment files and converted back to FASTQ files with tools such as Bedtools (Quinlan and Hall 2010) and Samtools (Li et al. 2009). The local gene-specific Tailer can also be used directly on large FASTQ files from global sequencing experiments, but this is not recommended as processing will be much slower than using global Tailer. The gene-specific mode downloads gene information from Ensembl along with 50 nucleotides of downstream sequence to aid in distinguishing between genome-encoded tails and post-transcriptionally added tails. Alternatively, a custom FASTA-formatted reference sequence can be provided instead with the "-f" flag (Fig. 2.2A). This reference sequence should contain genomic sequence downstream of the gene for accurate distinction between genomic-encoded and post-transcriptional tails. When including downstream sequence, the "-m" flag should be used to specify the number of downstream nucleotides included in the reference to ensure that the mature end is correctly annotated.

After building a BLAST formatted database with the downloaded sequences, gene-specific Tailer aligns each read in the FASTQ file and uses the alignment information to calculate the read 3' end position relative to the annotated gene 3' end and identify the composition of any predicted post-transcriptional tail, producing an

output Tail CSV file identical to that produced by global Tailer-Processing. It is important to note that for both the global and local methods, post-transcriptional tails are predictions based on absence of alignment. These generally represent the most conservative predictions for the actual post-transcriptional tail, since any post-transcriptionally added nucleotide that matches a nucleotide encoded from the genome will be assigned as genome-encoded by default.

2.3.4 Using Tailer on published datasets identifies ncRNA tails and compresses them into a human readable, portable, CSV format

To develop and validate this workflow, we used global 3' end sequencing data from two previously published datasets, Labno et al. 2016 (hereafter called the Labno dataset) which investigated targets of human DIS3L2, and Son et al. 2018 (the Son dataset) which investigated targets of human PARN and TOE1. We also used a gene-specific 3' end sequencing dataset from Lardelli and Lykke-Andersen 2020 (the Lardelli dataset), which investigated snRNA targets of human TOE1. After producing SAM alignment files using STAR with settings discussed above, Tailer, using a full genome annotation file (Ensembl 104), reduced gigabyte (Gb)-sized SAM files into megabyte (Mb)-sized Tail CSV files (Table 2.1), which makes uploading and analyzing on a web server practical. Output Tail CSV file sizes can be further reduced by using subset annotations with only genes of interest. The Tail CSV file can be used directly for visualization or analysis for users experienced with this type

of data, or it can be fed into the Tailer-Analysis web app described below, or used directly in R with individual Tailer-Analysis functions available from the GitHub repository.

Table 2.1: Dataset Summary

Reference	Method	Protein of Interest	Identified Targets	Average Read Depth	Average Alignment File Size	Average Tail File Size
Łabno, et al. 2016. "Perlman Syndrome Nuclease DIS3L2 Controls Cytoplasmic Non-Coding RNAs and Provides Surveillance Pathway for Maturing SnRNAs." Nucleic Acids Research 44 (21): 10437–53.	3' RACE	Dis3L2	snRNAs, Vault RNAs, YRNAs	6.25M	794Mb	3.4Mb
Son, et al. 2018. "PARN and TOE1 Constitute a 3 End Maturation Module for Nuclear Non-Coding RNAs End Maturation Module for Nuclear Non-Coding RNAs." CellReports 23: 888–98.	3' RACE	TOE1 and PARN	18S rRNA, snoRNAs, TERC, scaRNAs, snRNAs	21.4M	19.5Gb	80Mb
Lardelli and Lykke-Andersen 2020. "Competition between Maturation and Degradation Drives Human SnRNA 3′ End Quality Control." Genes & Development 34 (13–14).	Gene-specific 3' seq	TOE1	pol-II snRNAs	4.9K	NA	0.2 Mb

2.3.5 Tailer-Analysis: A Shiny webapp for candidate discovery and 3’ end data visualization

Tailer-Analysis provides a simple, user-friendly, and open-source GUI and is built using the R Shiny library (Chang et al. 2021). As an input, this analysis app takes the Tail CSV files generated from Tailer-Processing as described above. Multiple tail files can be uploaded in the “Tail File Upload” tab, including different experimental conditions to be compared, and experimental replicates (Fig. 2.3A). Using the table interface, users can enter metadata information which will group and average replicates, allowing for easy downstream comparisons and statistical analyses.

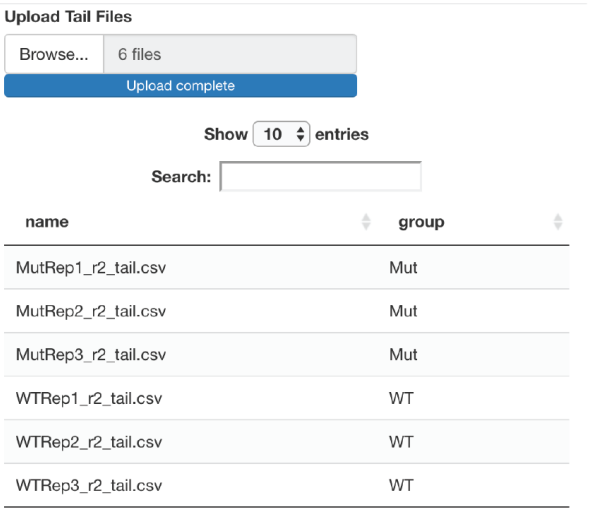
On the “Candidate Finder” tab, users can compare two of their experimental conditions (Fig. 2.3B). RNAs with significant changes at their 3’ ends are found by

comparing two conditions to look for statistically significant differences. The app first generates a list of all genes identified by Tailer-Processing. For each gene, the replicate data is then pooled and 3’ end positions and tail lengths are compared between experimental conditions using a Kolmogorov-Smirnov (KS) test. Pooled replicate data is used in the Candidate Finder to allow for greater computational throughput. However subsequent modules maintain separation between replicates for greater statistical power. Candidate genes are reported in order of p-value for changes in 3’ end position but can also be sorted based on changes in tail length. This helps distinguish between conditions that may be affecting the trimming or extension of RNA molecules in general, versus conditions that affect post-transcriptionally added nucleotides specifically. Candidates can be filtered by minimum number of observations, p-value, and magnitude of difference in end position.

Figure 2.3: Example screenshots of the Tailer-Analysis Shiny interface. (A) Individual sample tail files can be uploaded to the webserver. Using the table interface, users can set grouping metadata, which will be used to bin replicates. After selecting the format data option, the user is provided with feedback on the conditions provided and number of samples. The user is also able to alter the order in which samples will be displayed using a simple drag and drop interface. (B) After uploading and setting metadata, users can use the Candidate Finder tab to rank RNAs based on their changes in tailing between the uploaded datasets. Reads can be filtered by minimum number of observations, magnitude of difference between conditions, and p-value. Hits are reported and ordered initially by statistical changes in RNA 3' end positions but can also be re-ordered by statistical changes in post-transcriptional tail length by selecting the corresponding column. The candidate data can be downloaded and saved as a CSV file. (C) Every graph page contains an options side panel, which can be used to set the desired gene to be graphed and set different parameters for graphing. A checkbox for multi-locus genes is available to enable a slower but more accurate analysis of RNAs produced from multiple loci (see also Fig 2.6 below).

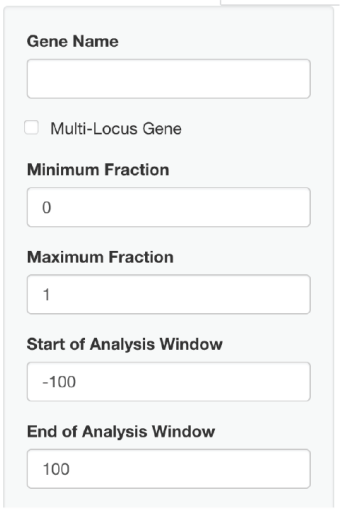
A

File Upload Interface



C

Options Panel



B

Candidate Finder Output

Gene	Δ End Position	p-value End Position	Δ PT Tail	p-value PT Tail	# Reads Condition 1	# Reads Condition 2
VTRNA1-1	7.65748	0	3.86457	0	76	205
RNU1-3|RNU1-1|RNU1-4|RNU1-2|U1	7.52727	0.00021	4.57364	0.00271	12	119
RNY3	6.30894	0.0055	0.88364	0.1304	504	669
RNU6ATAC	3.23707	0.00102	0.72633	0.14353	308	441
RMRP	-6.39112	0.01651	0.57754	0.4633	625	1131
RNY3|Y_RNA|Y_RNA	-20.77835	0	-1.66004	0.74934	384	106
RNVU1-27	13.41182	0.03912	0.26727	0.93107	66	190

Tail files generated from the Labno dataset were uploaded to the Shiny App web server and binned into a WT or a Mutant DIS3L2 condition. Using the built-in candidate discovery tool, a list of candidates with a minimum of 10 observed reads was generated (Supplemental Table 2.1). Among the top candidates with significantly altered 3' ends were Vault RNA-1 (VTRNA-1), Y3 RNA (RNY3), and U6atac snRNA (U6ATAC), all of which were identified by Labno et al. Similarly, tail files from the Son dataset were subjected to candidate analysis using the Tailer-Analysis webapp. Identified potential targets (Supplemental Table 2.2) included many snoRNAs and scaRNAs which were targets also identified by Son et al. Thus, the Tailer pipeline faithfully recapitulates the identification of small RNA targets of 3' end processing enzymes from published studies.

2.3.6 Rapid visualization of 3' end dynamics with the Tailer-Analysis webapp

The remaining tabs in the shiny app each correspond to graphs that can be used to individually explore RNA 3' ends. 3' ends of individual RNAs, either identified from the candidate discovery tool or of specific interest to the user, can be visually analyzed and compared between experiments as described in more detail below. These graph functions are written in ggplot2 (Wickham 2016; R Core Team 2021). For each graph, position 0 corresponds to the annotated mature RNA 3' end. In cases where the mature 3' ends of RNAs are incorrectly annotated in the provided annotation file, the position of the mature 3' end can be manually adjusted using the

options panel (Fig. 2.3C). Each plot also has an analysis window option whereby the user can limit their analysis to specific windows of 3' end mapping. This can be used to exclude potential truncated RNAs from the analysis. Plotting and examining individual RNAs of interest can help distinguish between spurious hits and actual biological targets and can confirm that length changes are in the predicted direction and are of a sufficient magnitude to warrant further investigation.

The first two graphs visualize 3' end positions of the sequenced population of the selected RNA. A bar graph gives a distribution of where the 3' ends of the sequence reads are mapping in relation to the annotated mature 3' end (Fig. 2.4A-C). Grey bars represent the positions, as fractions of the overall population, of the last genome-encoded nucleotide of the plotted RNA population. The colored bars represent the fraction of RNA molecules that contain post-transcriptionally added nucleotides at the indicated positions, broken down by nucleotide identity. It is important to emphasize, as detailed above, that post-transcriptional tails predicted by Tailer are the most conservative post-transcriptional tails based on the alignments.

Plotting VTRNA-1 from the Labno dataset recapitulates the presence of a post-transcriptional tail that consists primarily of uridines (dark blue bars) in the absence of DIS3L2 activity, which is observed on about half of the population and extends from the mature 3' end (position 0), ranging from one to over ten uridines (Fig. 2.4A). Plotting U1 snRNA from the Lardelli dataset demonstrates accumulation of extended U1 snRNAs that are partially tailed with adenosines (light blue bars) in the absence of TOE1 (Fig. 2.4B). Furthermore, plotting SCARNA-22 from the Son dataset recapitulates the accumulation of extended RNA species terminating at the

+10 position that accumulate with oligo-adenosine tails upon PARN and TOE1 depletion and a synergistic extension when both are depleted (Fig. 2.4C).

The second graph is a cumulative plot, which shows the cumulative fraction of RNA reads that map to specific 3' end positions (Fig. 2.4D). Solid lines represent cumulative 3' end positions of the RNA population when including post-transcriptional tails, and dotted lines represent the predicted 3' ends excluding the post-transcriptional tails, with shading in between representing the extent of post-transcriptional tailing. The cumulative plots are particularly useful for comparing effects of different experimental conditions on specific RNAs in single graphs. Visualizing VTRNA1, SCARNA22 and U1 snRNA using this tool in Figure 2.4D recapitulates the overall extension of these transcripts upon depletion of the respective exonucleases, as observed by the overall right-shifts of the corresponding step plots. It can also be seen, by examining the extent of shading, that in the case of DIS3L2 inactivation (top panel) most of the difference in the VTRNA-1 3' end is accounted for by differences in post-transcriptional tailing, consistent with the observations from the original study, whereas for PARN and TOE1 depletion (bottom two panels), effects are seen on both genome-encoded and post-transcriptional nucleotides of the target RNAs consistent with these enzymes trimming both post-transcriptional tails and genome-encoded nucleotides (Son et al. 2018; Lardelli and Lykke-Andersen 2020).

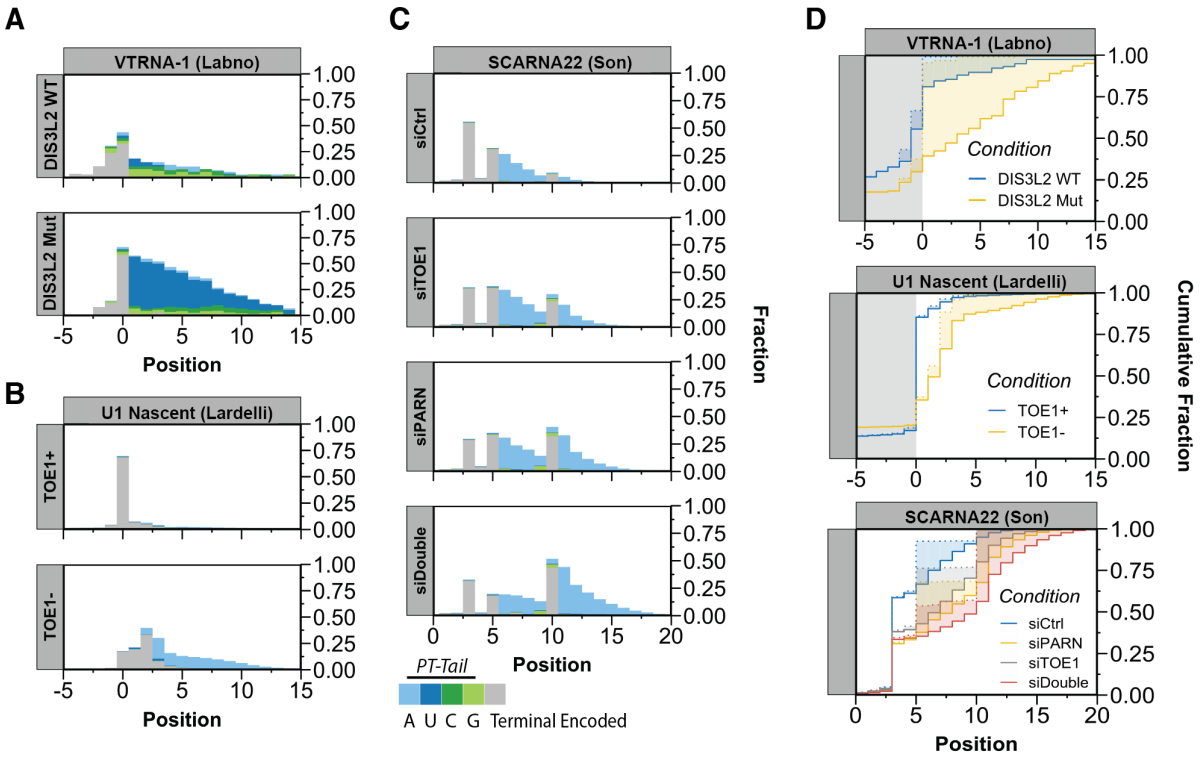


Figure 2.4: Sample plots of RNA 3' end dynamics in response to processing factor depletion. (A-C) Tail bar graphs for indicated RNAs from the Labno (*A*), Lardelli (*B*) and Son (*C*) datasets, with grey bars showing the position of the terminal genomic encoded nucleotide as a fraction of the RNA population, and stacked colored bars showing the fraction of the RNA population containing post-transcriptional nucleotides at the indicated positions. (D) Cumulative plots displaying the cumulative fraction of overall 3' end positions of the indicated RNA populations (including any post-transcriptional tail; solid lines) and the 3' terminal genome-encoded nucleotide (dotted lines) with shading in between indicating the extent of post-transcriptional tailing. Dots to mark individual experiments can be toggled on and off using the options panel (Supplemental Figure 2.1).

2.3.7 Using Tailer-Analysis to visualize post-transcriptional tails

The next set of graphs focus on information concerning predicted post-transcriptionally added tails. The first graph is a logo plot containing information about proportions and compositions of post-transcriptional tails, with the 1 position corresponding to the first nucleotide of the tail and the height of each nucleotide representing the fraction of the RNA population that contains the modification (Fig. 2.5A-C). Plotting the datasets from Figure 2.4 in this manner reveals oligo-U tails that accumulate on VTRNA-1 in the absence of DIS3L2 activity (Fig. 2.5A) and oligo-A tails that accumulate on U1 snRNA (Fig. 2.5B) and SCARNA22 (Fig. 2.5C) in the absence of TOE1 and/or PARN activities. A background of primarily guanosines (denoted by a star in Fig. 2.5A) observed on VTRNA-1 appears, upon inspection of individual reads, to originate from an unknown linker in the Labno dataset.

The final graph shows the average number of post-transcriptional adenosines, uridines, guanosines, and cytidines found per read for the RNA of interest (Fig. 2.5D). In cases where replicates are included, this graph will show dots representing each experiment, bars for standard deviation, and, optionally, a p-value from a Student's T-test. When applied to the analyzed datasets, these graphs again highlight the U-tailing observed for VTRNA-1, and A-tailing for SCARNA22 and U1 snRNAs upon depletion of the respective exonucleases. The source R code for generating all four types of graphs is available from our GitHub repository, are well documented, and can be imported and used in an active R session. Furthermore, below each graph is an option to download the raw plot data in CSV format. This option facilitates graphing using alternative software.

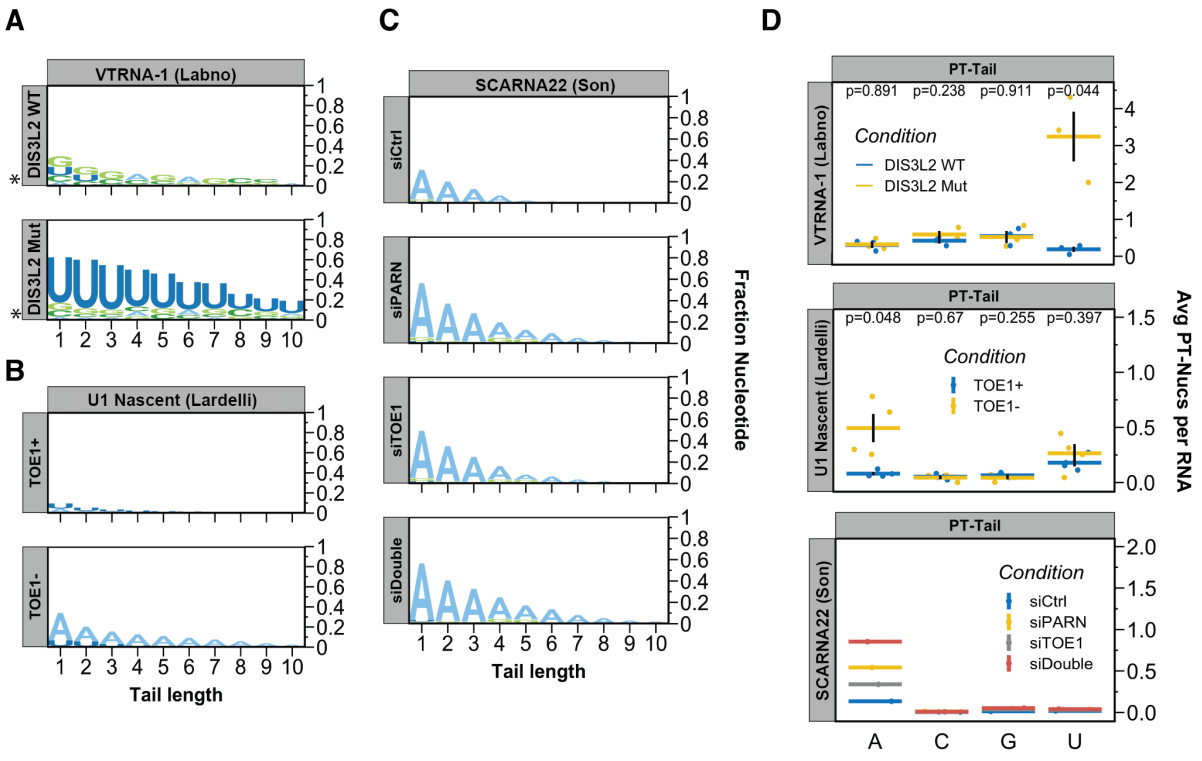


Figure 2.5: Sample plots of post-transcriptional tailing created by Tailer-Analysis. (A-C) Logo plots showing compositions of indicated RNA post-transcriptional tails as a fraction of the overall population. (D) Graphs showing average number of individual post-transcriptional nucleotides per RNA molecule as horizontal bars, with values for individual experiments shown by dots. In cases where multiple replicates are included (i.e. the Labno and Lardelli datasets), vertical lines show standard error of the mean (SEM) between experiments, and p-values from Student's T-tests are reported to monitor significance.

2.3.8 Statistical outputs

The final tab of Tailer-Analysis contains utilities for testing statistical significance between groups (Supplementary Figure 2.2). After selecting two conditions to compare, the user is presented with tables of pair-wise KS-tests between each replicate in each condition. Statistical testing is done for both overall end position and total post-transcriptional tail length, which, as noted above, can help

to distinguish between perturbations that affect post-transcriptional tailing only and those that also affect genome encoded tails. The page will also output a KS-test after pooling all replicates in each condition.

2.3.9 Inclusive alignments to multi-loci genes prevents spurious tailing calls

A subset of small ncRNAs are produced from multiple loci in the genome, which in many cases are identical to one another except for their downstream sequences. Forcing multi-locus RNA reads to map to single loci can lead to tails being falsely called that originate from the downstream sequence of a different locus from which the RNA was actually transcribed. As an example, human U1 snRNA originates from multiple active genes. Using any single locus in the Tailer analysis leads to the calling of spurious C- and U- post-transcriptional tails, which actually originate from other transcribed loci (Fig. 2.6A and B). In order to accurately assign reads, both global and local modes of Tailer allow for all loci to be considered when analyzing 3' end tails, in which case the spuriously called C- and U-tailing of U1 snRNA is much reduced as reads are mapped to their proper loci (Fig. 2.6A). This demonstrates the importance of considering information from all gene loci when analyzing 3' tailing data. Since post-transcriptional tail calls are conservative based on best alignment fits, analyses using multiple loci are more likely to miss a subset of actual short post-transcriptional tails, but, importantly, they help reduce the rate of false positive calls as observed for U1 snRNA.

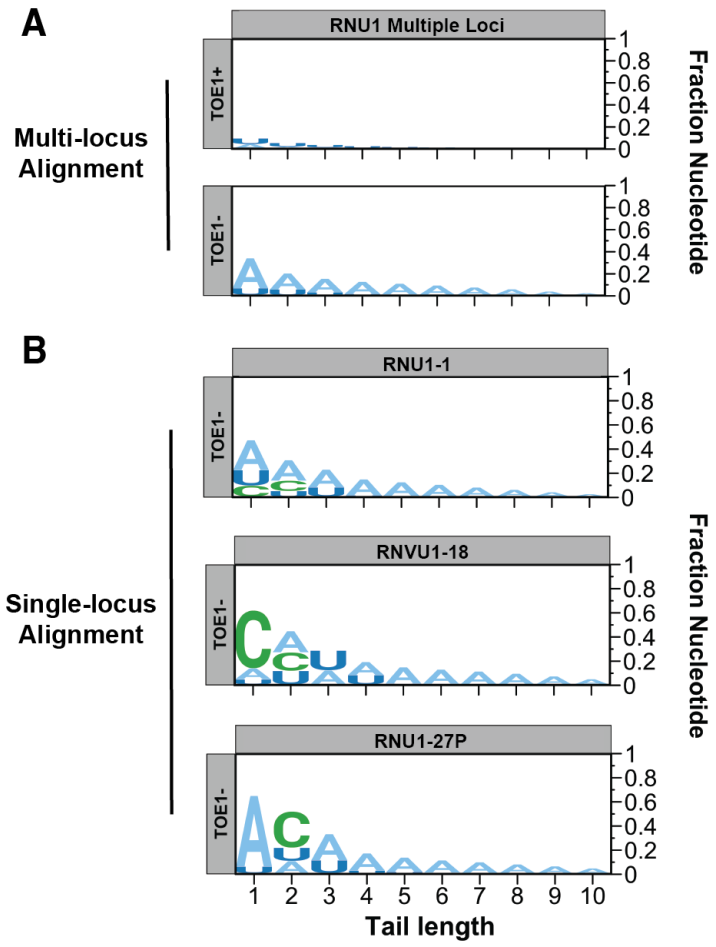


Figure 2.6: Logo plots of tails of a multi-locus RNA called with incomplete locus information. (A) U1 snRNAs from the Lardelli dataset aligned to all regular U1 loci of the genome, showing accumulation of post-transcriptional A-tails in the absence of TOE1. (B) The same dataset as in panel *A*, aligned to single U1 snRNA gene loci, leading to erroneous post-transcriptional tail calls.

2.4 Conclusions

Trimming and tailing of RNA 3' ends play key roles in the processing and quality control of non-coding RNAs. Advances in deep sequencing technologies and methods for library preparation have provided the tools to generate hordes of RNA 3'

end data. However, tools to analyze these types of data have remained limited. We developed Tailer to help spur inquiry into this important regulatory mechanism with a particular focus on ease of installation and use. Distribution of Tailer-Processing through PyPi allows for quick and easy installation in a wide variety of environments without end-users needing to manage dependencies and compatibility. Furthermore, users are not required to work with and manipulate genome or annotation data, such as making artificial genomes with singular loci for each gene. Tailer-Analysis as a web server allows users with no experience in R or coding to upload and explore datasets, and open-source distribution of the code allows for more advanced users to work more rapidly with their data using R.

It is important to note that with current protocols, because RNA 3' end information is typically preserved using a ligation step, biases in the data are likely introduced due to effects of RNA structure or sequence on ligation efficiencies (Fuchs et al. 2015). Furthermore, certain RNAs terminate with a 3' end modification that inhibits ligation (e.g. U6 snRNA terminating in a 2'-3' cyclic phosphate (Gu et al. 1997; Honda et al. 2016)), which needs to be removed prior to ligation to prevent exclusion of the RNA from the analysis, or bias of the analysis towards a state of maturation that does not contain the modification. In addition, since post-transcriptional tails are distinguished from nucleotides introduced during transcription by genome alignment, post-transcriptional tails can be missed, particularly short ones. Biases can also be introduced based on RNA length. For example, sequencing using current Illumina platforms requires short amplicons, which necessitates a truncation step after the initial 3' end ligation step (such as Rnase treatment or

internal upstream priming), for analysis of 3' ends of long transcripts. Lastly, depending on the choice of library preparation, some amplicons could arise from mis-priming events within the cDNA sequence rather than the adapter (Roy and Chanfreau 2020). This should reveal itself as an apparent truncated transcript in Tailer-Analysis, which can be excluded from the analysis by limiting the analysis window. Thus, analysis strategies based on ligation-mediated sequencing lend themselves best to monitoring RNA-specific changes to 3' ends between different cellular conditions (such as depletion of processing factors), rather than measuring accurate levels of tailing of one RNA over another. However, the development of direct RNA sequencing methods (Byrne et al. 2017), which can also readily be analyzed by the tools developed here, promise to alleviate many of these concerns.

Through a combined approach of local and global alignment, Tailer can, reproducibly and transparently, address many of the issues common with working with non-coding RNA sequencing data including the analysis of RNAs produced from multiple loci. Other approaches for analysis of RNA 3' end processing have been published including one specifically for microRNAs (Newman et al. 2011), one published for use with circular RACE data (Pirouz et al. 2019) and AppEnD (Welch et al. 2015), which was used to examine histone mRNAs but in principle could be applied to many different types of RNAs with non-templated additions. Compared to AppEnD, Tailer does not require a linker to be present in the sequencing data to identify 3' ends, which facilities analysis of sequencing data deposited without linker information. Tailer also comes with a robust graphing and visualization suite for 3' end data that is unique to this pipeline. The Tailer suite is validated by extensive

analysis of datasets from three different studies using three different methods of library preparation. Thus, Tailer should allow more investigators to enter this research space and improve our understanding of this important mechanism of RNA regulation.

2.5 Materials and Methods

2.5.1 Tailer-Processing and Tailer-Analysis Access

Source code for Tailer-Processing, the Tailer-Analysis Shiny app, further examples, and usage instructions can be found on our GitHub (https://github.com/TimNicholsonShaw/tailer and https://github.com/TimNicholsonShaw/tailer-analysis) and are available for use under the MIT license. Tailer-Analysis is available as a web server at https://timnicholsonshaw.shinyapps.io/tailer-analysis/.

2.5.2 Data pre-processing

Data for these analyses was obtained from the NCBI GEO repository using the FASTQ-dump utility (Labno: GSE82336, Son: GSE111511, Lardelli: GSE141709). The Labno dataset, which was reported in the NCBI GEO repository trimmed of linkers, was aligned without modification using the STAR aligner as described above. The Son dataset contained a four nucleotide 3’ adapter sequence which was trimmed using the FASTQ/A Trimmer from the FASTX-Toolkit and then aligned as above. The

Lardelli dataset needed to have a 13-nucleotide barcode trimmed from the 3' end which was performed using options provided by Tailer-Processing's local mode and described in more detail in the repository's readme (-x 13 flag).

2.5.3 Tailer-Processing global

Global Tailer-Processing begins by generating a Search Query Language (SQL) database of all genes in the annotation file using the GFFutils module which allows for rapid look up. Tailer then reads in the provided SAM/BAM file using the PySam module, iterates through every read, discarding members of read pairs that originate from the 5' end of the RNA (typically read 2 in a paired-end sequencing experiment, which does not provide reliable information about the 3' end of the original molecule), and tags them with every gene that they overlap within all their possible alignments using the SQL database, while combining identical reads together. For each aligned gene, tail position and composition are inferred using the soft-clipping flag in the CIGAR string of the SAM/BAM file and the annotated 3' end of the gene from the GTF annotation file. This analysis of soft-clipping approach is fundamentally identical to the approach taken by the authors of the Labno and Son datasets. Tail information is compared for all possible genes and the gene that gives an alignment closest to the annotated 3' mature end is reported. In cases where multiple genes produce identical tail information, all genes are reported. The resulting tail information is then written to a CSV file referred to as a Tail CSV file.

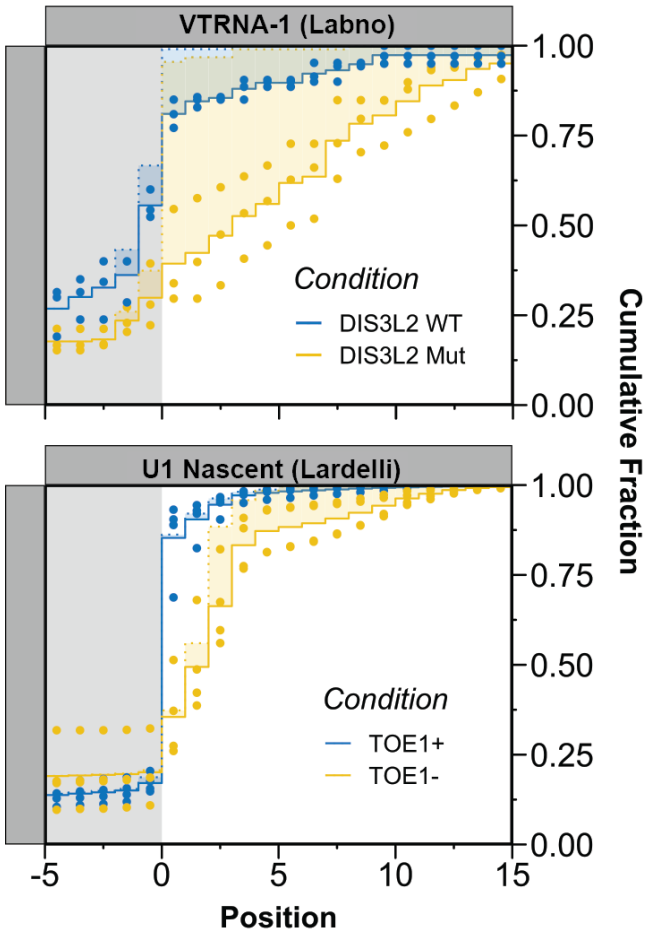
2.5.4 Tailer-Processing local

If provided with Ensembl IDs of interest, Tailer-Processing in local mode will contact Ensembl servers (requiring an internet connection), download their gene sequences via Ensembl's REST API, and build a FASTA file. If provided with a reference FASTA file, it will instead use that reference, which should include downstream sequence for accurate distinction between encoded and post-transcriptional tails, in which case the length of the downstream sequence should be indicated with the -m flag for accurate annotation of the mature 3' end. This is done automatically when providing an Ensembl ID. The mature end can also be adjusted later in the options panel of Tailer-Analysis. Using the command line BLAST utility makeBlastDb, Tailer creates a database compatible with BLAST searches. Tailer then uses the query FASTQ to generate a BLAST compatible query file and, using command-line blast, searches the query against the reference outputting the results in JSON format. After parsing the output, Tailer infers tails for each aligned read using alignment to the reference sequence and reports the tail for the gene(s) whose 3' end is closest to the 3' end of the gene. The resulting tail information is then written to a Tail CSV file. With the largest dataset, the Son dataset (Table 2.1), on a 2 GHz Quad-Core Intel i5 processor, Tailer-Processing takes approximately 30 minutes to complete.

2.5.5 Tailer-Analysis

Tailer-Analysis takes Tail CSV files generated above and metadata provided by the user indicating replicate groups and creates a singular data frame in long format (1 observation per row). This data frame is then fed into the other tools provided by Tailer-Analysis. For the candidate finder, replicates from two different groups are pooled and compared with a K-S test which is reported for End Position and for Tail Length. This list is sorted by End Position P-value and reported to the user. Tail bar graphs are initiated by creating a matrix of frequencies of each nucleotide or genome encoded end at every requested position. This matrix is fed to ggplot's geom_bar function and faceted based on the experimental condition. Cumulative plots are created by calculating cumulative sums at each position for both End Position (total tail length with post-transcriptional additions) and End Position minus Tail Length (location of the genome-encoded end). This data is summarized and averaged based on condition and position and fed to a geom_step ggplot function. The tail logo grapher calculates nucleotide frequencies at all requested positions and feeds the frequency matrix to ggseqlogo's geom_logo function (Wagih 2017). The post-transcriptional nucleotide graph is created by first finding the number of each nucleotide in the Tail Sequence column for each sample and calculating the mean count of each nucleotide per RNA molecule. Data is then summarized based on condition and nucleotide and fed to ggplot using geom_jitter for dots, geom_segment for lines, and geom_errorbar (SEM reported). Uniform theming is accomplished with a single defined common theme that is applied to all graphs and can be reviewed on the GitHub repository for Tailer-Analysis.

2.6 Supplemental Figures



Supplementary Figure 2.1: Recreation of cumulative plots from Figure 2.4D with optional replicate dots. Replicate dots can be toggled on and off using the options panel.

VTRNA1-1 (Labno)

End Position KS-test Matrix

	DIS3L2 WT_1	DIS3L2 WT_2	DIS3L2 WT_3
DIS3L2 Mut_1	4.829023E-05	4.339632E-04	2.187131E-01
DIS3L2 Mut_2	7.084303E-04	1.295540E-03	3.120647E-01
DIS3L2 Mut_3	9.286356E-06	1.645482E-04	2.532410E-01

Pooled End Position KS-test

1.852351e-09

Tail Length KS-test Matrix

	DIS3L2 WT_1	DIS3L2 WT_2	DIS3L2 WT_3
DIS3L2 Mut_1	2.377555E-05	7.524774E-05	2.187131E-01
DIS3L2 Mut_2	1.222320E-03	2.284022E-03	6.107870E-01
DIS3L2 Mut_3	5.426585E-06	4.051894E-05	1.465849E-01

Pooled Tail Length KS-test

1.34335e-09

Number of Observations

Condition	n
DIS3L2 WT_1	21
DIS3L2 WT_2	20
DIS3L2 WT_3	35
DIS3L2 Mut_1	54
DIS3L2 Mut_2	118
DIS3L2 Mut_3	33

RNU1-1 (Lardelli)

End Position KS-test Matrix

	TOE1+_1	TOE1+_2	TOE1+_3	TOE1+_4
TOE1-_1	<1e-50	<1e-50	<1e-50	<1e-50
TOE1-_2	<1e-50	<1e-50	<1e-50	<1e-50
TOE1-_3	<1e-50	<1e-50	2.81840850657744e-09	<1e-50
TOE1-_4	<1e-50	<1e-50	<1e-50	<1e-50

Pooled End Position KS-test

<1e-50

Tail Length KS-test Matrix

	TOE1+_1	TOE1+_2	TOE1+_3	TOE1+_4
TOE1-_1	<1e-50	<1e-50	3.30846461338297e-14	8.8159308742064e-09
TOE1-_2	<1e-50	<1e-50	9.64485158405637e-11	4.19450031485802e-07
TOE1-_3	<1e-50	<1e-50	0.000124527617351933	0.000211237539438636
TOE1-_4	<1e-50	<1e-50	6.50890452646991e-12	2.96518487430397e-07

Pooled Tail Length KS-test

<1e-50

Number of Observations

Condition	n
TOE1+_1	3495
TOE1+_2	4576
TOE1+_3	2567
TOE1+_4	8044
TOE1-_1	1704
TOE1-_2	1355
TOE1-_3	1644
TOE1-_4	1285

Supplementary Figure 2.2: Example output from Tailer-Analysis' statistics tab. After selecting a gene of interest, each replicate in a condition is compared pairwise to every other replicate in the comparison condition using a KS test. Comparisons are performed for the End Position metric and the Tail Length metric described above. Statistics are also performed after pooling all samples to get a singular pooled KS test metric. The number of observation in each condition is also reported in a table.

2.7 Acknowledgments

We would like to thank the Triton Shared Compute Cluster (TSCC) at the San Diego Supercomputer Center for use of their hardware for alignments. We would also like to thank Dr. Elly Poretsky for enlightening conversations concerning Shiny apps, Dr. Brian Tsu for Pythonic-based encouragement and general enthusiasm, and

members of the Lykke-Andersen lab, Alberto Carreño, Cody Ocheltree, and Tiantai Ma for feedback and testing. This work was supported by National Institutes of Health (NIH) grant R35 GM118069 awarded to J. L.-A.

Chapter 2, in full, is a reprint of material that is in press at *RNA*, “Tailer: A Pipeline for Sequencing-Based Analysis of Non-Polyadenylated RNA 3’ End Processing,” Nicholson-Shaw, T. Lykke-Andersen, J., 2022. I was the primary author.

Chapter 3

The 2',3' cyclic phosphatase Angel1 facilitates mRNA degradation during human ribosome-associated quality control

3.1 Abstract

The Ribosome-associated Quality Control (RQC) pathway serves to resolve ribosomes stalled during the translation process and degrade the associated mRNA and nascent polypeptide. The machinery responsible for RQC-mediated mRNA degradation in human cells remains poorly understood. Here we identify the 2',3' cyclic phosphatase Angel1 as a rate-limiting factor in this process. We find that Angel1 associates with proteins of the RQC pathway and with mRNA coding regions, consistent with a factor that monitors the translation process. Angel1 depletion causes stabilization of reporter mRNAs that are targeted for RQC by the absence of stop codons, but not a reporter mRNA targeted for nonsense-mediated decay, an unrelated mRNA quality control pathway that is also translation-dependent. Angel1 catalytic activity is critical for its function in RQC, as a catalytic inactivating mutation causes loss of RQC function. We also identify a rate-limiting role in RQC for N4BP2, a human homolog of the recently identified RQC endonuclease. Given the biochemical activity of Angel1 as a 2',3' cyclic phosphatase, our observations suggest

that the resolution of a cyclic phosphate is a rate-limiting step in RQC-mediated mRNA decay.

3.2 Introduction

Faithful and accurate expression of the cell's repertoire of protein coding genes is vital to proper cellular function. However, certain cell contexts and problematic mRNA substrates can, if unresolved, lock up translational machinery in unproductive events or create potentially toxic non-functional protein products. Quality control pathways for these aberrant translation events are necessary to maintain the integrity of the proteome (Ermolaeva et al. 2015; Doma and Parker 2007; Chen et al. 2011). One such pathway is ribosome-associated quality control (RQC), which is promoted by local stalling of a ribosome during translation (Joazeiro 2017) and has been shown to be sensed by a trailing ribosome colliding with the stalled ribosome (Simms et al. 2017). The collided ribosomes create a unique surface that is recognized by the E3 ubiquitin ligase ZNF598 (Hel2 in the budding yeast *Saccharomyces cerevisiae*) (Juszkiewicz et al. 2018; Ikeuchi et al. 2019). This activates a cascade of events that results in degradation of the nascent polypeptide as well as the mRNA (D'Orazio and Green 2021).

mRNAs targeted by RQC are broadly categorized into two pathways for historical reasons rather than any clear mechanistic distinction. One is No-Go Decay (NGD), which is defined by translational stalls that occur in the coding region (Doma and Parker 2006), for example due to strong RNA structures including G-

quadruplexes (Bao et al. 2020; Endoh and Sugimoto 2016), certain amino-acid tracts (Huter et al. 2017), or oxidative damage (Simms et al. 2014). Another is Non-Stop Decay (NSD) in which the ribosome stalls at the end of the mRNA because it never encounters a stop codon (Frischmeyer et al. 2002; Van Hoof et al. 2002); this can occur due to mRNA truncation (Pisareva et al. 2011), or an improper polyadenylation event, such as at a coding region cryptic poly(A) site, which leads to translation into the poly(A) tail causing a stall through a recently elucidated mechanism (Chandrasekaran et al. 2019). In either case, collided ribosomes are what appear to be functionally sensed by the RQC system.

Our best understanding of the decay of RQC-targeted mRNAs comes from budding yeast. It has been long understood that decay of NGD substrates in budding yeast involves endonucleolytic cleavage at the site of the stall (Doma and Parker 2006). Endonucleolytic cleavage is believed to also be an initiating event for NSD substrates (Glover et al. 2020), but cleavage would occur close to the mRNA 3' end which is technically difficult to monitor. The 5’ RNA fragment is subsequently degraded in a process dependent on the exosome and its associated Ski complex, whereas the 3’ fragment is degraded in a manner dependent on the 5'-to-3' exonuclease Xrn1 (Frischmeyer et al. 2002; Tsuboi et al. 2012). A recent study found that RQC mRNA substrates in budding yeast also undergo degradation independently of endonucleolytic cleavage in an Xrn1-dependent manner (D’Orazio et al. 2019).

The RQC-mediated mRNA decay pathway is poorly understood in mammals and likely has departures from the pathway in budding yeast given that mammals

lack an ortholog of Ski7, an important component of budding yeast NSD (Van Hoof et al. 2002). Ribosome collisions or pauses appear to be common in mammals and occur at predictable motifs (Han et al. 2020). There also appears to exist a large amount of cleaved, native mRNAs created potentially by endonucleolytic cleavage during stall resolution (Ibrahim et al. 2018). Furthermore, NSD mRNA substrates have been shown to be unstable in Hela cells and require the ribosome rescue factors HBS1 and DOM34, and the Exosome-SKI complex for decay (Saito et al. 2013).

Recent studies have identified Cue2 in budding yeast, and its homolog NONU-1 in the worm *Caenorhabditis elegans*, as the endonuclease responsible for the initial cleavage event of NGD substrates (D'Orazio et al. 2019; Glover et al. 2020). Humans have a putative homolog of Cue2/NONU-1, called N4BP2, but its function in RQC has yet to be confirmed. Cue2 and NONU-1 share homology with endonucleases that cleave RNA in a divalent metal ion-independent manner and generate RNA fragments terminating in 5' hydroxyl and 2',3' cyclic phosphates (D'Orazio et al. 2019; Glover et al. 2020). Indeed, endonucleolytic cleavage of RQC substrates was found to produce 5' hydroxylated 3' RNA fragments (Navickas et al. 2020).

RNAs terminating with 2',3' cyclic phosphates are produced by many RNA processing events, including a subset of endonucleolytic cleavage and 3' end trimming events. Yet, little is known about how this modification affects RNA metabolism. A recent study identified the human Angel2 and Angel1 proteins as 2',3' cyclic phosphatases (Pinto et al. 2020). The Angel protein family is widely conserved in eukaryotes and evolutionarily related to the catalytic CCR4 subunit of the CCR4-

NOT deadenylase complex (Goldstrohm and Wickens 2008; Kurzik-Dumke and Zengerle 1996). Yet, Angel2 was not observed to possess deadenylase activity in biochemical assays, and instead both Angel1 and Angel2 were found to hydrolyze 2',3' cyclic phosphates, (Pinto et al. 2020).

In this work, we identify the 2',3' cyclic phosphatase Angel1 as a rate-limiting factor in RQC-mediated mRNA decay in human cells. Angel1 associates with proteins known to be involved in RQC and is enriched in mRNA coding regions, including near sequences that are associated with ribosome stalling. Depletion of Angel1 and the human RQC endonuclease homolog N4BP2 causes stabilization of RQC reporter mRNAs targeted for NSD, and the catalytic activity of Angel1 is critical for this activity. These observations implicate Angel1 in human RQC-mediated mRNA decay and suggest that a rate-limiting step of the pathway is removal of a 2',3' cyclic phosphate.

3.3 Results

3.3.1 Angel1 associates with components of the RQC pathway

To gain insight into possible functions for Angel1, we first established an assay to identify Angel1 protein binding partners. Using the Flp-In T-REx system, we constructed stable human embryonic kidney (HEK) 293 cell lines expressing N-terminally FLAG-tagged Angel1 under the control of a tetracycline-regulated promoter

that we titrated to express Angel1 at close to endogenous levels (Supplementary Figure 3.1A). We performed immunoprecipitation (IP) against the FLAG-tag (Supplementary Figure 3.1B) and identified associated proteins by liquid chromatography followed by tandem mass spectrometry (LC-MS/MS). IPs were performed with or without prior RNase A treatment to help distinguish between RNA- and protein-mediated interactions. To identify interactions specific to Angel1, the IPs were compared to IPs from a parental Flp-In T-REx cell line expressing no FLAG-tagged fusion protein, and a cell line expressing FLAG-tagged TOE1, a better understood DEDD-type deadenylase with a role in snRNA processing (Lardelli et al. 2017) (Figure 3.1A).

Among the most abundant proteins that specifically co-purified with FLAG-tagged Angel1 (Supplementary Table 3.1), was the mRNA cap-binding protein eIF4E, which, importantly, reproduces a previously described interaction (Gosselin et al. 2013). Other proteins that specifically co-purified with Angel1 included additional mRNP components (LARP4, LARP4B, DDX6, LSM14A, ATXN2, and PABPC), all components of the GATOR2 complex (MIOS, WDR24, WDR59, and SEH1L) involved in activation of mTORC1 (Cai et al. 2016), and components of a complex important for cytoskeletal functions of neurons and synaptic plasticity, DISC1, NDE1 and NDEL1 (Tropea et al. 2018).

A striking subset of Angel1-associated proteins were components of the RQC pathway (Figure 3.1A). These included RACK1, LTN1, ubiquitin, and all three components of the SKI complex, a cytoplasmic adapter and RNA helicase for the RNA exosome. With the exception of RACK1, these all associated with Angel1 in a

manner resistant to RNase A treatment (Figure 3.1A), suggesting protein-mediated interactions that are independent of RNA. We confirmed the association of Angel1 with eIF4E and the SKIV2L subunit of the SKI complex by IP followed by Western blotting (Figure 3.1B). Given the homology of Angel1 with 3' RNA processing factors and, in particular, its association with components of the RQC pathway, we explored the hypothesis that Angel1 is involved in quality control of mRNAs with stalled ribosomes.

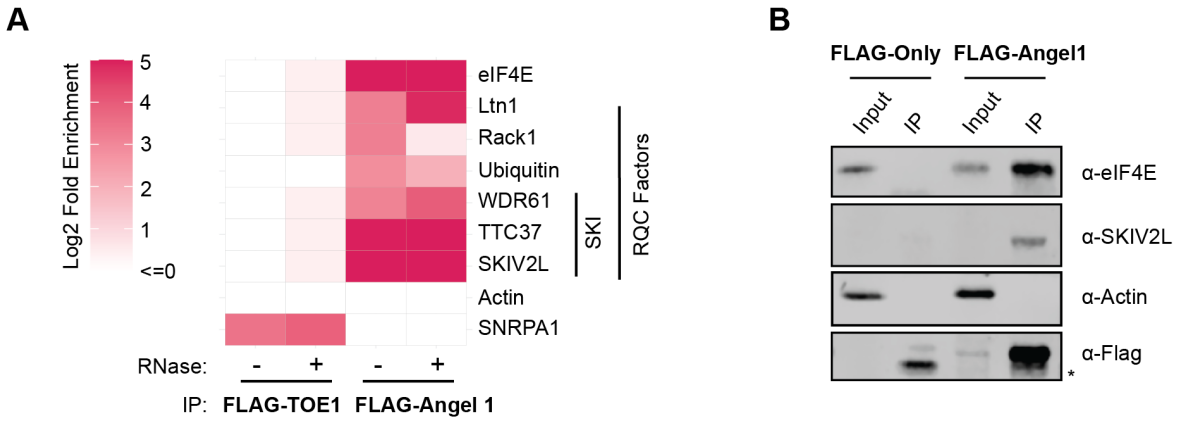


Figure 3.1: Angel1 associates with components of the ribosome-associated quality control pathway. (A) Select proteins enriched in IP-MS/MS for FLAG-tagged Angel1 or TOE1 over an IP performed with a cell line expressing no FLAG-tagged protein. IPs were performed in the absence (-) or presence (+) of RNase A. Fold enrichment was calculated as number of peptides per 10,000 total observed in the test IP over the negative control IP, after adding a pseudocount of 1 to each. For all peptide counts, see Supplementary Table S1. (B) Co-IP assays followed by Western blotting monitoring specific proteins associated with Angel1. Actin served as a negative control. Input: 10% of the total cell extract used for IP. *: non-specific band.

3.3.2 Angel1 associates with mRNA coding regions and sequence features correlated with stalled ribosomes

Pursuing this hypothesis, we were interested in understanding what RNA sequences and motifs Angel1 interacts with, reasoning that it may show preference for regions of transcripts associated with stalled ribosomes. To that end, we performed enhanced cross-linking and immunoprecipitation followed by sequencing (eCLIP-Seq) (Van Nostrand et al. 2016). Two replicates of FLAG-Angel1 eCLIP-Seq were performed from our Flp-In T-REx HEK293 cell lines expressing Angel1 close to endogenous levels. We also performed eCLIP-Seq with the parental cell line as a comparison control for background noise in the assay. We found high agreement between the two eCLIP replicates (Figure 3.2A) with a higher correlation than with the negative control (Supplementary Figures 3.2A and 3.2B). Consistent with a factor that may monitor translation, Angel1 CLIP reads were strongly enriched for coding regions of mRNAs (Figure 3.2B). Using an eCLIP-Seq analysis pipeline (Van Nostrand et al. 2016), we identified CLIP peaks for each replicate ($p<0.05$). Limiting the analysis to CLIP peaks that were reproducible between the two replicates (cutoff threshold $p < 0.001$) showed a further increase in the percentage of peaks mapping to coding regions.

Typically, nucleotide motifs of RNA binding proteins are identified by applying motif finding algorithms to the sequences of reproducible peaks. However, this analysis failed to produce strong nucleotide or codon motifs in Angel1-associated peaks. We reasoned that, if involved in RQC, Angel1 might be recruited to regions upstream or downstream of ribosome stalls and therefore examined regions within 50 nucleotides of each peak. Motif analysis (Bailey et al. 2009) of these Angel1-associated regions revealed an abundance of guanosine-rich sequences

(Supplementary Figure 3.2C) and several amino acid-coding motifs that have been associated with stalled ribosomes (Chyzyńska et al. 2021), including poly-Glycine, poly-Glutamate/Aspartate, and, less enriched, poly-Proline codons (Figure 3.2C). We also examined nucleotide content of regions around peaks and found that they contained on average higher GC content than regions surrounding reads in the control (Figure 3.2D) and were more likely to contain what are predicted to be stronger structures (Supplementary 3.2D). Using a G-quadruplex prediction algorithm (Kikin et al. 2006), we found that these regions were also predicted to more likely form G-quadruplexes (Figure 3.2D), a secondary structure element that has also been associated with ribosome stalling (Endoh and Sugimoto 2016). These associations are consistent with a factor involved in RQC and the wide variety of nucleotide and nascent oligopeptide sequences that may induce ribosome stalling.

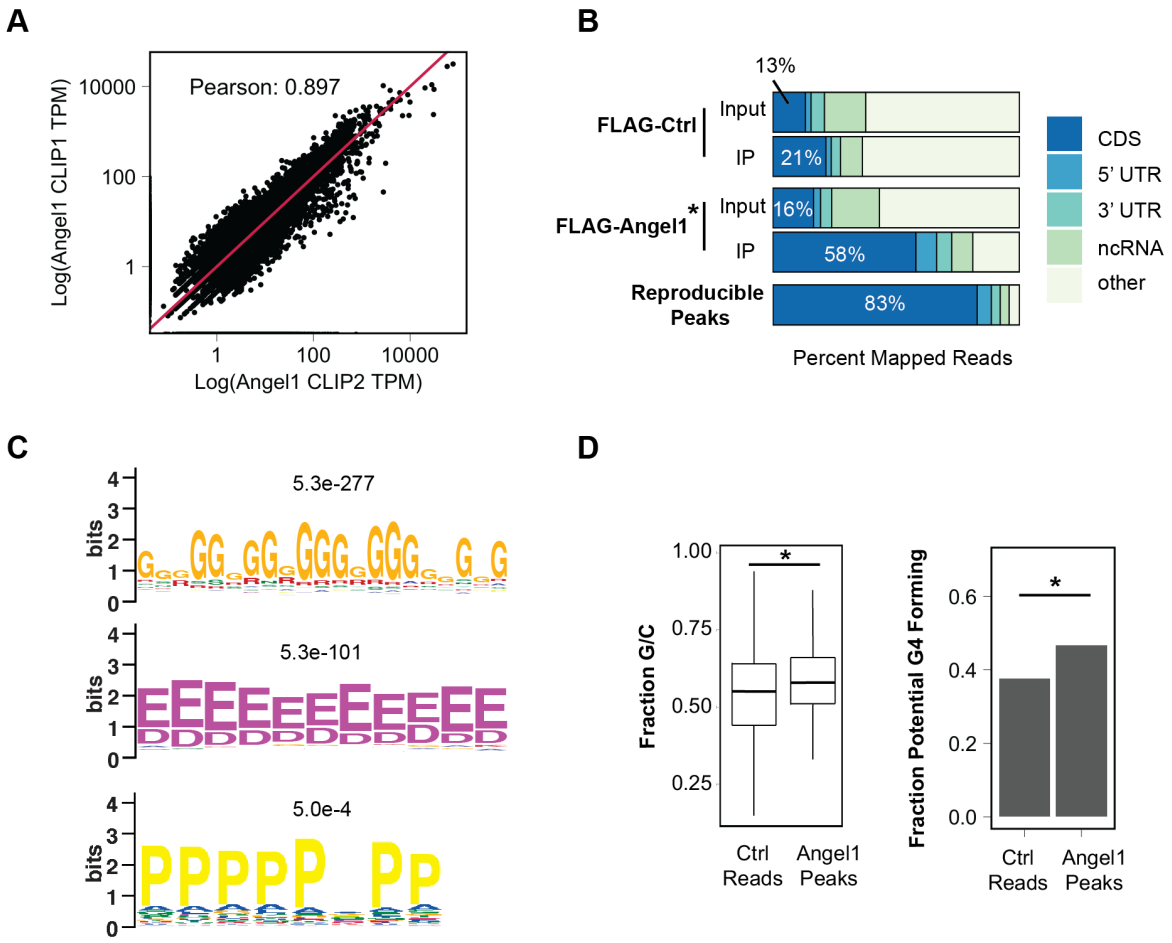


Figure 3.2: Angel1 associates with coding regions of mRNAs and with sequences associated with ribosomal stalling. A) Calculated transcripts per million (TPM) compared between the two FLAG-tagged Angel1 eCLIP replicates. Calculated Pearson correlation between the replicates is shown. (B) Fraction of reads mapping to different functional regions of RNAs in control versus FLAG-tagged Angel1 input and eCLIP (IP) samples, and in reproducible peaks ($p<0.001$) between the two FLAG-tagged Angel1 eCLIP experiments. (C) Peptide motifs enriched in areas within 50 nucleotides upstream or downstream of identified peaks as compared to areas around reads from the control sample. E-values are listed. The top two shown motifs were the most highly enriched motifs in the MEME analysis. (D) GC content (left) and calculated guanosine quadruplex (G4) formation capacity (right) of sequences between 50 nucleotides upstream or downstream of identified Angel1 eCLIP peaks as compared to areas around reads from the control sample. *: $p<2.2e-22$ (KS-test).

3.3.3 N4BP2 and other human homologs of budding yeast RQC factors contribute to NSD

While many factors involved in RQC-mediated mRNA degradation have been identified in budding yeast (D'Orazio et al. 2019; Van Hoof et al. 2002; Frischmeyer et al. 2002; Tsuboi et al. 2012) and in C. elegans (Glover et al. 2020), much less is known about factors in human cells (Saito et al. 2013). To establish an assay to monitor mRNA degradation mediated by RQC, we adapted the well-characterized β-globin mRNA pulse-chase system (Lykke-Andersen et al. 2000) by creating an NSD reporter mRNA with no stop codons. In this system, wild-type β-globin mRNA is highly stable with a half-life of over 600 minutes (Durand and Lykke-Andersen 2013). Removal of all stop codons created an unstable mRNA (BG-NSD) that degraded at a rate faster than the well-characterized β-globin Nonsense-Mediated Decay (NMD) reporter mRNA containing a premature termination codon at position 39 (BG-NMD) (Durand et al. 2016) (Figures 3.3A and 3.3B). A single point mutation reintroducing a stop codon a few nucleotides upstream of the cleavage and polyadenylation site of the BG-NSD reporter (BG-NSD+Stop) rescued stability (Figures 3.3A and 3.3B). These substrates provided a platform for investigating the effects of perturbations of human RQC machinery.

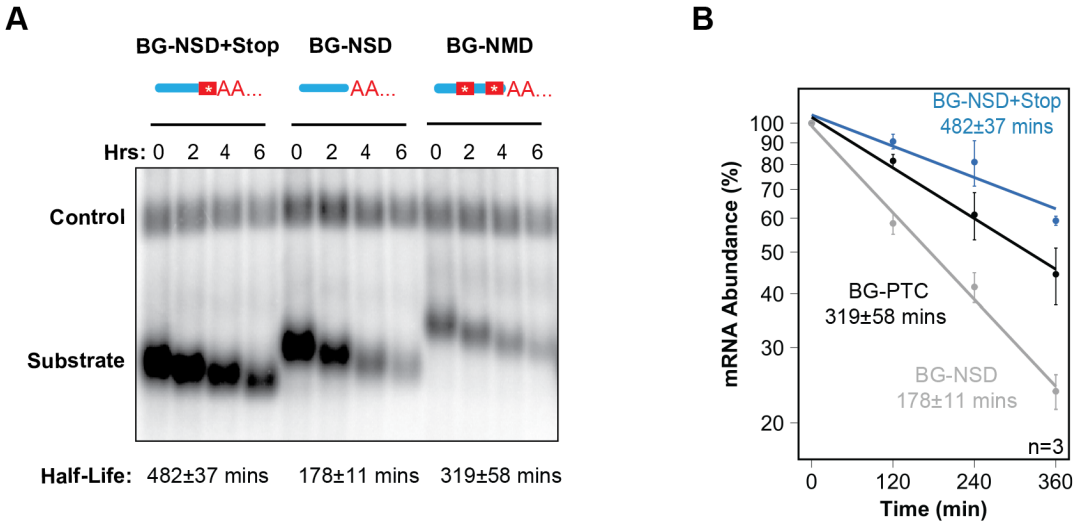


Figure 3.3: Establishment of a human NSD reporter mRNA assay. (A) Representative Northern blot of a pulse-chase mRNA decay assay in HeLa Tet-off cells monitoring degradation of BG-NSD+STOP, BG-NSD and BG-NMD mRNAs (Substrate) as compared with constitutively expressed β-globin-GAP3 control mRNA (Control). Numbers above lanes refer to hours after transcription shut-off of the substrate mRNAs by tetracycline. Bands were quantified using a phosphorimager and normalized to the constitutively expressed β-globin-GAP3 mRNA to calculate mRNA half-lives assuming first-order kinetics, which are given below the blot with standard error of the mean from three experiments. (B) Exponential decay plots of the experiment in panel A performed in triplicate (n=3). Error bars represent standard error of the mean.

Using this system, we examined changes in the stability of the NSD reporter mRNA when depleting human homologs of RQC factors identified in budding yeast (Supplementary Figures 3.3A and 3.3B). Depletion of SKIV2L, a component of the human SKI complex, led to stabilization of the NSD reporter (Figure 3.4A), consistent with a previous report (Saito et al. 2013). Similarly, depletion of human RACK1 caused stabilization of the NSD reporter, although the level of stabilization did not reach statistical significance.

Having thus confirmed that our NSD reporter responds to known RQC factors, we next tested whether N4BP2, a mammalian homolog of the initiating RQC endonuclease Cue2/NONU-1, is rate-limiting for degradation of the NSD substrate. Indeed, depletion of N4BP2 led to stabilization of the NSD reporter (Figure 3.4B). In contrast to the other tested RQC factors, depletion of XRN1, using knock-down conditions that stabilizes a cleavage intermediate in the NMD pathway (Franks et al. 2010), did not stabilize the β-globin NSD substrate (Figure 3.4C). These observations demonstrate that the NSD reporter mRNA is indeed a target of degradation by RQC machinery, and that N4BP2 is a human ortholog of the *S. cerevisiae* and *C. elegans* RQC endonuclease.

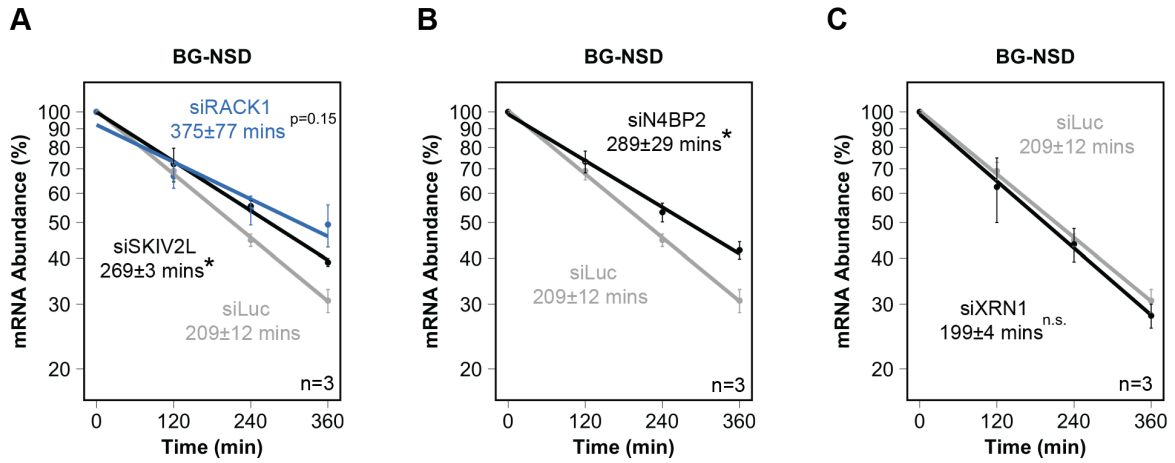


Figure 3.4: Depletion of human homologs of yeast RQC factors stabilize the NSD reporter. (A) Exponential decay plots of the BG-NSD substrate after depletion of known RQC factors, SKIV2L and RACK1. (B) Exponential decay plots of the BG-NSD substrate after depletion of N4BP2. (C) Exponential decay plots of the BG-NSD substrate after depletion of XRN1. Error bars represent standard error of the mean (n=3). *: $p<0.05$, calculated by one-tailed Student's t-test compared to the control knockdown targeting luciferase (siLuc).

3.3.4 Angel1 and its catalytic residues are limiting for NSD

We next tested whether Angel1 contributes to human NSD. Indeed, siRNA-mediated depletion of Angel1 (Supplementary Figure 3.4A) resulted in stabilization of the β-globin NSD reporter mRNA (BG-NSD) (Figure 3.5A). This effect was observed for two independent siRNAs targeting Angel1 (Supplementary Figure 3.4B). By contrast, depletion of Angel1 did not alter the stability of the β-globin NMD reporter mRNA (BG-NMD) (Figure 3.5B), showing that the effect is specific to turnover of the NSD substrate and not due to general repression of mRNA turnover or translation. Depletion of Angel1 also caused stabilization of an NSD reporter based on TPI mRNA (Supplementary Figure 3.4C), showing that the effect of Angel1 is not specific to a β-globin mRNA substrate.

Finally, to test whether Angel1 catalytic activity is important for its function in RQC, we generated an Angel1 protein containing a glutamate to alanine substitution previously shown to disrupt 2',3' cyclic phosphatase activity of the Angel1 homolog Angel2 (Pinto et al. 2020). Exogenous expression (Supplementary Figure 3.4D) of wild-type Angel1 partially rescued the effect of Angel1 depletion on NSD reporter stability (Figure 3.5C). By contrast, no rescue of activity was observed upon expression of the catalytically inactive Angel1 mutant (Angel1 EA), demonstrating that the activity of Angel1 in NSD depends on its catalytic residues. Thus, the 2',3' cyclic phosphatase Angel1 and its catalytic activity is a rate-limiting component of RQC-mediated mRNA degradation in human cells.

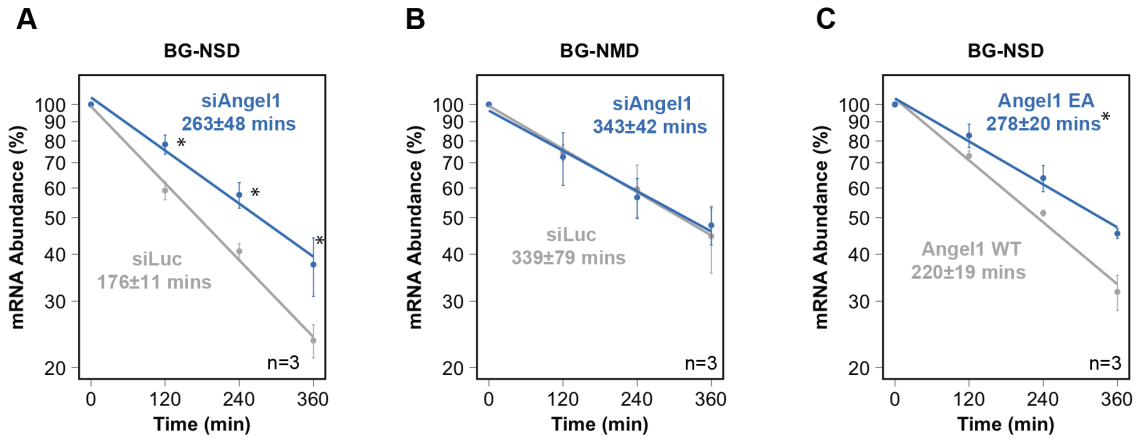


Figure 3.5: Angel1 and its catalytic activity is rate-limiting for the degradation of an NSD target mRNA. (A) Exponential decay plots of the BG-NSD substrate after depletion of Angel1 or using a control siRNA (siLuc). (B) Exponential decay plots of a substrate containing a premature termination codon that is targeted for Nonsense-Mediated Decay (BG-NMD) after depletion of Angel1 or using a control siRNA. (C) Exponential decay plots of the BG-NSD substrate after depletion of Angel1, and complementing with exogenous Angel1 WT or catalytic inactive Angel1 EA. *: $p<0.05$ calculated by a one-tailed Student's t-test. Error bars represent standard error of the mean (n=3).

3.4 Discussion

The mechanism by which mRNAs subjected to RQC in human cells are targeted for degradation remains poorly understood. In this work, we present evidence that the 2',3' cyclic phosphatase Angel1 facilitates decay of mRNAs targeted for RQC. Indeed, Angel1 associates with proteins involved in RQC (Figure 3.1) and with coding regions of mRNAs, including RNA sequences that have been associated with translational stalling (Figure 3.2). Depletion of Angel1, as well as of N4BP2, a human homolog of the RQC endonuclease, stabilizes NSD reporter mRNAs, and the catalytic activity of Angel1 is critical for this activity (Figures 3.4 and 3.5). Thus, Angel1, and its catalytic activity, contributes to RQC-mediated mRNA

decay in human cells, suggesting that hydrolysis of a 2',3' cyclic phosphate is a rate-limiting step in the pathway.

3.4.1 By what mechanism does Angel1 facilitate RQC?

Angel1 is a homolog of EEP-type CCR4 deadenylases, but no deadenylase activity has been observed for human Angel proteins in biochemical assays (Supplementary Figure 3.5) (Pinto et al. 2020). Instead, a recent study found that Angel1 and its homolog Angel2 have activities as 2',3' cyclic phosphatases, dependent on a highly conserved catalytic glutamate residue (Pinto et al. 2020). Our observation that Angel1's function in RQC is dependent on the same glutamate residue suggests that Angel1 functions as a cyclic phosphatase in the pathway. Evidence has shown that the initial cleavage of the RNA substrate during RQC occurs independently of divalent metal ions and leaves a 5'OH on the 3' RNA fragment (Navickas et al. 2020). This suggests that this cleavage generates a 2',3' cyclic phosphate on the 5' RNA fragment, consistent with the general biochemical activity of divalent metal ion-independent RNases with homology to N4BP2 (Yang 2011). Thus, a likely function for Angel1 in RQC is the hydrolysis of this cyclic phosphate, which may have to be resolved before the RNA can be degraded in the 3'–5' direction. Another potential substrate for a cyclic phosphatase in the RQC pathway could be the P-site tRNA, which, after removal and degradation of the nascent polypeptide, is left with a 2',3' cyclic phosphate (Yip et al. 2019), although resolution of this phosphate seems a less likely explanation for the impact of Angel1

on mRNA degradation. While its demonstrated cyclic phosphatase activity is the most parsimonious explanation for the function of Angel1 in RQC, it cannot be ruled out that Angel1 functions by a different mechanism, such as by acting as a deadenylase. However, such an activity has not been observed for Angel proteins *in vitro*. If the function of Angel1 in RQC indeed is as a cyclic phosphatase, this would be the first known example of a cyclic phosphatase involved in a mRNA decay process. Such a process could be wide-spread. For example, U6 snRNA is known to be stabilized by a terminal 2',3' cyclic phosphate (Gu et al. 1997) and a subset of RNA endonucleases generate 2',3' cyclic phosphates that may have to be resolved prior to exonucleolytic degradation of RNA fragments.

3.4.2 Possible functions for Angel1 outside of RQC

In addition to RQC factors, we also found association in our IP-MS/MS analysis of Angel1 with the Gator2 complex which, along with Sestrins and Gator1, is important for sensing amino acid deprivation and signaling through mTORC1 (Bar-Peled et al. 2013; Kowalsky et al. 2020). The association of Angel1 with these components suggests a potential role in sensing or modulating amino acid deprivation. Given Angel1's biochemical activity, such a function could be related to tRNAs, which can be cleaved during tRNA splicing and stress conditions to create 2',3' cyclic phosphate-containing species (Zillmann et al. 1991; Shigematsu and Kirino 2020). Furthermore, Angel1's association with DISC1-NDE1/NDEL1 suggests a function for Angel1 in neurons and neuronal plasticity, although these proteins have

no currently known role in RNA metabolism. Angel1 could also be involved in additional processes that involve cyclic phosphates, such as the metabolism of RNAs that feature cyclic phosphates during their life-cycles, such as U6 snRNA (Gu et al. 1997), spliced tRNAs (Zillmann et al. 1991), or XBP1 mRNA (Jurkin et al. 2014).

3.4.3 What are the endogenous substrates of Angel1 and the human RQC pathway?

While our reporter assays show that Angel1 is rate-limiting for decay of engineered human NSD substrates (Figure 3.5), our eCLIP experiments suggest that Angel1 associates broadly with protein coding regions of mRNAs (Figure 3.2A). Indeed, depletion of RQC factors such as ZNF598 have shown broad, low-level, effects on the transcriptome (Tuck et al. 2020; Kalisiak et al. 2017; Weber et al. 2020). Identification of endogenous substrates of RQC has been generally unsuccessful with only a few potential examples, including the ER stress-induced XBP1, which is upregulated at the protein level upon depletion of ZNF598 (Han et al. 2020). These observations suggest broad pleiotropic effects of perturbations in this system, perhaps reflecting a process that occurs stochastically at individual mRNAs in normal conditions.

3.4.4 Potential differences between the degradation of RQC mRNA substrates in budding yeast and humans

Information on RQC-mediated mRNA decay in mammalian cells has so far been limited. A previous study identified a component of the SKI complex and ribosome recycling factors HBS1 and DOM34 as limiting for degradation of an NSD reporter mRNA (Saito et al. 2013). Here, in addition to Angel1, we identified N4BP2 as a component of the pathway, a homolog of the recently identified RQC endonuclease Cue2/NONU-1. In a potential departure from the yeast pathway, we observed no effect on degradation of our NSD reporter mRNA upon depletion of XRN1, despite using conditions that stabilized a cleavage intermediate of the NMD pathway (Franks et al. 2010). That XRN1 was not rate-limiting for degradation of the β-globin NSD substrate is perhaps not unexpected given the predicted position of the ribosome stall and endonucleolytic cleavage at the extreme 3' end of the tested substrate. However, this contrasts the predominance of Xrn1-mediated degradation over a cleavage-initiated pathway observed for NGD mRNA substrates in budding yeast (D'Orazio et al. 2019). This may indicate that an endonucleolytic cleavage-mediated pathway is more predominant in mammalian cells, at least for the NSD substrate tested here, which is consistent with our observation that N4BP2 is rate-limiting for its degradation. Altogether, our study identifies Angel1 and N4BP2 as factors involved in human RQC-mediated mRNA decay, suggesting a rate-limiting role in RQC mRNA degradation for the hydrolysis of a 2',3' cyclic phosphate, perhaps one generated upon the initial cleavage of the mRNA by N4BP2.

3.5 Materials and Methods

Table 3.1: Oligo sequences

Name	Sequence
siNEDD4BP2	GGGUUGAAUUUGAAAGAAAUU
siRack1	UCUGGGAUUUAGAGGGAAAUU
siSki	GCCUUAGCUGUAUGUUGGAUU
siXRN1	AGAUGAACUUACCGUAGAAUU
siAngel1 #1 (3'UTR)	CCACAGACCUGGUGUAAUUUU
siAngel1 #2 (CDS)	AUGGAAGGAGUGACAGAUAUU
siLuciferase	CGUACGCGGAAUACUUCGAUU
N4BP2_qPCR_Fwd	CCCGCTCCAGAAGCAGTAAG
N4BP2_qPCR_Rev	CGATGCTGTCAACATGCTCATT
Rack1_qPCR_F	CTTCTGGAGGCAAGGATGGC
Rack1_qPCR_R	CACACAGCCAGTAGCGGTTA
PELO_qPCR_F	GGATGTGAAAAGCGTGGAGC
PELO_qPCR_R	TTTCTCTTCCATCCCTGTCAGT
SKIV2L_qPCR_F	CGGGAGCGAATGCAGATACA
SKIV2L_qPCR_R	GTTCGGAGCACCTCTACTCG
DNA Loading Control	TCTCTCTCTCTCTAAATAAAAAAAAAAAAAAAAAA
GAPDH_qPCR_F	CATGGCCTTCCGTGTTCCTA
GAPDH_qPCR_R	CCTGCTTCACCACCTTCTTGA
PolyA Substrate	UCUAAAUAAAAAAAAAAAAAAAAAAA

3.5.1 Antibodies

Western blotting was performed with anti-FLAG (Sigma F7425; 1:1,000), anti-eIF4E (Cell Signaling Technologies 9742; 1:1,000), anti-SKIV2L (Thermo Fisher 11462-1-AP; 1:500), anti-β-actin (Cell Signaling Technologies 4967; 1:1,000).

3.5.2 Plasmids

Gibson assembly (New England Biolabs) was used to insert the coding region of Angel1 with an engineered N-terminal FLAG-tag into pcDNA5/FRT/TO and pcDNA3 to create pcDNA5/FRT/TO-FLAG-Angel1 WT and pcDNA3-FLAG-Angel1 WT. Site-directed mutagenesis (New England Biolabs, E0554S) was used to create an E298A catalytically inactivating mutation, generating pcDNA3-FLAG-Angel1 EA. Pulse-chase constructs were created from the previously described plasmid, pcTet2-BWT (Damgaard and Lykke-Andersen 2011). pcTet2-BG-NSD was created by 3 rounds of site-directed mutagenesis that removed all in-frame stop codons before the cleavage and polyadenylation site through deletion of a 70 nucleotide region and 6 point mutations. A point mutation was introduced 10 nucleotides upstream of the cleavage and polyadenylation site to create a stop codon, generating pcTet2-BG-NSD+stop. pcTet2-BG-NMD was created by site-directed mutagenesis introducing a stop codon at codon 39 of pcTet2-BWT. pcBGAP3 was used as a loading control for pulse-chase experiments (Clement and Lykke-Andersen 2008). pcTet2-TPI-NSD was generated from pcTet2-TPI (Singh et al. 2008) using a synthesized double stranded gene fragment (IDT, gBlock) of the 3'UTR of pcTet2-TPI edited to remove stop codons. The gene fragment was used to replace the 3'UTR of pcTet2-TPI by Gibson assembly. Plasmid sequences are available upon request.

3.5.3 Stable cell line construction and titration of FLAG-Angel1 levels

pcDNA5/FRT/TO-FLAG-Angel1 WT was used to generate stable HEK FLp-In T-REx-293 cell lines (Invitrogen) according to the manufacturer's protocol, in which FLAG-Angel1 expression can be titrated with tetracycline. In the absence of an Angel1 antibody, we used an anti-FLAG antibody to estimate FLAG-Angel1 expression levels in comparison to FLAG-TOE1, which had been titrated to endogenous levels as monitored by a TOE1 antibody (Wagner et al. 2007; Lardelli et al. 2017). TOE1 is approximately 25 times more abundant than Angel1 in HeLa cells according to global mass spectrometry measurements (Nagaraj et al. 2011). We therefore titrated FLAG-Angel1 expression with tetracycline to match a level of approximately 1:25 relative to TOE1, which was reached at 5 ng/ml of tetracycline. This concentration of tetracycline was used in all experiments when expressing FLAG-Angel1 in the stable HEK Flp-In T-REx-293 line.

.

3.5.4 Cell growth and depletions

Cells were maintained in Dulbecco's Modified Eagle Medium (DMEM, Gibco, 11965092) with 10% fetal bovine serum (Gibco, 10437028). Flp-In TREx lines were induced with 5 ng/ml tetracycline 24 hours before harvest. Knockdowns were performed using 20 nM of small interfering RNAs (siRNAs) custom ordered from Horizon Discovery (Table 3.1). The control siRNA targeted luciferase mRNA. Knockdowns were performed with siLentFect reagent (Bio-Rad, 703362) according to the manufacturer's specifications.

3.5.5 Pulse chase mRNA decay assays

Hela Tet-off cells were plated at $15x10^4$ cells per well in a 6-well plate. siRNA-mediated knockdowns were performed at 72 hours and 24 hours prior to cell harvest. In addition, 48 hours prior to cell harvest, cells were transfected with 0.5 µg of the test construct (pcTet2-BG-NSD, -BG-NSD+Stop, or -BG-NMD), 0.5 µg of pcDNA3-based Angel1 addback construct (if applicable), 0.1 µg of a pcBGAP3 loading control plasmid, and empty pcDNA3 plasmid stuffer to a total of 2 µg. Cells were maintained with 50 ng/ml tetracycline to prevent expression from the test plasmid. 72 hours after the initial siRNA transfection, cells were rinsed with PBS, and transcription from the test plasmids was pulsed by addition of 2 ml of fresh medium free of tetracycline for 6 hours. Medium was subsequently replaced with DMEM/10% FBS containing 1000 ng/ml tetracycline to shut off test plasmid transcription and cells were collected every 2 hours thereafter in Trizol reagent (Thermo Fisher, 15596026). RNA was isolated and substrate levels were analyzed by Northern blotting as previously described (Clement and Lykke-Andersen 2008).

3.5.6 Immunoprecipitation assays

Flp-In Trex lines expressing FLAG-tagged Angel1, FLAG-tagged TOE1, or no FLAG-tagged fusion protein were grown to approximately 50% confluency and induced with 5 ng/ml tetracycline for 24 hours. Cells were harvested by scraping into ice cold PBS and flash frozen in liquid nitrogen. Pellets were resuspended in isotonic

lysis buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.2 mM EDTA, 0.5% Triton X-100 with 80 units/ml RNaseOUT (Thermo Fisher, 10777019) or 125 µg/ml Rnase A (Sigma, R4875), and 1 tablet/10 ml of protease inhibitor (Thermo Fisher, 88666)) for 10 minutes on ice. Lysates were spun down at 20,000 g for 15 minutes at 4˚C. FLAG peptide (ApexBio, A6002) was added to lysates to a concentration of 1 µg/ml to reduce non-specific interactions. Samples were incubated with pre-washed anti-FLAG M2 agarose beads (Sigma, A2220) for 2 hours at 4˚C with rotation. Beads were washed 9 times with NET2 buffer (10 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.1% Triton X-100). Protein was eluted by treating beads three times for 30 minutes at 4˚C with NET2 containing 200 µg/ml FLAG peptide and elutions were subsequently pooled. Samples from input, the unbound fraction, and elution were separated by gel electrophoresis and visualized by silver staining (Thermo Fisher, 24580) according to the manufacturer's protocol. Protein amounts for deadenylation assays were estimated against BSA standards (New England Biolabs, B9000S).

3.5.7 LC-MS/MS and analysis

LC-MS/MS was performed as previously described (Sundaramoorthy et al. 2017). To calculate the fold enrichment of individual proteins in the Angel1 IP over the matched FLAG control, the number of peptides for each protein were normalized to counts per 10,000 in the total count for each sample, and the normalized counts for Angel1 IP were divided by normalized counts for the control after adding a

pseudocount of 1 to every normalized peptide count to prevent division by zero errors.

3.5.8 eCLIP

Flp-In Trex lines expressing FLAG-tagged Angel1 or no FLAG-tagged fusion protein were grown to approximately 50% confluency and induced with 5 ng/ml tetracycline for 24 hours. Cells were crosslinked to preserve protein-RNA interactions by treatment with UV (Stratalinker, 254nm, 400 mJ/cm2, on ice). One sample was not exposed to UV as a no-crosslink control. eCLIP library preparation was performed as previously detailed (Van Nostrand et al. 2017). Samples were mapped to the hg37 human genome and features from the Gencode 19 annotation were counted with featureCounts (Liao et al. 2014). Areas 50 nucleotides upstream and downstream from peaks were extracted with custom python scripts that used transcripts tagged as Appris principal (Rodriguez et al. 2013) to limit genes to one transcript. In cases where genes had multiple principal transcripts, the longest transcript was selected. G/C content was calculated with custom scripts for those regions. G quadruplex formation potential was measured for those sequences using QGRS (Kikin et al. 2006). Significance was tested with a Kalmagorov-Smirnov (KS) test.

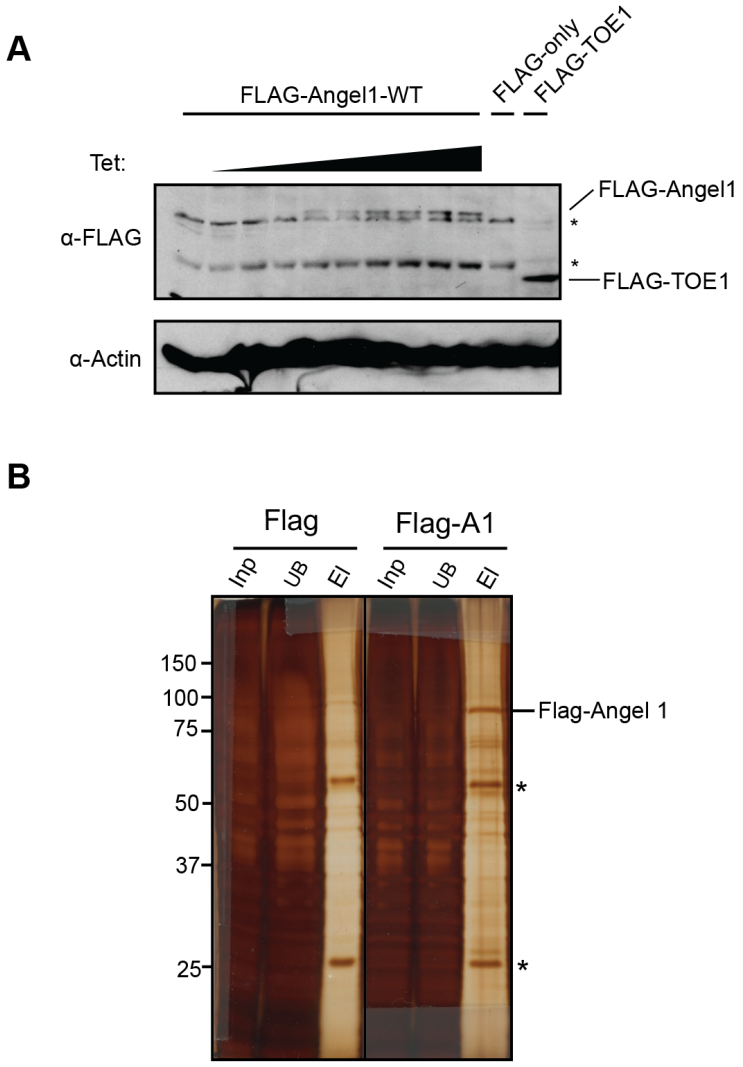
3.5.9 RT-qPCR

After total RNA isolation from cells, reverse transcription was performed using Superscript III (Thermo Fisher, 18080044) according to manufacturer's protocol. qPCR was performed with a master mix (Thermo Fisher, 4385612) and using a Quantstudio 3 machine according to manufacturer specifications. Angel1 qPCR was carried out using pre-validated primers (Bio-Rad, 10025636). All other qPCR primer pairs are listed in Table 3.1.

3.5.10 Deadenylation assay

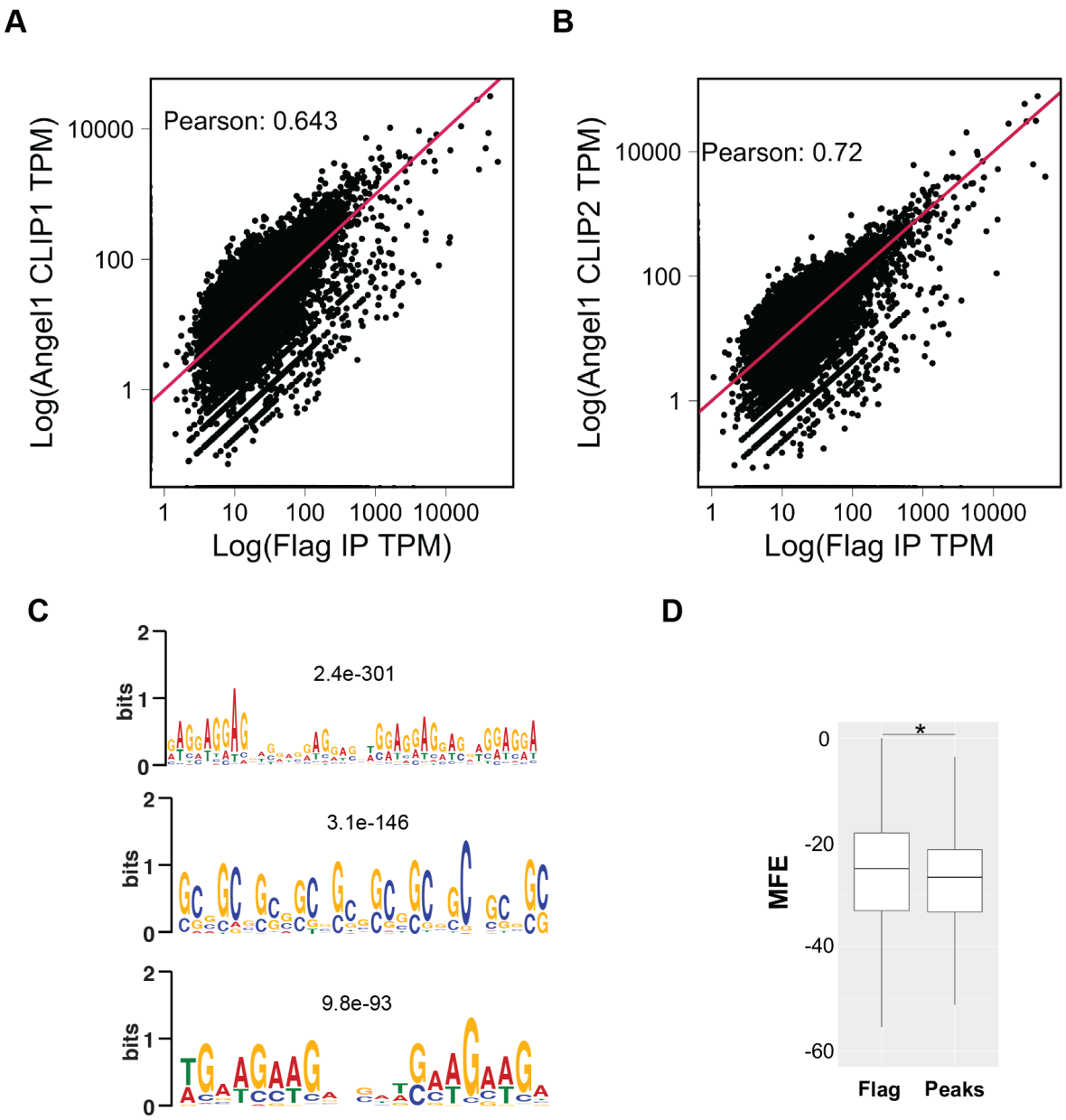
A custom poly-A RNA substrate terminating in 20 adenosines (Dharmacon) and a DNA loading control also terminating in 20 adenosines (IDT) (Table 3.1) were 5' labelled with [γ-^{32}P]-ATP (PerkinElmer) using T4 poly-nucleotide kinase (NEB) according to manufacturer protocol. The deadenylation assay was adapted from a previously described protocol (Wagner et al. 2007). Approximately 50 nM of the indicated protein was added to approximately 5,000 CPM each of DNA loading control and RNA substrate and incubated at 37°C in deadenylation buffer (20mM HEPES, pH7.4, 2mM $MgCl_2$, 0.1 mg/ml bovine serum albumin, 1mM spermidine, 0.1% Igepal CA-630 (Sigma), 0.5 units/µl Rnase-Out, and 0.5 µg/µl yeast total RNA). Formamide loading buffer was added to stop the reaction and samples were loaded on a 9% polyacrylamide-6M urea denaturing gel. The gel was dried and imaged using a phosphorimager.

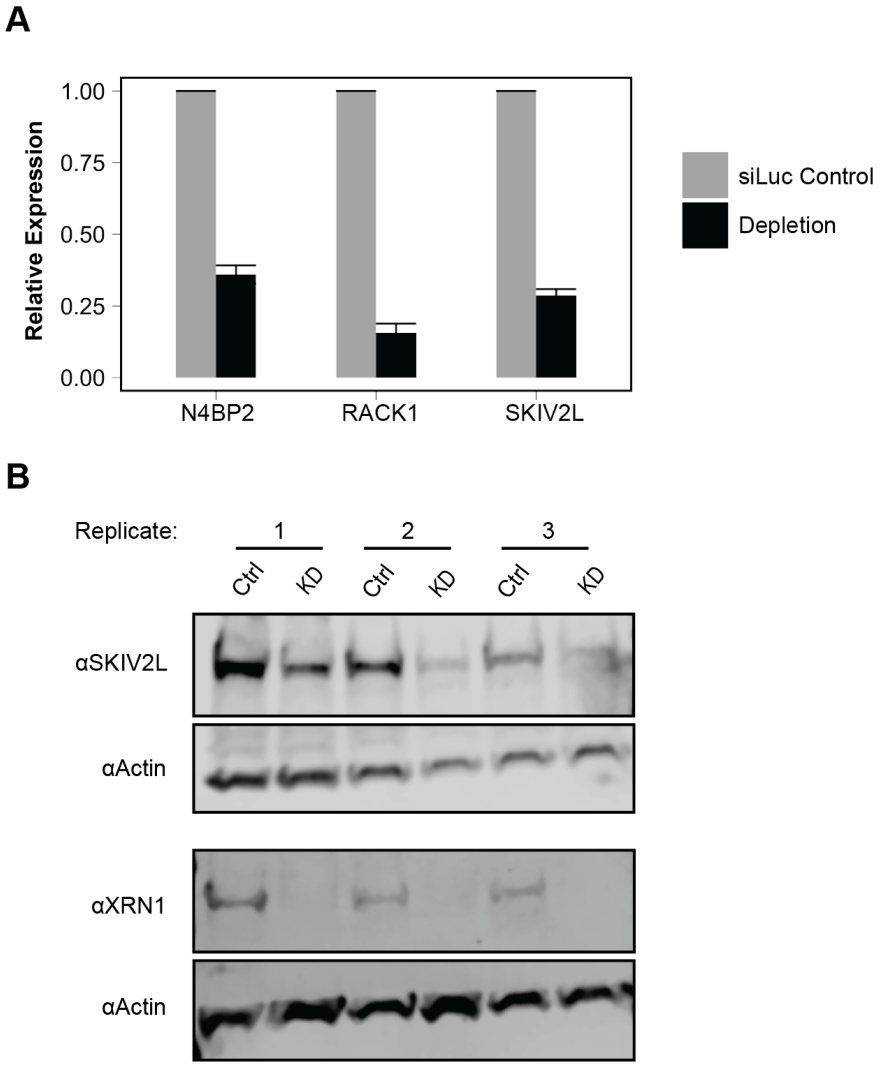
3.6 Supplemental figures



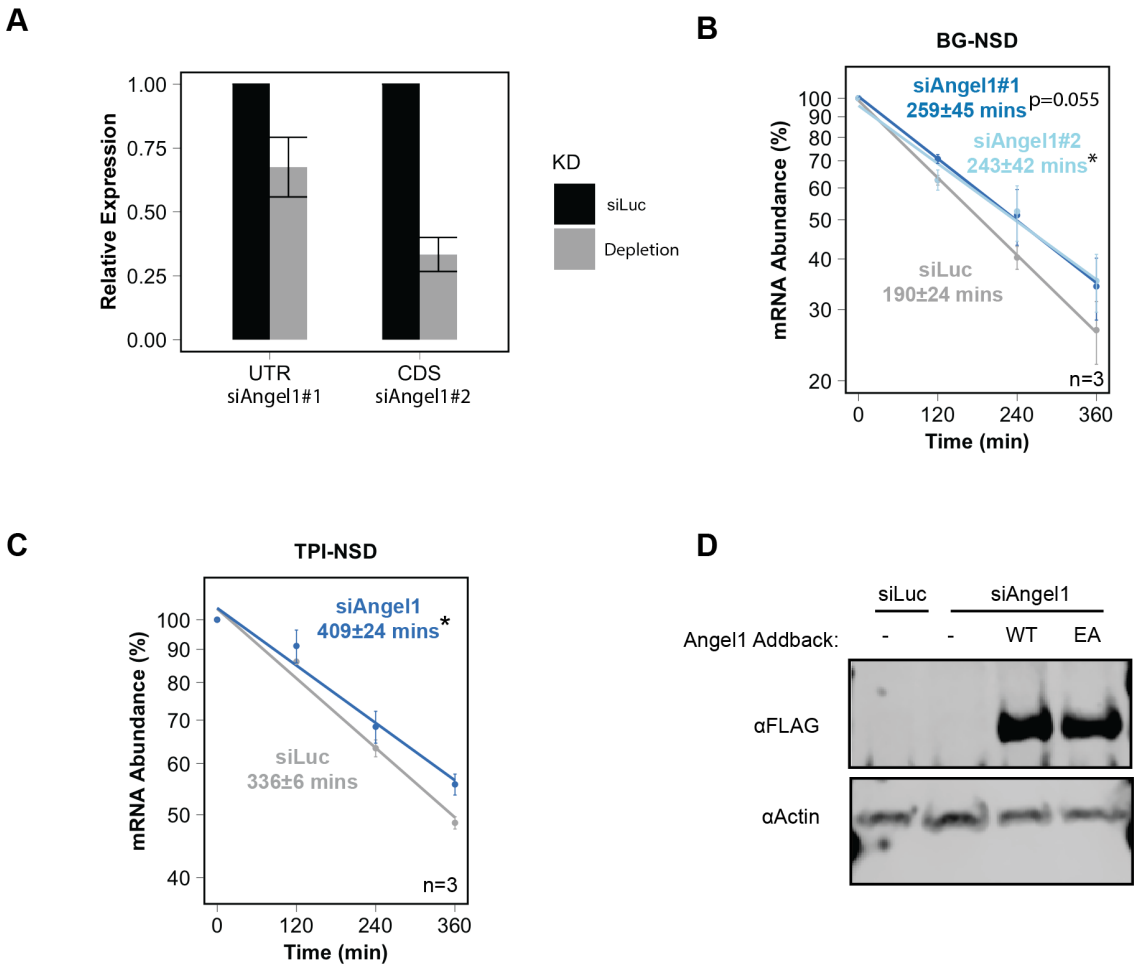
Supplementary Figure 3.1: FLAG-Angel1 HEK293 FlpIn T-REx cell line validation. (A) Western blot showing expression of FLAG-Angel1-WT with titration of tetracycline (Tet) as indicated above lanes. The parental FlpIn T-REx line expressing no fusion protein serves as a negative control (FLAG-only), and FlpIn T-REx expressing FLAG-TOE1 at near endogenous levels was included as a reference. B-Actin was used as a loading control. *: non-specific bands. (B) Silver-stained SDS-PAGE gel examining protein from FLAG-Angel1-WT IPs. An input sample of 10% of the total cell extract used for the IP (Input), a sample from the remaining lysate after IP containing unbound proteins (Unbound), and a sample of the pooled elutions (Elution) were run. A control IP using a FlpIn T-REx line expressing no fusion protein (FLAG) was run alongside. *: denotes the antibody heavy and light chain.

Supplementary Figure 3.2: Extended eCLIP analysis. (A) Calculated transcripts per million (TPM) compared between eCLIP replicate 1 and a control sample expressing no fusion protein (FLAG). (B) Calculated transcripts per million (TPM) compared between eCLIP replicate 2 and a control sample expressing no fusion protein (FLAG) plotted against one another. Red lines represent equal TPM values between samples. Calculated Pearson correlations are shown. (C) Logo plots of top three enriched nucleotide motifs identified by MEME analysis of sequences within 50 nucleotides upstream or downstream of Angel1 CLIP peaks as compared to sequences within 50 nucleotides of mapped reads in the FLAG input sample. E-values are displayed above each plot. (D) Box plots comparing predicted mean free energy (MFE) calculated by RNAfold for regions within 50 nucleotides upstream or downstream of Angel1 CLIP peaks and within 50 nucleotides of mapped reads in the FLAG input sample. Lower MFE scores are associated with stronger predicted secondary structure. *: p-value < 2.2e-22 (KS-test).

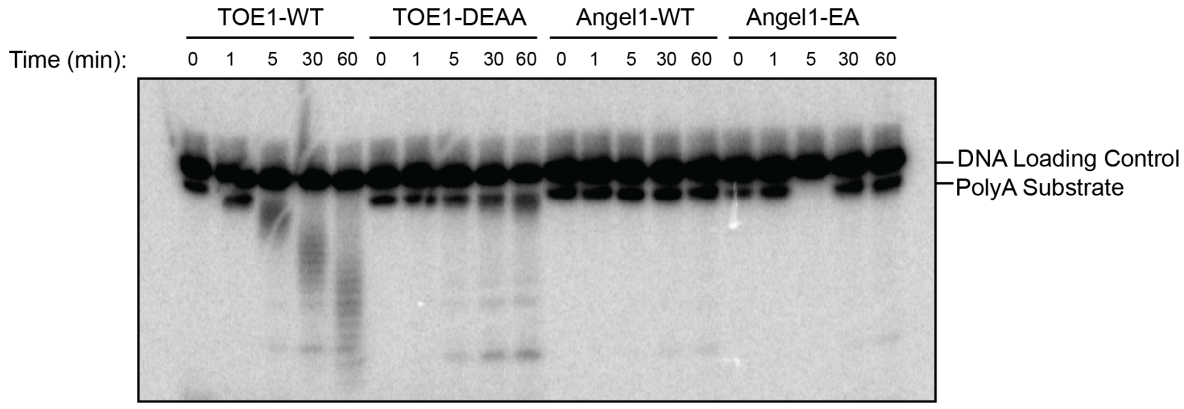

A
Pearson: 0.643
Log(Angel1 CLIP1 TPM)
Log(Flag IP TPM)
10000
100
1
1
10
100
1000
10000
B
Pearson: 0.72
Log(Angel1 CLIP2 TPM)
Log(Flag IP TPM
C
2.4e-301
3.1e-146
9.8e-93
bits
D
MFE
0
-20
-40
-60
*
Flag
Peaks



Supplementary Figure 3.3: Validation of siRNA-mediated depletions. (A) Relative expression of knockdown targets compared to a control knockdown (siLuc) obtained from RT-qPCR. Error bars represent standard error of the mean (n=3). Comparisons between each knockdown and its control had a p-value < 0.05 as calculated by a one-tailed Student's t-test. (B) Western blots examining levels of SKIV2L and XRN1 proteins after siRNA-mediated knockdown (KD) in triplicate or knockdown with a non-targeting control siRNA (Ctrl). B-Actin is shown as a loading control.



Supplementary Figure 3.4: Additional BG-NSD validation. (A) Relative expression of Angel1 mRNA levels determined by RT-qPCR for two independent Angel1 siRNAs compared to a non-targeting control (siLuc). Comparisons between each knockdown and its control had a p-value of < 0.05 as calculated by a one-tailed Student's t-test. Error bars represent standard error of the mean (n=3). (B) Exponential decay graphs of pulse-chase mRNA decay data generated from Northern blots using the BG-NSD substrate with depletion using the Angel1-targeting siRNAs used in panel A and siLuc depletion. (C) Exponential decay graphs of pulse-chase mRNA decay data generated from Northern blots using the TPI-NSD substrate with depletion of Angel1 or a non-targeting control. Error bars represent standard error of the mean. *: p-value < 0.05. n=3. (D) A Western blot examining total protein from Hela Tet-off cells either depleted by a non-targeting siRNA (siLuc) or an siRNA targeting endogenous Angel1 (siAngel1). Lanes 3 and 4 represent cells that were transfected with constructs expressing siRNA-resistant active (WT) or catalytically dead (EA) FLAG-Angel1 to complement depletion. B-Actin is shown as a loading control.



Supplementary Figure 3.5: Angel1 has no activity in a deadenylation assay. Phosphorimager scan of a 5' 32P-labelled poly-A RNA substrate incubated with the indicated active (WT) or catalytically dead (DEAA, EA) proteins (≈ 50nM) for the indicated amounts of time and subsequently separated in a denaturing gel. A 5' 32P-labelled DNA substrate was included in each reaction as an internal loading control.

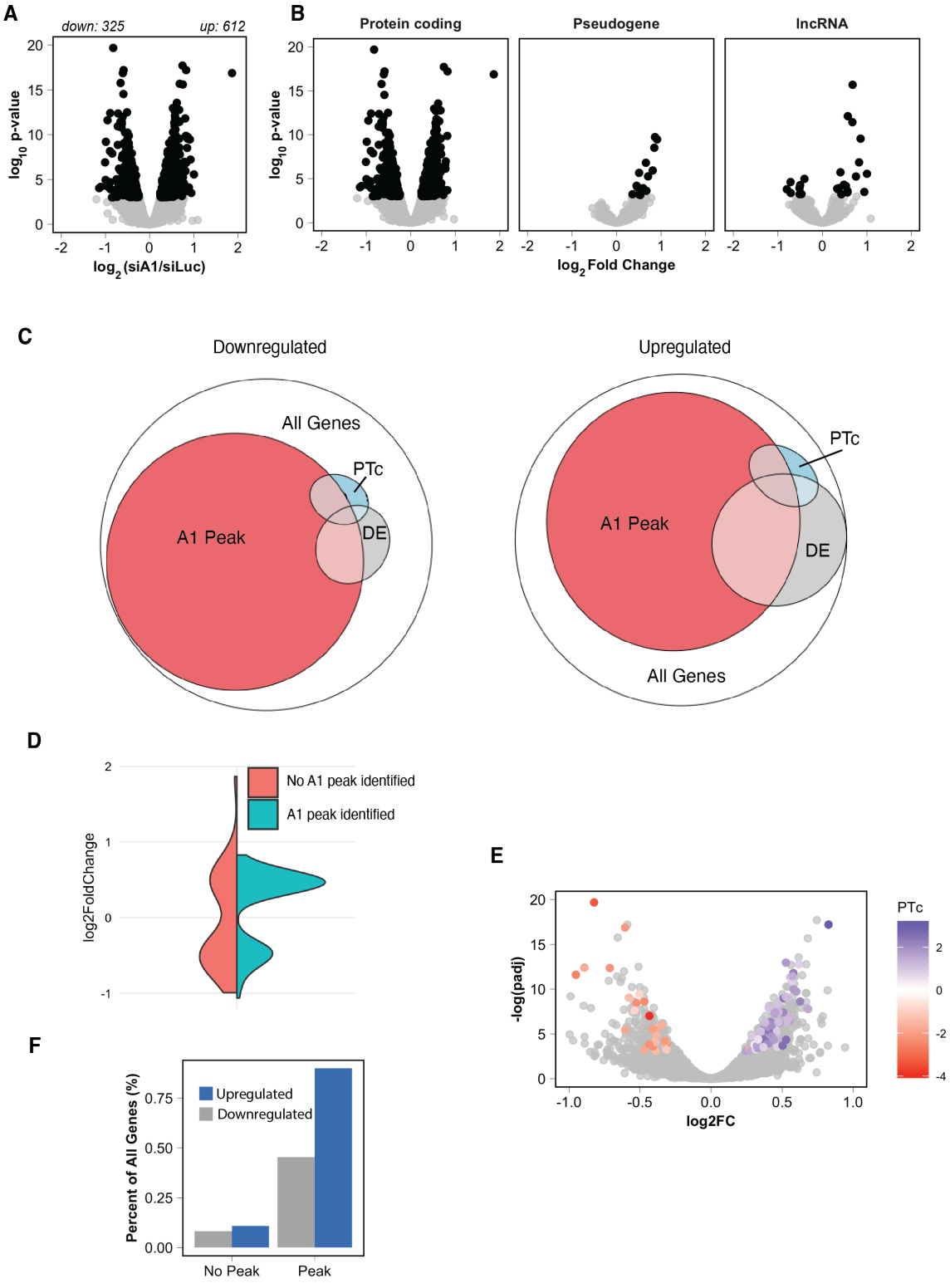
3.7 Supplemental analysis

3.7.1 Angel1 depletion causes minor upregulation of distinct subsets of RNAs

We sought to identify endogenous RNAs upregulated upon Angel1 depletion and performed RNA sequencing in the absence or presence of Angel1-targeting siRNA in HEK293 cells (Supplementary Figure 3.4), the same background in which we performed eCLIP. Similar to experiments performed previously for established RQC factors (Kalisiak et al. 2017; Weber et al. 2020), gross changes in transcript

levels upon Angel1 knockdown were subtle (Supplementary Figure 3.6A), but we found that more genes were increased (612) than decreased (325) consistent with a decay factor. A subset of upregulated genes were predicted regulated at the post-transcriptional level based on exon-intron split analysis (Gaidatzis et al. 2015) (Supplementary Figure 3.6E). Comparing the RNA-Seq dataset to our eCLIP dataset showed that RNAs that were significantly upregulated upon Angel1 depletion were more likely to be associated with Angel1 than those that were downregulated (Supplementary Figure 3.6C), consistent with a subset of upregulated transcripts being direct targets of Angel1. When genes were broken down by biotype, we noticed transcribed pseudogenes and long non-coding (lnc)RNAs as categories with a skew towards upregulation (Supplementary Figure 3.6B). While these classes of RNAs are typically considered non-translated, reports have documented translation of a subset of these (Brunet et al. 2020).

Supplementary Figure 3.6: Angel1 depletion upregulates distinct subsets of genes. (A) A volcano plot of differential expression after depletion of Angel1. Black dots represent significant genes ($p<0.05$). (B) Volcano plots of the same differential expression data subsetted by gene biotype. Black dots represent significant genes ($p<0.05$). (C) Venn diagrams showing overlap between genes that showed at least one significant CLIP peak (A1 Peak), were calculated to be significantly post-transcriptionally regulated (PTc), and significantly differentially expressed (DE). Genes were separated by whether they were significantly upregulated in DE or downregulated in DE. (D) Violin plots showing magnitude of change for significantly differentially regulated genes. Plots are split by genes that did not contain a significant CLIP peak (blue) and those that did not (red). I Differential expression overlaid with calculated strength of post-transcriptional regulation based on exon-intron split analysis. Only genes that were calculated to be significantly post-transcriptionally regulated and significantly differentially expressed are colored. Blue are predicted positively post-transcriptionally regulated and red dots are predicted negatively post-transcriptionally regulated. (F) A bar plot splitting all genes identified in RNA-seq by whether or not they found to have an eCLIP peak (peak vs no peak) and then by whether or not they were significantly upregulated (grey vs blue). Normalized to total number of identified genes.


A
down: 325
up: 612
log10 p-value
log2 (siA1/siLuc)
B
Protein coding
Pseudogene
lncRNA
log10 p-value
log2 Fold Change
C
Downregulated
All Genes
PTc
A1 Peak
DE
Upregulated
PTc
A1 Peak
DE
All Genes
D
No A1 peak identified
A1 peak identified
log2FoldChange
E
-log(padj)
log2FC
PTc
F
Upregulated
Downregulated
Percent of All Genes (%)
No Peak
Peak

3.7.2 Knockout of Angel1 using a Synthego kit only recovered in-frame indel mutants

siRNA-mediated depletion is an invaluable tool for understanding gene function; however, this depletion is incomplete and leads to a hypomorphic phenotype. A complete knockout of Angel1 could potentially lead to more readily detectable phenotypes and would also allow us to avoid the double-knockdown treatments that can negatively impact the health of cells. Using an established gene knockout procedure (Synthego, Gene Knockout v2), I proceeded to generate Angel1 knockout lines using Hela Tet-off cells and Flp-In TREx cells according to the manufacturer's protocol. While individually plated cells were growing, I sent pooled cells for Sanger sequencing and used Synthego's ICE tool (Inference of CRISPR Editing) to estimate editing at the Angel1 locus (Supplementary Figure 3.7). ICE analysis showed high editing efficiency; however, the majority of predicted indels were in-frame and would preserve the catalytic domain. There was a small percentage of predicted out-of-frame indels which are predicted to knock out the gene, so I proceeded to screen ~40 clones (data not shown). However, none of the screened clones contained out-of-frame deletions or insertions.

One potential reason explaining why I was unable to recover Angel1 knockouts would be if the Angel1 gene is essential. More experiments would need to be performed before I could attach any confidence to that conclusion, but if true, it would be exciting evidence of the fundamental importance of Angel1 to the cell. If

Angel1 were essential, it may also be because of functions of Angel1 unrelated to its catalytic activity. Therefore, it could be possible to knock-in a catalytic domain containing the inactivating E298A mutation (Figure 3.5). This mutant would also be an excellent tool to examine Angel1 function.

Supplementary Figure 3.7: A screenshot of Synthego ICE analysis of T/O cell line knockouts of Angel1. ICE analysis utilizes Sanger sequencing data to estimate CRISPR editing efficiency.

3.8 Acknowledgements

We would like to thank the members of the Lykke-Andersen lab for their input and useful discussions. We also would like to thank the Triton Shared Compute Cluster (TSCC) at the San Diego Supercomputer Center for use of their hardware for alignments. Sequencing was conducted at the IGM Genomics Center, University of California, San Diego, La Jolla, CA. This work was supported by National Institutes of Health (NIH) grant R35 GM118069 awarded to J. L.-A.

Chapter 3, sections 1-6, are a reprint of material that is being prepared for submission. “The 2’,3’ cyclic phosphatase Angel1 facilitates mRNA degradation during human ribosome-associated quality control,” Nicholson-Shaw, T., Ajaj, Y., Perelis, M., Fulzele, A., Yeo, G., Bennett, E., Lykke-Andersen, J., 2022. I was the primary author.

Chapter 4

An exploration of potential functional roles of Angel2

4.1 Introduction

The Angel gene was first discovered in the fruit fly *Drosophila melanogaster* as a member of a tumor suppressor locus (Kurzik-Dumke and Zengerle 1996). The unique name of this protein appears to come from the senior author of that study, Angelika Zengerle, rather than any celestial fly phenotype. From studies of protein homology, we have learned that Angel proteins are related to EEP-type deadenylases, typified by the catalytic CCR4 proteins of the CCR4-Not complex, the main eukaryotic cytoplasmic deadenylase (Goldstrohm and Wickens 2008). However, proteins containing EEP catalytic domains have shown activity as endonucleases, exonucleases, and phosphatases (Laothamatas et al. 2020; Winkler and Balacco 2013; Wallace et al.) Angel proteins are widely distributed among eukaryotes; humans have two, Angel1 and Angel2. A discussion of Angel1 can be found in chapter 3 of this dissertation, but in this chapter, I present unpublished data I have collected regarding Angel2 in the hope that it can serve as a foundation for a future study on Angel2 function.

Since the discovery of the *D. melanogaster* Angel protein in 1996, virtually no work has been published on Angel2, until 2020 when the Martinez lab published

evidence that Angel2 functions as a 2',3' cyclic phosphatase (Pinto et al. 2020). The authors also found no activity of Angel2 as a deadenylase which caused a huge shift in my thinking regarding this protein because I had been pursuing hypotheses that focused on potential function as a deadenylase. As of the writing of this dissertation, a cellular function for Angel2 has yet to be established.

In this chapter, I present unpublished, discovery-based experiments that I performed targeting Angel2. With the identification of Angel2 as a cyclic phosphatase, I was able to go back and examine this data through a different lens that provided new insights into old data. Analysis of IP-MS/MS data collected in collaboration with the Bennett lab shows evidence that Angel2 associates with components of the tRNA splicing ligase. Analysis of eCLIP data collected in collaboration with the Yeo lab shows evidence that Angel2 associates with tRNAs, but not necessarily a greater proportion of intron-containing tRNAs, which undergo a splicing process with a 2',3' cyclic phosphate intermediate. RNA-seq data shows that depletion of Angel2 upregulates a subset of mRNAs depleted of rare codons and codons recognized by spliced tRNAs. Finally, I present an assay that, with some improvements, could be used to measure proportions of cyclic phosphates in candidate RNA species.

4.2 Results

4.2.1 Angel 2 associates with components of the tRNA splicing ligase

To gain a better understanding of potential functions of Angel2, I wanted to examine its protein binding partners. To facilitate this assay, I used the Flp-In TRex (FITR) system to create a HEK 293 cell line expressing N-terminally FLAG-tagged Angel2 under the control of a tetracycline-regulated operator. Using this system, I expressed Angel2 at approximately endogenous levels and performed immunoprecipitation (IP) against the FLAG tag. I performed these IPs with and without prior treatment with RNase A to help distinguish between protein-mediated and RNA-mediated interactions. Furthermore, as a control I performed an identical IP in cell lines expressing FLAG-tagged TOE1, a better understood protein involved in snRNA maturation (Lardelli et al. 2017).

In collaboration with the Bennett lab, we performed liquid chromatography followed by tandem mass spectroscopy (IP-MS/MS). We identified a wide variety of proteins enriched by Angel2 pull-down, but among the top ten proteins identified (Figure 4.1A), which include Angel2 itself, five were members of the tRNA splicing ligase, RTCB, FAM98B, FAM98A, C14orf166, and DDX1. Importantly, these proteins were not enriched in the control FLAG-TOE1 IP indicating that this interaction was specific to Angel2. Apart from RTCB and C14orf166, these interactions were sensitive to RNase treatment, suggesting that these interactions may be RNA mediated while the RTCB and C14orf166 interactions may be direct protein interactions. Given this association with a tRNA processing factor, and Angel2's homology to RNA processing enzymes, I sought to ask what RNAs Angel2 may bind to.

A

	Log2 Fold Change Over Input			
IP:	FLAG-TOE1		FLAG-Angel 2	
RNase:	-	+	-	+
ANGEL2	1.00000000	1.00000000	180.13285080	123.96253930
DDX1	1.76196282	0.18672739	42.67580610	7.30625426
RTCB	1.00000000	1.00000000	33.17079769	33.40873129
HNRNPH3	2.52392563	0.12108002	24.39694377	3.23723716
FAM98A	2.52392563	0.25617376	20.74117131	2.20964822
C14orf166	1.00000000	1.00000000	17.08539884	21.01715756
AKAP8L	1.17632265	0.14690326	15.80063755	2.10729371
HNRNPUL2	6.33373971	0.04127319	14.16078087	2.59847202
PRMT1	0.23301046	0.10297811	12.82288749	32.39713412
FAM98B	1.00000000	0.18672739	12.69847189	1.61063277

B

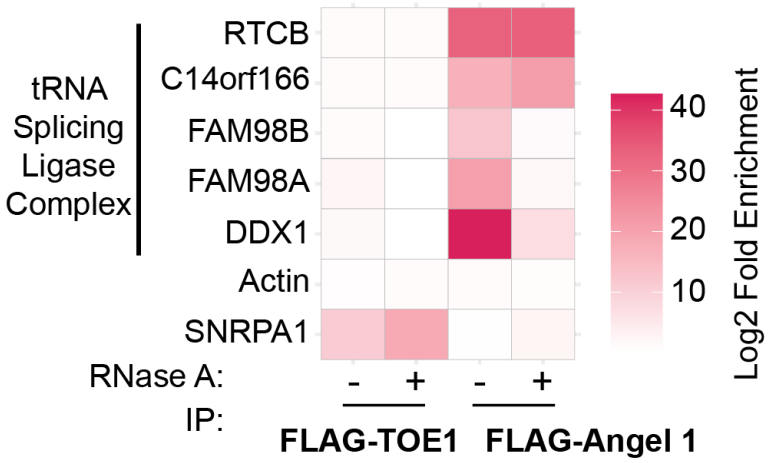


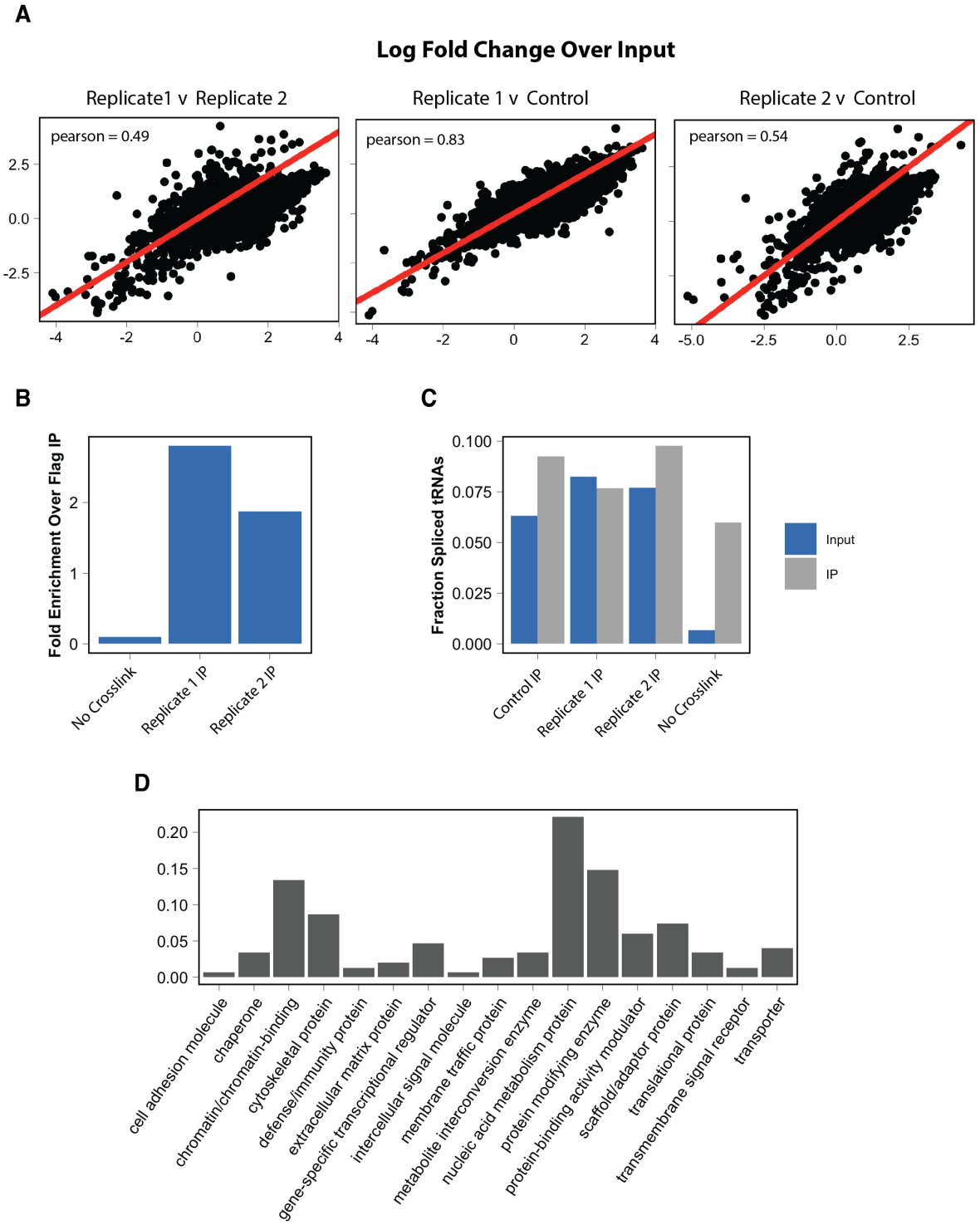
Figure 4.1: IP-MS/MS identifies key components of the tRNA splicing ligase as interaction partners of Angel2. (A) The top ten proteins identified as sorted by no RNase treatment. Highlighted proteins are members of the tRNA splicing ligase complex. (B) A heatmap of Angel2 and TOE1's association with proteins of interest. Samples were treated with or without RNase A. Fold enrichment was calculated over peptide counts found in the FLAG control IP.

4.2.2 Angel 2 eCLIP replicates show poor correlation

Using the same FLAG-Angel2 FITR cell lines, in collaboration with the Yeo lab, we performed enhanced cross-linking and immuno-precipitation (eCLIP) (Van Nostrand et al. 2017). As controls, we performed eCLIP using a cell line that only expressed the FLAG peptide and another FLAG-Angel2 preparation that left out the key step of UV cross-linking. The data was aligned to the hg38 human genome.

One predicted outcome from this type of experiment is that the sequencing data from the two replicates should be more alike than either are to the control. Fold change of the reads in the IP over reads in the input was calculated. When replicates were graphed against each other (Figure 4.2A), I found a middling correlation (left panel); however, when either replicate was mapped against the negative control, I find a greater correlation (middle and right panels). This evidence is inconsistent with our expectations for this experiment and suggests that there might be experimental issues that should be considered in the context of the data presented below.

Figure 4.2: eCLIP data shows poor correlation but suggests Angel2 interacts with tRNAs. (A) Log2 fold change of TPM over input for samples plotted against one another. Pearson correlations are displayed. (B) Display of Angel2 tRNA association by eCLIP over FLAG control. (C) Measurement of the fraction of tRNAs that are produced from spliced pre-tRNAs found in individual eCLIP assays. (D) GO protein enrichment analysis of peaks reproducible between the two Flag-Angel2 eCLIP experiments.


A
Log Fold Change Over Input
Replicate1 v Replicate 2
pearson = 0.49
Replicate 1 v Control
pearson = 0.83
Replicate 2 v Control
pearson = 0.54
B
Fold Enrichment Over Flag IP
No Crosslink
Replicate 1 IP
Replicate 2 IP
C
Fraction Spliced tRNAs
Control IP
Replicate 1 IP
Replicate 2 IP
No Crosslink
Input
IP
D
cell adhesion molecule
chaperone
chromatin/chromatin-binding
cytoskeletal protein
defense/immunity protein
extracellular matrix protein
gene-specific transcriptional regulator
intercellular signal molecule
membrane traffic protein
metabolite interconversion enzyme
nucleic acid metabolism protein
protein modifying enzyme
protein-binding activity modulator
scaffold/adaptor protein
translational protein
transmembrane signal receptor
transporter

4.2.3 Angel2 shows association with reads mapping to tRNA genes

Because Angel2 showed association with the tRNA splicing ligase I was interested in examining tRNA reads in our eCLIP data. Typically, tRNAs are poorly represented in sequencing experiments because mature tRNAs are highly modified and often stall reverse transcriptases used to prepare libraries. However, we were able to find some representation of reads that mapped to tRNAs and calculated enrichment over tRNAs found in the FLAG control sample (Figure 4.2B). We found that both Angel2 IPs were enriched over the negetive control IP for tRNA reads while the non-crosslinked control was depleted. This suggests that Angel2 directly binds to tRNAs.

Only a small fraction of genomic tRNAs contain introns and require the activity of the tRNA splicing ligase for full maturation. I was therefore interested in the proportion of spliced tRNAs in our eCLIP samples. I found little difference in the proportion of spliced tRNAs found in input and IP samples suggesting that Angel2 does not prefer intron-containing tRNAs (Figure 4.2C). However, counter to our expectations we did see a slight increase in the proportion of spliced tRNAs in the control sample and a robust increase for the no crosslinking control. This may be indicative of the unreliability of examining tRNA reads using current sequencing technology.

Using the eCLIP pipeline, I called peaks and examined ones that were reproducible between the two datasets. Performing gene ontology (GO) enrichment

analysis (Thomas et al. 2022), I found that the most enriched class of RNAs were ones that coded for nucleic acid metabolism proteins (Figure 4.2D). It is unclear why Angel2 may be interacting with these types of RNAs.

4.2.4 RNA-seq analysis of cells depleted of Angel2

To understand what transcripts may be regulated by Angel2, I examined changes in the transcriptome after siRNA-mediated Angel2 depletion (Supplementary Figure 4.1). I used the same FITR FLAG-Angel2 (FLAG-Angel2 WT) cell lines and created a new FITR HEK 293 line expressing a catalytically dead (Pinto et al. 2020) mutant of Angel2 (FLAG-Angel2 EA). I performed control and Angel2 knockdowns in the parental HEK 293 line and knocked down endogenous Angel2 in the cell lines that exogenously express Angel2 WT or mutant. I used these lines to rescue the Angel2 depletion with either a WT or catalytically dead, siRNA-resistant, Angel2 expressed at close to endogenous levels. This approach could allow us to find changes in gene signatures that were a direct result of the catalytic activity of Angel2.

Sequencing reads were aligned to the human genome and examined with principal component analysis (PCA) to find agreement between the samples. I reasoned that the Angel2 knockdown and the catalytically dead addback would be more alike in this analysis than the control knockdown and WT addback. However, the samples did not cluster as expected (Figure 4.3A). Opposite samples clustered together along principal component 1 (PC1) which would suggest that those two samples are more alike than their counterpart sample: The Angel2 knockdown and

the catalytically dead addback samples clustered together in PC1 as expected; however, the WT and control knockdowns were the farthest apart, counter to expectations. These issues may be due to variability when exogenously expressing a factor, be indicative of a poor-quality dataset, or that there are off-target Angel2 siRNA effects that are not rescued by Angel2 addback. The analyses presented below should be viewed in that context.

4.2.5 The gene signature upon Angel2 depletion is consistent with a decay factor

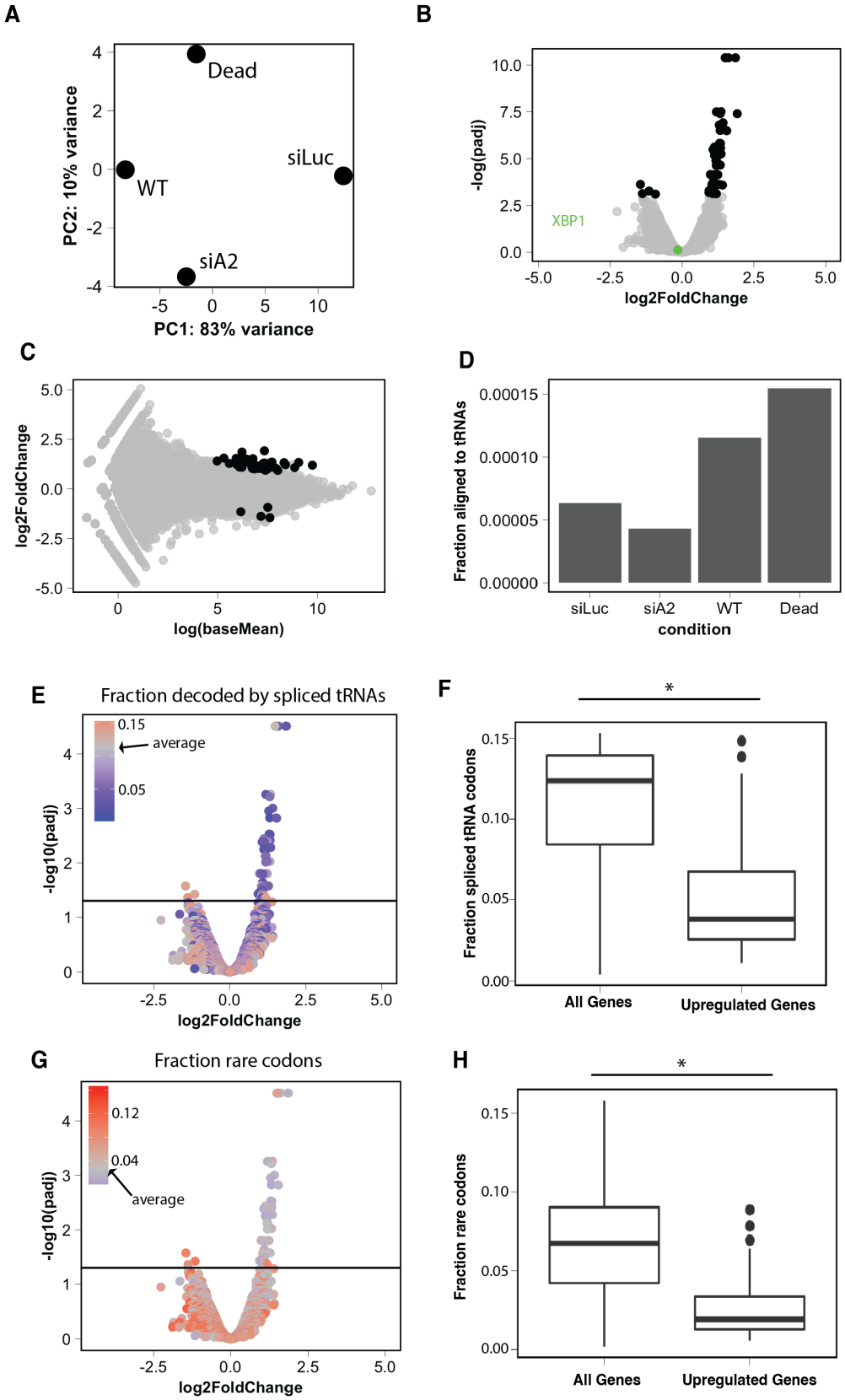
Differential expression between the samples was determined using DE-seq (Love et al. 2014). To perform differential expression analysis, the siLuc and WT addback samples were grouped as control, while siAngel2 and catalytic dead add-backs were grouped as perturbation. This experimental design is less than optimal and will in future studies need to be conducted with multiple replicates for each treatment. Consistent with a decay factor, I found that most genes that were significantly affected by Angel2 depletion were upregulated (Figure 4.3B). These upregulated genes were some of the most abundant genes in the transcriptome (Figure 4.3C). Gene Ontology analysis of these upregulated genes gave few insights (data not shown).

Because Angel2 has been identified as a cyclic phosphatase, I was especially interested in changes in the XBP1 mRNA, since splicing of this mRNA proceeds

through a cyclic phosphate intermediate (Jurkin et al. 2014). I did not find any significant change in the levels of the XBP1 mRNA (Figure 4.2B green dot). This may suggest that Angel2 does not affect the accumulation of XBP1 mRNA, but it may also be due to testing in a context that did not activate the unfolded protein response, which promotes splicing of XBP1 mRNA.

I also mapped reads to genomic tRNA regions and calculated the fraction of those reads to the total number of reads in the sample (Figure 4.3D). Exogenous expression of either WT or catalytic dead Angel2 correlated with a greater fraction of tRNA reads. However, increased fraction of tRNAs did not correlate with catalytic activity of Angel2.

Figure 4.3: RNA-seq analysis shows upregulation of genes containing fewer spliced tRNA codons and rare codons. (A) A PCA plot examining clustering of the samples. Dots closer in distance are more alike. Note that 83% of variability is described by principal component 1 along the x-axis. (B) A volcano plot of differential expression as determined by DE-seq. siLuc and WT addback were grouped as control and siAngel2 and Dead addback were grouped as perturbation. (C) An MA plot (ratio intensity) of the same differential expression data. (D) Fraction of reads that map to tRNAs. (E) The volcano plot in panel *B* overlayed with information about codons decoded by spliced tRNAs. Blue dots represent genes that have a lower-than-average proportion of codons recognized by spliced tRNAs. (F) Boxplot showing the fraction of codons recognized by spliced tRNAs of all protein-coding genes versus protein-coding genes that were upregulated by Angel2 depletion. p-value was calculated with a Kalmogorov-Smirnov (KS) test. (G) Volcano plot overlayed with rare codon information. (H) Boxplot showing the fraction of rare codons in all protein-coding genes compared to rare codons in upregulated protein-coding genes. p-value was calculated with a KS-test.


A
Dead
siLuc
WT
siA2
PC2: 10% variance
PC1: 83% variance
B
-log(padj)
XBP1
log2FoldChange
C
log2FoldChange
log(baseMean)
D
Fraction aligned to tRNAs
siLuc
siA2
WT
Dead
condition
E
Fraction decoded by spliced tRNAs
0.15
average
0.05
-log10(padj)
log2FoldChange
F
*
Fraction spliced tRNA codons
All Genes
Upregulated Genes
G
Fraction rare codons
0.12
0.04
average
-log10(padj)
log2FoldChange
H
*
Fraction rare codons
All Genes
Upregulated Genes

4.2.6 Upregulated genes are less likely to contain codons translated by spliced tRNAs and rare codons

If Angel2 were to regulate or affect tRNAs, one potential way that this may lead to dysregulation of genes is by altering the composition of tRNAs in the cell. This could lead to upregulation or downregulation of mRNAs depending on the composition of their codons due to decoding tRNA availability. Because of Angel2's association with components of the tRNA splicing ligase, I was particularly interested in correlations with the proportion of codons decoded by spliced tRNAs. I measured the fraction of codons recognized by intron-containing tRNAs across all protein-coding genes observed and marked genes that contained less than the average fraction (Figure 4.3E). I found that almost all significantly upregulated genes had a lower-than-average fraction of codons recognized by spliced tRNAs (dark blue dots). When comparing the distribution of codons recognized by spliced tRNAs genome-wide to just upregulated genes, I found that upregulated genes had far lower fractions of spliced tRNA codons (Figure 4.3F).

Similarly, I was interested in proportions of rare codons. These codons are particularly sensitive to changes in tRNA availability. Overlay of the fraction of rare codons revealed that most upregulated mRNAs had a low proportion of rare codons (Figure 4.3G). Comparing these proportions genome-wide showed that upregulated mRNAs had a significantly lower proportion of rare codons (Figure 4.3H). This data suggests that depletion of Angel2 may have a upregulatory effect on mRNAs that have low levels of codons recognized by intron-containing tRNA and of rare codons.

These observations will have to be verified with more complete datasets in future studies.

4.2.7 A protocol for investigating proportions of cyclic phosphate

Angel2 and Angel1 were recently identified as cyclic phosphatases (Pinto et al. 2020). One prediction based on this function is that depletion of Angel2 or Angel1 should increase the proportion of target RNAs terminating in a 2',3' cyclic phosphate. To examine potential changes in proportions of cyclic phosphate, I sought to establish an assay that could measure the fraction of a particular RNA species that terminated in a cyclic phosphate. I made use of steps from a published cyclic phosphate sequencing protocol (cP-seq) (Honda et al. 2016) that leveraged sodium periodate. Treatment with periodate destroys the ends of species terminating in 3'OH, leaving them unable to undergo RNA ligation. The protocol, schematized in Figure 4.4A and presented in more detail in Appendix B, begins with splitting the RNA sample into two parts. The first receives a mock treatment while the second is treated with periodate. Afterwards, both are treated with poly-nucleotide kinase (PNK). In the mock/PNK-treated sample, all RNAs should be available for ligation, while in the periodate/PNK treatment, only species originally terminating with a cyclic phosphate will be available for ligation. After ligation and reverse transcription from the adapter, specific RNA species can be monitored by qPCR. The qPCR signal from the mock/PNK treatment (representing the total RNA population) is compared to the

signal from periodate/PNK treated sample (representing cyclic phosphate species only) which gives a proportion of cyclic phosphates for a particular RNA. A periodate-treated sample without PNK serves as a control for the effectiveness of the periodate treatment.

To optimize this protocol, I made use of U1 and U6 snRNAs. A large proportion of U6 snRNA terminates in a cyclic phosphate created by the enzyme USB1 (Mroczek and Dziembowski 2013; Gu et al. 1997). Conversely, U1 snRNA terminates with a 3'OH. After treatment with or without PNK and periodate, I separated total RNA in a denaturing gel and examined the position of U6 snRNA by Northern blotting (Figure 4.4B). Treatment with PNK was necessary for the majority of U6 snRNA to be ligated while periodate alone did not allow anything to be ligated, consistent with expectations. Using the above-described qPCR assay, I examined U6 and U1 snRNA signals under the same conditions (Figure 4.4C). Treatment with periodate alone showed no signal for either U1 or U6 snRNAs. Treatment with both periodate and PNK showed no signal for U1 snRNA and increased signal for U6 snRNA, consistent with U6 snRNA terminating in a cyclic phosphate. With more optimization, particularly for mRNAs, this assay has the potential to measure proportions of cyclic phosphate RNA species in the cell, including those targeted for RQC (see Chapter 3).

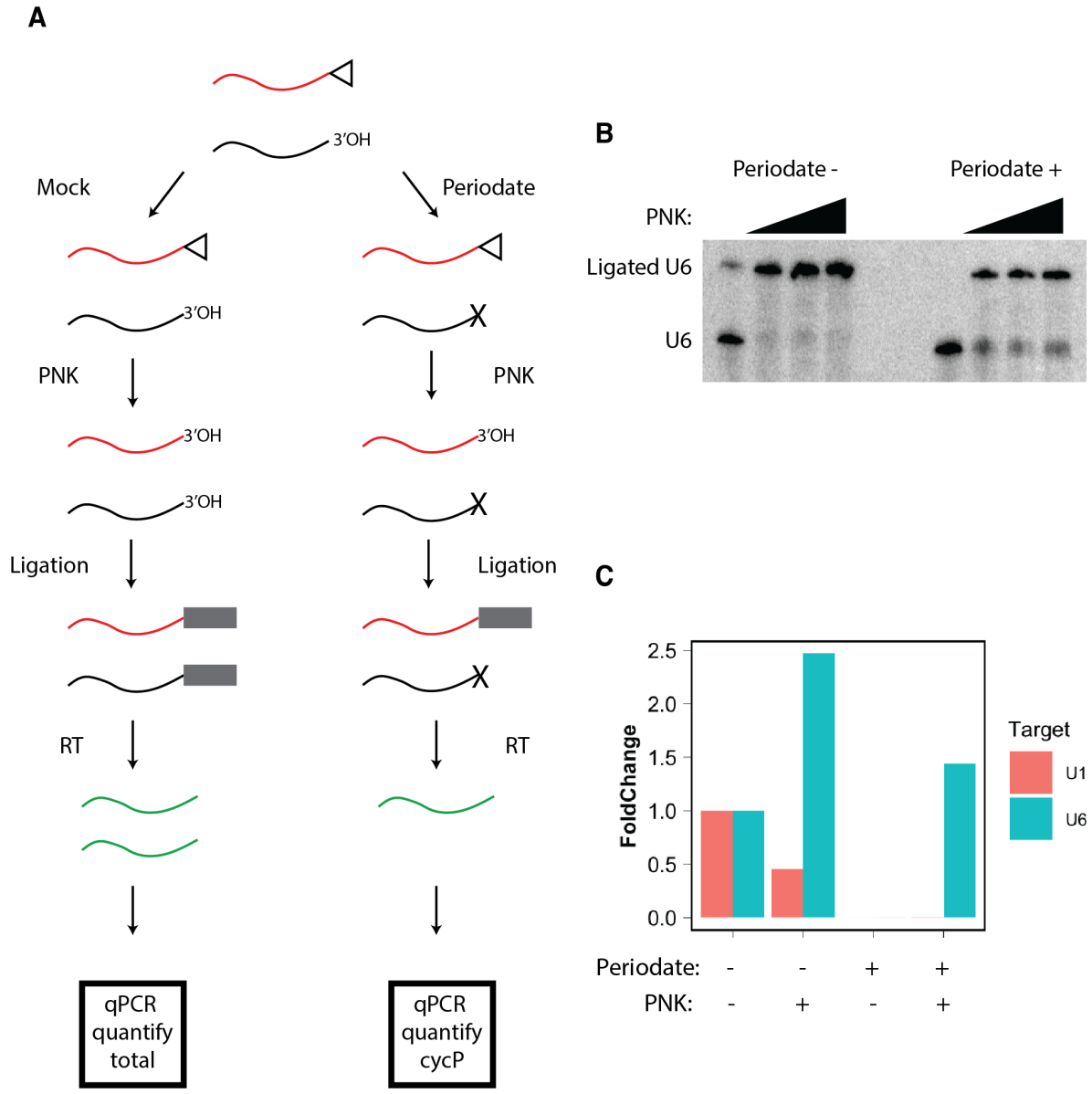


Figure 4.4 A qPCR-based assay to examine proportions of cyclic phosphate in the cell. (A) An experimental schematic. Periodate leaves 3'OH ends unligatable while PNK treatment allows 2',3' cyclic phosphate terminating species to be ligated. Using an adapter and gene-specific reverse transcription, cDNA populations can be generated representing all RNA and just RNAs terminating in 2',3' cyclic phosphate. (B) A northern assay to validate treatments. The upper band represent ligated U6 snRNA while the lower band is not ligated. (C) qPCR quantification of U1 and U6 snRNAs in the same conditions. Fold change over no treatment is shown.

4.3 Discussion

This chapter presents evidence that Angel2 associates with components of the tRNA splicing ligase, may bind directly to tRNAs, and that Angel2 depletion upregulates a set of mRNAs that are depleted for spliced tRNA codons and rare codons. I also present a protocol to examine proportions of 2',3' cyclic phosphates for targeted RNA species.

How reliable is this data?

The data presented here concerning Angel2 was created from discovery-based experiments. The quality control metrics for IP-MS/MS looked excellent. Angel2 was the most enriched protein in the dataset and enriched proteins were also not pulled down in the TOE1 control (Appendix A) (or the Angel1 control, which was carried out at the same time, see chapter 3). This dataset seems to be highly reliable, but the evidence would be strengthened by testing associations through orthogonal methods such as Western blotting for associated proteins after IP. It would also be important to perform these co-IP assays with prior RNase treatment to confirm which interactions are RNA-mediated.

The data for Angel2 eCLIP should be considered preliminary and would require repetition before confidence in the conclusions. The individual Angel2 CLIP assays were more similar to the FLAG control CLIP rather than the replicate CLIP. This could potentially indicate problems with the experiment and that most interactions were not specific to Angel2. All CLIP peaks were called as enriched over

input which should control for non-specific interactions. However, follow-up experiments need to be performed to confirm the key eCLIP results, for example by Northern blotting for tRNAs in RNA Immuno-Precipitation samples (RIPs) to verify the association of Angel2 with tRNAs (see Schaffer et al. 2014 for a potential tRNA blotting protocol).

The RNA-seq data also showed issues, specifically with clustering in the PCA plot. The WT and control knockdown samples failing to cluster suggests that there may be substantial off-target effects of the Angel2 siRNA that are not rescued by the WT addback and that many observed effects may not be due to Angel2 depletion. Alternatively, Angel2 add-back could have causes a squelching effect. Either way, the experimental design is insufficiently robust and needs to be further optimized in future study. Also, orthogonal methods such as qRT-PCR or northern blotting should be used to confirm level changes observed in the RNA-seq data.

What is Angel2's function?

Angel2's association with the tRNA splicing ligase, potential association with tRNAs, and documented activity as a 2',3' cyclic phosphatase paint an enticing picture of Angel2 functioning in tRNA splicing. Spliced tRNA introns are cleaved by TSENs which leave behind a 2',3' cyclic phosphate that is used by the tRNA splicing ligase to ligate the two exons together (Abelson et al. 1998). If Angel2 is associated with this process, we might predict it to inhibit splicing by removing the cyclic phosphate, thereby interrupting ligation. In this model, Angel2 depletion might lead to increased levels of intron-containing tRNAs and potentially decreased levels of 5'

tRNA fragments. Specifically sequencing tRNAs and monitoring how levels change and which parts of tRNAs are stabilized may be an excellent line of inquiry to investigate Angel2's function here. However, inconsistent with this model is that Angel2 did not seem to favor intron-containing tRNAs over non-intron-containing tRNAs (Figure 4.2C). This may indicate that Angel2 interacts more generally with tRNAs.

It is important to note that tRNAs are typically not well-represented in sequencing data using normal workflows because of their high level of post-transcriptional modifications, which can stall reverse transcriptases. It may be that the species that we are observing in these sequencing datasets are immature species before they have had a chance to be modified, or tRNA fragments (TERFs) which may not be modified in the same way. The best experiment to distinguish between these types of hypotheses would be a small RNA library prep on RNA that has been demethylated and PNK-treated so that tRNAs are correctly represented.

The most curious piece of evidence that does not seem to fit in with the other observations is that Angel2 depletion caused upregulation of mRNAs with a low proportion of rare codons and codons recognized by intron-containing tRNAs. If Angel2 is involved in biogenesis of spliced tRNAs, it might make sense to observe downregulation of genes that have a high proportion of codons recognized by spliced tRNAs, but my observation of upregulation of genes with a low proportion does not fit with this model. It is important to keep in mind that these observations are preliminary and that it is possible that the Angel2 siRNA had off-target effects.

Another potential function for Angel2 (or Angel1) is in the non-conventional splicing of the mRNA for the unfolded protein response (UPR) promoting transcription factor XBP1. During UPR, the endonuclease IRE1 cleaves out a small intron from XBP1 pre-mRNA (XBP1u) leaving behind a cyclic phosphate. The exons are spliced together by the tRNA splicing ligase creating an isoform that promotes UPR. Indeed Pinto et al, found that overexpression of Angel2 inhibited splicing of XBP1 mRNA which is consistent with its activity as a cyclic phosphatase (Pinto et al. 2020). However, inconsistent with this function is that Angel2 has a putative nuclear localization signal (NLS), potentially placing it in the nucleus, while XBP1 splicing takes place in the cytoplasm. I believe that the most likely hypothesis that should be pursued first is one involving tRNAs.

Future Directions

The most important next experiments would be to perform small RNA sequencing after RNA demethylation on samples depleted of Angel2. These samples could also be split off to perform traditional RNA-sequencing looking for differential expression. Some of the questions that could be asked are: Do tRNA levels change upon depletion of Angel2? Is there any difference in the level of changes for intron-containing tRNAs versus tRNAs that do not contain introns? Are specific regions of tRNAs upregulated, such as the 5’ versus the 3’ end? Do any upregulated or downregulated tRNAs map to regions from which TERFs are produced? Do tRNA level changes correlate with any of the codons in genes that are differentially expressed? Does Angel2 depletion change the expression of any genes involved in

the unfolded protein response? Do proportions of tRNAs terminating in a cyclic phosphate change with Angel2 depletion?

4.4 Methods

4.4.1 Cell lines, maintenance, and knockdowns

Angel2 was inserted into the pcDNA5/FRT/TO plasmid by Gibson assembly (NEB) according to manufacturer recommendations to generate pcDNA5-FLAG-Angel2-WT. Site directed mutagenesis (NEB) was used to make a catalytically inactivating mutation pcDNA5-FLAG-Angel2-EA. HEK293 Flp-In TRex (FITR) cell lines were used to make stable integrants according to the manufacturer's recommendations. All cells were maintained in Dulbecco's Modified Eagle Medium (DMEM, Gibco) supplemented with 10% fetal bovine serum (FBS). Cells were induced with 10 ng/ml tetracycline 24 hours before harvest. Knockdowns were performed 72 hours and 24 hours before harvest using siLentFect transfection reagent (BioRad) according to manufacturer protocols.

4.4.2 Immunoprecipitation

Flp-In TRex lines were grown to ~50% confluency and induced with 1 ng/ml tetracycline for 24 hours. Cells were harvested by scraping into ice-cold PBS and flash frozen in liquid nitrogen. Pellets were lysed with isotonic lysis buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.2 mM EDTA, 0.5% triton x-100, 80 U/ml RNase out

(for samples not treated with RNase), and 1 tablet/10 ml Pierce Protease Inhibitor Mini Tablets) for 10 minutes on ice. Lysates were spun down at 20,000*g* for 15 minutes at 4C. Half of the samples were treated with 125 µg/ml RNaseA for 30 minutes at 37˚C while the others received a mock treatment. FLAG peptide was added to lysates to a concentration of 1 µg/mL to reduce non-specific interactions. Samples were incubated with washed anti-FLAG M2 beads for 2 hours at 4˚C with rotation. Beads were washed 9 times with NET2 buffer (10 mM Tris, pH 7.5, 150 mM NaCl, 0.1% Triton X-100). Beads were eluted for 3 times for 30 minutes at 4˚C with NET2 containing 200 µg/mL FLAG peptide. Elutions were pooled.

4.4.3 IP-MS/MS and analysis

Samples were given to the Bennett lab who performed IP-MS/MS as described in Sundaramoorthy et al. 2017. IPs were also performed alongside using a Flp-In-TRex line that only expressed the FLAG peptide. A pseudo-count of 1 was added to every peptide count to prevent division by zero errors. Counts were normalized to the total count for each sample and then fold change was calculated over the matched negative control.

4.4.4 eCLIP and analysis

eCLIP was performed as outlined in Van Nostrand et al. 2017. Samples were mapped to the hg38 human genome and features from Ensembl 105 were counted with featureCounts (Liao et al. 2014). TPM and graphs were generated with custom

R scripts. Fraction mapping to tRNAs was measured by featureCounts against a database of genomic tRNAs obtained from gtRNAdb (Chan and Lowe 2016). GO enrichment analysis was performed using pantherdb (Thomas et al. 2022).

4.4.5 RNA-seq

Table 4.1: siRNA duplex sequences produced by Horizon Discovery

Target	Sequence
Luciferase	CGUACGCGGAAUACUUCGAUU
Angel2	GGGAAAUGUUUGGGAAAUAUU

Flp-In TRex lines were knocked down with Angel2 siRNA (Table 4.1) and either Angel2 WT or Angel2 EA was added back by induction with 10 ng/ml tetracycline 24 hours before harvest. After harvest in cold PBS, RNA was isolated according to the trizol protocol. Samples were submitted to UCSD's IGM genomics core for ribo-depleted, stranded RNA library preparation and sequencing on an Illumina HiSeq machine. FASTQ files were aligned to the hg38 human genome and features were counted with featureCounts against Ensembl 105. Differential expression analysis was performed with DEseq2 (Love et al. 2014). Fraction aligned to tRNAs was calculated using featureCounts and the above mentioned gtRNA database. Fraction spliced tRNA and fraction rare codons were calculated using custom python scripts and canonical CDS (CCDS) annotated FASTA sequences from Ensembl.

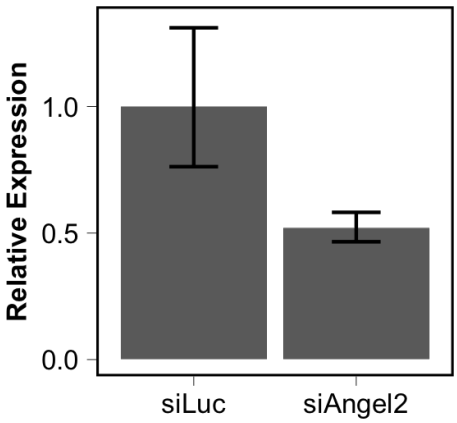
4.4.6 Cyclic Phosphate qPCR

This protocol was adapted from Honda et al. 2016. Total RNA was harvested from a 6-well plate of HeLa cells with Trizol (Thermo Fisher) according to the manufacturer's protocol. Samples were treated with 5 units of DNase I (Zymo) for 30 minutes at 37˚C. Samples were subsequently treated with Quick CIP (NEB) at 37˚C for 40 minutes. RNA was extracted using spin columns (Zymo) and split into two equal samples. Each sample was treated with 100 mM of sodium periodate or mock treated and extracted by spin column and split into two. Samples were then treated with 2 units T4 PNK (NEB) or mock treated for 40 minutes at 37˚C and then extracted again with spin columns (Zymo). Ligation was performed with 400 units of T4 RNA Ligase 2 truncated (NEB) and 2 μM of pre-adenylated adapter (5'APP-AG_fixed). These were subsequently extracted using spin columns (Zymo). After splitting samples, reverse transcription was performed with Superscript III (Life Technologies) using either an adapter-specific primer or random hexamers. A detailed protocol can be found in Appendix B.

4.5 Acknowledgements

Chapter 4 contains unpublished material prepared in collaboration with Isabel Breuer, Amit Fulzele, Mark Perelis, Eric Bennett, Gene Yeo, and Jens Lykke-Andersen. I was the primary author of this material.

4.6 Supplemental Figures



Supplementary Figure 4.1: Angel2 depletion validation. Expression of Angel2 mRNA levels in the knockdown condition relative to a non-targeting control (siLuc). Values were determined by RT-qPCR in technical triplicate but only 1 biological replicate. Error bars represent one standard deviation from the mean of the.

Chapter 5

Conclusions

Aside from the Nintendo 64, there is no perfect system; and so, cells need to maintain robust mechanisms to ensure the functional quality of cellular products, DNA, RNA, and proteins. One of these mechanisms, the Nonsense-Mediated mRNA Decay (NMD) pathway, has been the focus of decades of research leading to a solid understanding of its mechanisms and how premature termination codons (PTCs) promote surveillance. The rules and factors that govern No-Go Decay (NGD) and Non-Stop Decay (NSD) are less well understood. Whether NGD/NSD is triggered appears to depend on the context and the organism. For example, what constitutes a rare codon is determined by tRNA availability which is organism specific. So, substrates or sequences that trigger NGD/NSD in one organism, might not trigger it in another organism. Furthermore, there seems to be some decoupling between nascent peptide decay and mRNA decay. Substrates that trigger degradation at the protein level do not necessarily trigger decay at the mRNA level. All these nuances to NSD/NGD underscore the fact that we do not fully understand the features and factors that determine the functional outcomes of this process. The features, factors, and mechanisms that determine the maturation and degradation of small non-coding RNAs is similarly poorly understood.

In this work, I have sought to provide tools to further our knowledge in this arena and push forward how we understand NSD/NGD. Tailer (chapter 2) will make it easier for researchers to ask questions in the arena of non-coding RNA 3' end maturation and degradation. Chapter 3 examines Angel1, a cyclic phosphatase, that I identify as an important factor in NSD/NGD. Excitingly, this finding suggests that removal of a cyclic phosphate may be rate-limiting for NSD/NGD and potentially other RNA decay pathways. Finally, chapter 4 lays out all the data I have collected regarding Angel2 and raises interesting questions about Angel2's involvement in tRNA biogenesis or decay.

5.1 RNA quality control in health and human disease

Perturbations in pathways such as NMD, NSD, and NGD and dysfunctions in other machinery meant to degrade RNA products have been implicated in a substantial number of human diseases. Approximately 3-20% of genetic diseases are caused by a mutation that introduces a premature termination codon (PTC) (Frischmeyer and Dietz 1999; Nguyen et al. 2014) including diseases like cystic fibrosis, β-thalassemia, and Duchenne's muscular dystrophy (Kim et al. 2022; Forster et al. 2015; Mah et al. 2014). mRNAs containing these PTCs are subject to quality control by the NMD machinery which can at times improve the disease severity by clearing potentially harmful products, or intensify the disease by removing mRNAs that could make partially functional products (Nguyen et al. 2014; Miller and Pearce 2014).

Mutations in the factors important for quality control have also been implicated in a wide number of diseases. Mutations in the NMD factor Upf3b have been associated with intellectual disorders, autism, and other neuro-developmental disorders (Xu et al. 2013; Laumonnier et al. 2009; Nguyen et al. 2013). Mutations in factors important for trimming, tailing, and maturing non-coding RNAs have similarly been implicated in diseases like retinitis pigmentosa, pontocerebellar hypoplasia, cerebellar atrophy, myeloma, and dyskeratosis congentia (Lardelli et al. 2017; Di Donato et al. 2016; Wan et al. 2012; Boczonadi et al. 2014; Rudnik-Schöneborn et al. 2013; Benyelles et al. 2019; Eggens et al. 2014; Burns et al. 2018; Weißbach et al. 2015; Tummala et al. 2015). Increased information on the mechanisms of these pathways is improving our understanding of the etiology of these diseases, why their presentation can be so variable, and provides new avenues for potential treatment.

In cases where a genetic disease is caused by introduction of a PTC, modulation of NMD could be a viable therapeutic target. If the PTC containing protein is still partially functional, suppressing NMD by inhibiting components of the system or by encouraging read-through of stop codons could have positive effects (Baradaran-Heravi et al. 2016). Furthermore, inhibition of NMD may improve the amount of antigens that make it to the surface of cancer cells (Pastor et al. 2010). For diseases that are caused by mutations in factors that mature or decay ncRNAs, inhibition of an opposing protein may improve aspects of the disease (Shukla et al. 2016, 2020).

It has been shown that knockout of factors important for RQC is detrimental to an organism (Chu et al. 2009). Issues with ribosomal stalling and recycling have also

been shown to contribute to neurodegeneration (Terrey et al. 2020). With the rapidly lowering cost of next generation sequencing and the advent of personalized medicine, we may recognize more diseases that are caused by RQC or that could be improved by modulating the system. Thus, understanding these pathways is of vital importance to human health.

5.2 Outstanding Angel1 questions

The most interesting observation to come out of my Angel1 study is the finding that a cyclic phosphatase is rate-limiting for a decay process and the idea that removal of a cyclic phosphate may be a critical step in 3'–5' RNA decay. While this mechanism makes the most sense given the evidence, I cannot rule out other potential activities that may be responsible for Angel1's role in RQC. If Angel1 were to act in this system by resolving a cyclic phosphate to facilitate 3'–5' decay, we would make two key predictions. First, Angel1 depletion should stabilize the 5' fragment of the RNA while having no effect on the 3' fragment. And second, that Angel1 depletion should increase the proportion of substrate RNAs that terminate in a cyclic phosphate.

Examining stability of the 5' fragment versus the 3' fragment has been technically challenging for several reasons. First, with my BG-NSD substrate, we cannot separately monitor the 5' fragment and the 3' fragment; collision and cleavage likely occur in the poly-A tail which I am unable to probe for. Using an NSD substrate for this purpose therefore does not seem feasible. This should be possible with an

NGD substrate that contains a stalling sequence in the middle of the mRNA; we would be able to probe for unique sequences on either side of the cleavage site. Production of an NGD substrate has been challenging because sequences known to induce NGD (including poly-proline, G-quadruplex, and poly-A insertion) while having been shown to destabilize the nascent polypeptide (Endoh et al. 2013; Huter et al. 2017; Sundaramoorthy et al. 2017), did not have a robust effect on the stability of a β-globin mRNA reporter in the context of human HeLa Tet-off cells in my hands (data not shown).

It might be possible to create an mRNA substrate with a very firm blockage in the body of the coding sequence. Perhaps through chemical blockage that the ribosome is incapable of getting through. This could create a potent NGD substrate that the cell does not have tools to resolve. This substrate would have to be examined in an *in vitro* translation reaction (Witherell 2001) and, if such a reaction reproduced RQC-dependent RNA degradation, fragments could be quantified by Northern blotting with the prediction that Angel1 depletion from translation extracts would stabilize the 5' fragment, while Xrn1 depletion from translation extracts will stabilize the 3' end. It is also possible that NGD surveillance of stall-inducing mRNAs would be potentiated in cell lines lacking ribosome rescue factors such as HBS1, RACK1, or PELO. In that case, it could be easier to monitor dysfunction in decay of the mRNA because the mRNA would be less stable in this background.

Another issue with examining the 5' fragment is the presence of repeated cleavages in the RQC pathway. After cleavage occurs, the 5' fragment is still a substrate for RQC which can lead to more collisions and repeated cleavages along

the body of the transcript (Han et al. 2020). We would predict that this would be exacerbated by depletion of a factor important for decay of the 5' fragment. So, when we attempt to measure levels of the 5' fragment, we may actually be observing levels of numerous ribosome-sized fragments which is much more technically challenging.

One way to examine 5' fragment stability with an NSD substrate in light of this reality could be an adaptation of the 5' Akron-seq protocol (Ibrahim et al. 2018). One could potentially treat RNA with terminator which would remove all 3' RNA fragments produced upon cleavage because they do not contain a 5' cap. Subsequently, one could use long-read sequencing, such as Nanopore sequencing, to examine these 5' capped fragments. Using long reads, one would be able to tell where cleavages occurred and what proportion of species are still polyadenylated, with the prediction that depletion of Angel1 should increase the proportion of deadenylated 5' fragments. There are many technical challenges to overcome with this type of experiment (how to enrich for NSD substrates, dealing with bias toward poly-A species, depth), but an experiment like this could give a very detailed picture of cleavages.

The second prediction that could be pursued is that Angel1 depletion should increase the proportion of 5' mRNA fragments that terminate in 2',3' cyclic phosphates. One way I have attempted to examine this is to use the CycP qPCR protocol described in appendix B. This has also proved to be challenging mainly due to variation in the assay, which could be caused by the fact that any RNA hydrolysis during sample preparation would lead to creation of de novo cyclic phosphates. Another approach may be to use the already established cP-seq protocol (Honda et al. 2016). If one were to sequence with sufficient depth, one may be able to monitor

an NSD reporter mRNA, map where 2',3' cyclic phosphates occur in the body of the mRNA and observe how those cleavages change in response to Angel1 depletion. One may also be able to develop a protocol using a cyclic-phosphate specific RNA ligase (such as RtcB Ligase; NEB, M0458S), which would bypass the need for periodate and PNK treatment of samples.

5.3 Endogenous substrates of ribosome-associated quality control

One fascinating avenue of exploration in the story of RQC is identifying potential endogenous targets of RQC. Depletion of NMD factors leads to dysregulation of as much as 20% of transcripts (Frischmeyer and Dietz 1999). This would seem to support the idea that NMD is not only a quality control pathway, but also a general pathway for transcriptome regulation. Whether the same is true for NSD/NGD is not well understood.

Depletion of RQC factors seems to have a subtle effect on the transcriptome (Tuck et al. 2020; Kalisiak et al. 2017; Weber et al. 2020). Direct endogenous substrates are not well characterized. The one example of a substrate in humans is XBP1, which has been observed at the protein level to be regulated by the RQC factor LTN1 (Han et al. 2020). This difficulty in identifying substrates likely speaks to the stochasticity of RQC in that ribosome collisions happen often across the transcriptome, but they do not always lead to decay. It is possible that there are only few transcripts that are endogenously regulated and RQC is invoked primarily when

something goes wrong with the transcript, for example by truncation, cryptic poly-A site usage, or RNA damage.

There are many interesting questions that can be examined in this context. One may be able to analyze ribosome profiling data, RQC factor depletion gene signatures, disome profiling, and 5’ Akron-seq data to better understand which parts of transcripts are monitored by RQC. Furthermore, RQC may only regulate particular, potentially truncated, isoforms of a transcript, and changes in its levels are diminished in RNA-seq data by the more abundant canonical transcript. Advances in the depth and quality of long-read sequencing will make asking questions about isoform levels and distribution much more tractable.

Lastly, other interesting evidence from my Angel1 study was that Angel1 depletion seemed to specifically upregulate pseudogene RNAs (Figure 3.3). Many pseudogenes are created by re-integration of a spliced transcript back into the genome (Sisu 2021). It stands to reason that some of the pseudogenes may be incorporated back into the genome in a way that would make them substrates of NSD if they become transcribed and translated. Indeed, it has recently been shown through highly sensitive mass spectrometry methods that translation products of some of these pseudogenes exist (Brunet et al. 2020). An interesting avenue for future study would be if these RNAs are targeted by RQC.

5.4 Impact

The work presented in this dissertation advances our understanding of the critical processes of RNA quality control. It also suggests exciting potential directions for the field in the realm of cyclic phosphates and their clearance. My hope is that this work inspires new ideas for future researchers in the Lykke-Andersen lab and elsewhere.

Appendix

A: Angel1 and Angel2 IP-MS/MS raw peptide counts for proteins greater than 5-fold enrichment

Table Appendix A1: Angel1 IP-MS/MS raw peptide counts

Gene	FLAG-Ctrl No RNase Peptide Counts	FLAG-Ctrl RNase treated Peptide Counts	FLAG-Angel1 No RNase Peptide Counts	FLAG-Angel1 RNase Peptide Counts	No RNase Fold Change	RNase Treated Fold Change
ANGEL1	0	0	182	344	184.6687	216.2807
DISC1	0	0	88	125	89.81157	78.98947
EIF4E	0	0	58	56	59.53801	35.73333
NDE1	0	1	52	60	53.48329	19.12047
LARP4	0	0	45	0	46.41946	0.626901
SKIV2L	0	0	36	79	37.33739	50.15204
ANKRD13A	0	0	34	42	35.31916	26.95672
ATXN2	0	0	32	2	33.30092	1.880702
GCN1	0	0	31	89	32.2918	56.42105
GRIPAP1	0	0	31	86	32.2918	54.54035
TTC37	0	0	30	64	31.28268	40.74853
XPO5	0	0	27	25	28.25532	16.29941
CCT7	0	0	22	34	23.20973	21.94152
NDEL1	0	1	19	29	20.18237	9.403508
GRAMD1A	0	0	16	23	17.15502	15.04561
LSM14A	4	4	83	122	16.95319	15.42175
ATP2A2	0	1	15	32	16.1459	10.34386
DNAJB1	0	0	15	18	16.1459	11.91111
CCT3	0	0	14	29	15.13678	18.80702
DDX6	0	0	12	20	13.11854	13.16491
AKAP1	0	0	12	0	13.11854	0.626901
CCT4	2	1	33	28	11.43668	9.090058
HUWE1	1	0	21	55	11.10031	35.10643

Table Appendix A1: Angel1 IP-MS/MS raw peptide counts (cont.)

Gene	FLAG-Ctrl No RNase Peptide Counts	FLAG-Ctrl RNase treated Peptide Counts	FLAG-Angel1 No RNase Peptide Counts	FLAG-Angel1 RNase Peptide Counts	No RNase Fold Change	RNase Treated Fold Change
TRIM24	0	0	10	13	11.10031	8.776607
PLOD1	0	0	1 0	0	11.10031	0.626901
LARP4B	6	0	72	11	10.52367	7.522806
CCT6A	2	0	30	25	10.42756	16.29941
SEH1L	0	0	9	40	10.09119	25.70292
DNAJC7	1	2	19	34	10.09119	7.313839
SLC25A5	0	2	9	12	10.09119	2.716569
LTN1	0	0	8	46	9.082069	29.46432
FANCI	0	0	8	31	9.082069	20.06082
STUB1	0	0	8	23	9.082069	15.04561
RACK1	0	4	8	11	9.082069	1.504561
WDR61	1	0	16	24	8.577509	15.67251
HNRNPUL2	0	14	7	39	8.07295	1.671735
BAG5	1	0	15	25	8.07295	16.29941
TFRC	0	0	7	13	8.07295	8.776607
LRPPRC	2	0	22	5	7.736577	3.761403
UBA52	6	6	50	43	7.352151	3.940518
Ubiquitin	6	6	50	43	7.352151	3.940518
WDR59	0	0	6	34	7.063831	21.94152
SMC2	1	1	13	20	7.063831	6.582456
IRAK1	0	0	6	15	7.063831	10.03041
UBL4A	0	0	6	5	7.063831	3.761403
FUBP3	0	3	6	4	7.063831	0.783626
PPP1CB	0	1	6	3	7.063831	1.253801
PPP1CA	0	2	6	3	7.063831	0.835867
HELZ	0	0	6	1	7.063831	1.253801
STAU2	0	0	6	0	7.063831	0.626901
SLIRP	0	0	6	0	7.063831	0.626901
HSPBP1	0	0	5	27	6.054712	17.55321
WDR24	0	0	5	23	6.054712	15.04561
ARAF	0	0	5	21	6.054712	13.79181
THOC2	0	12	5	17	6.054712	0.868016
COPA	0	0	5	11	6.054712	7.522806
AKAP8	0	0	5	10	6.054712	6.895906
VBP1	0	0	5	9	6.054712	6.269005

Table Appendix A1: Angel1 IP-MS/MS raw peptide counts (cont.)

Gene	FLAG-Ctrl No RNase Peptide Counts	FLAG-Ctrl RNase treated Peptide Counts	FLAG-Angel1 No RNase Peptide Counts	FLAG-Angel1 RNase Peptide Counts	No RNase Fold Change	RNase Treated Fold Change
TRIM33	0	0	5	6	6.054712	4.388304
PAN3	0	0	5	0	6.054712	0.626901
PUM2	0	0	5	0	6.054712	0.626901
ELAVL2	0	0	5	0	6.054712	0.626901
PHGDH	1	3	10	13	5.550153	2.194152
FIP1L1	1	0	10	2	5.550153	1.880702
PUM1	1	0	10	1	5.550153	1.253801
RALY	0	22	4	27	5.045594	0.763183
WDR6	0	0	4	23	5.045594	15.04561
C14orf166	1	14	9	21	5.045594	0.919454
POMGNT2	0	0	4	19	5.045594	12.53801
RAF1	0	0	4	13	5.045594	8.776607
PKP2	0	1	4	11	5.045594	3.761403
GET4	0	0	4	11	5.045594	7.522806
SALL2	0	0	4	9	5.045594	6.269005
MCM7	0	0	4	8	5.045594	5.642105
ATP5C1	0	0	4	6	5.045594	4.388304
SLTM	0	0	4	2	5.045594	1.880702
DPY30	0	1	4	0	5.045594	0.31345
CASC3	0	0	4	0	5.045594	0.626901
S100A9	0	0	4	0	5.045594	0.626901
ELAVL4	0	0	4	0	5.045594	0.626901
FOXK1	0	0	4	0	5.045594	0.626901
LSM12	0	2	4	0	5.045594	0.208967
GATAD2A	0	0	4	0	5.045594	0.626901

Table Appendix A2: Angel2 IP-MS/MS raw peptide counts

Gene	FLAG-Ctrl No RNase Peptide Counts	FLAG-Ctrl RNase treated Peptide Counts	FLAG-Angel1 No RNase Peptide Counts	FLAG-Angel1 RNase Peptide Counts	No RNase Fold Change	RNase Treated Fold Change
ANGEL2	0	0	245	129	180.13	123.96
DDX1	0	3	57	40	42.68	7.31
RTCB	0	0	44	34	33.17	33.41
HNRNPH3	0	5	32	27	24.40	3.24
FAM98A	0	2	27	8	20.74	2.21
C14orf166	0	0	22	21	17.09	21.02
AKAP8L	1	4	45	14	15.80	2.11
HNRNPUL2	0	16	18	65	14.16	2.60
PRMT1	20	6	418	329	12.82	32.40
FAM98B	0	3	16	8	12.70	1.61
SEC16A	6	17	123	21	11.55	0.82
AKAP8	0	0	14	10	11.24	10.53
DDX54	0	0	14	6	11.24	6.72
RBM4	0	0	12	2	9.77	2.91
PRMT8	5	0	86	63	9.49	61.05
SEC13	0	0	11	0	9.04	1.00
TP53BP1	0	4	11	4	9.04	0.71
GTF3C4	0	0	11	0	9.04	1.00
SERBP1	0	6	10	10	8.31	1.08
ADAR	0	6	10	41	8.31	4.13
DEK	0	0	9	0	7.58	1.00
RBM4B	0	0	8	1	6.85	1.95
PRKDC	5	14	61	65	6.78	2.95
SSB	0	0	7	0	6.12	1.00
EMD	0	5	7	4	6.12	0.58
DICER1	0	0	7	0	6.12	1.00
CTR9	0	0	7	0	6.12	1.00
XRN2	3	0	33	7	5.66	7.67
DYNLL2	0	3	6	8	5.39	1.61
HSD17B10	0	0	6	14	5.39	14.34
MCM7	0	0	6	9	5.39	9.58

Table Appendix A2: Angel2 IP-MS/MS raw peptide counts (cont.)

Gene	FLAG-Ctrl No RNase Peptide Counts	FLAG-Ctrl RNase treated Peptide Counts	FLAG-Angel1 No RNase Peptide Counts	FLAG-Angel1 RNase Peptide Counts	No RNase Fold Change	RNase Treated Fold Change
ZC3H4	0	1	6	0	5.39	0.41
MLF2	0	0	6	2	5.39	2.91
ZNF512	0	0	6	0	5.39	1.00
NMNAT1	0	0	6	2	5.39	2.91
MCM2	0	0	6	6	5.39	6.72
CDKN2AIP	0	0	6	0	5.39	1.00
HNRNPUL1	19	0	155	19	5.02	19.11

B: Cyclic Phosphate qPCR protocol

Table Appendix B1: Oligonucleotide sequences from IDT

Name	Sequence	Notes
AR17	ACACGACGCTCTTCCGA	
5'APP-AG_fixed	/5rApp/AGTCGTATCCATAGATCGGAAGAGCGTCGTG	Rnase Free HPLC cleanup
Beta Globin qPCR fwd	AGGAGAAGTCTGCCGTTACTG	
Beta Globin qPCR rev	CAGGGGAAAGCGCGAACGCAGCCGAGCACTTTCTTGCCATGA	
U1 qPCR fwd	GCACTCCGGATGTGCTGACCC	
U1 qPCR rev	CAGGGGAAAGCGCGAACGCAG	
U6 qpCR fwd	GCTTCGGCAGCACATATACTAAAAT	
U6 qPCR rev	CGCTTCACGAATTTGCGTGTCAT	

Table Appendix B2: Materials and reagents

Reagent	Manufacturer	Cat no.
Trizol	Thermo Fisher	15596026
RNA Clean and Concentrator 5	Zymo Research	R1016
Quick CIP	New England Biolabs	M0525L
Sodium Periodate	Sigma Aldrich	311448
T4 PNK	New England Biolabs	M0201L
Rnase Out	Invitrogen	10777019
T4 RNA Ligase 2 Truncated	New England Biolabs	M0242L
Superscript III Reverse Transcriptase	Life Technologies	18080085
Fast SYBR Green Master Mix	Life Technologies	4385617

Protocol; adapted in part from (Honda et al. 2016)

1) RNA isolation
    a. Isolate RNA using standard trizol methods
    b. Take ~10ug of RNA in 40uL of water in a 1.5mL epi
2) DNaseI treatment
    a. To each 40uL sample add:

        i. 5uL cutsmart 10x buffer
        ii. 5uL DNaseI (Included in zymo RNA clean and concentrator kit)
    b. Incubate at 37C for 30min
3) Quick CIP treatment
    a. Make master mix
        i. 1uL cutsmart 10x
        ii. 0.5uL RNaseOut
        iii. 1uL Quick CIP (NEB)
        iv. 7.5uL water
    b. Add 10uL to each sample
    c. Incubate at 37C for 40min
4) Zymo extraction
    a. Follow manufacturer protocol
    b. Elute in 15uL
    c. Split 7uL into two tubes for each sample (for +/- periodate)
    d. Add 38uL to each sample (45uL total)
5) Periodate Treatment
    a. Prepare fresh sodium periodate by dissolving 21.3mg/mL in water
    b. Add 5uL periodate solution to periodate+ samples and 5uL water to periodate- samples
    c. Incubate for 40min at room temperature in the dark
6) Zymo extraction
    a. Follow manufacturer protocols and elute in 11uL of water
    b. Take 10uL into a new tube
7) PNK treatment
    a. Make PNK master mix
        i. 3uL T4 PNK 10x Buffer
        ii. 0.5uL ATP (10mM)
        iii. 0.5uL RNaseOut
        iv. 2uL PNK
        v. 14uL water
    b. Add 20uL of master mix to each sample
    c. Incubate at 37C for 40min
8) Zymo Cleanup
    a. Follow manufacturer procedure and elute in 6uL of water
    b. Move 5uL to a new tube
9) Ligation
    a. Make Ligation master mix
        i. 5uL water
        ii. 2uL T4 RNA ligase buffer
        iii. 4uL 50% PEG8000
        iv. 2uL 5’APP-AG_Fix(20uM)
        v. 2uL RNA Ligase 2 Truncated
    b. Add 15uL to 5uL of sample

c. Incubate for 2 hours at 25C

10)Zymo cleanup
   a. Standard protocol, elute in 12uL of water
   b. Split 5uL into two epi tubes

11)Reverse transcription (Gene-specific)
   a. To the first set of samples add
      i. 1uL AR17 (20uM)
      ii. 1uL dNTPs (10mM each)
      iii. 6uL H2O
   b. Heat to 65C for 5min and cool on ice for >1min
   c. Add to each tube
      i. 4uL 5x First Strand Buffer
      ii. 1uL 0.1M DTT
      iii. 1uL RNaseOut
      iv. 1uL SSIII
   d. Run program
      i. 55C 1hr
      ii. 70C 15min
      iii. 4C for all time

12)Reverse transcription (random hexamer)
   a. To the remaining set of samples add:
      i. 2uL random hexamers (100ng/uL)
      ii. 1uL dnTPs (10mM each)
      iii. 5uL H2O
   b. Heat at 65C for 5 min and cool on ice for >1min
   c. Add to each tube
      i. 4uL 5x First strand buffer
      ii. 1uL 0.1M DTT
      iii. 1uL RNaseOut
      iv. 1uL SSIII
   d. Run program
      i. 25C 5min
      ii. 50C 1hr
      iii. 70C 15min
      iv. 4C until the inevitable heat death of the universe

13)Analysis
   a. Quantify target in gene-specific and random hexamer
   b. Normalize gene-specific Ct to random hexamer Ct to get proportion of cyclic phosphate

References

Abelson J, Trotta CR, Li H. 1998. tRNA Splicing *. *J Biol Chem* **273**: 12685–12688.

Allmang C, Kufel J, Chanfreau G, Mitchell P, Petfalski E, Tollervey D. 1999. Functions of the exosome in rRNA, snoRNA and snRNA synthesis. *EMBO J* **18**: 5399–5410.

Aphasizhev R, Suematsu T, Zhang L, Aphasizheva I. 2016. Constructive edge of uridylation-induced RNA degradation. *RNA Biol*.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202–W208.

Bao C, Loerch S, Ling C, Korostelev AA, Grigorieff N, Ermolenko DN. 2020. mRNA stem-loops can pause the ribosome by hindering A-site tRNA binding. *Elife* **9**: 1–67.

Bar-Peled L, Chantranupong L, Cherniack AD, Chen WW, Ottina KA, Grabiner BC, Spear ED, Carter SL, Meyerson M, Sabatini DM. 2013. A Tumor Suppressor Complex with GAP Activity for the Rag GTPases That Signal Amino Acid Sufficiency to mTORC1. *Science (80-)* **340**: 1100–1106.

Baradaran-Heravi A, Balgi AD, Zimmerman C, Choi K, Shidmoossavee FS, Tan JS, Bergeaud C, Krause A, Flibotte S, Shimizu Y, Anderson HJ, Mouly V, Jan E, Pfeifer T, Jaquith JB, Roberge M. 2016. Novel small molecules potentiate premature termination codon readthrough by aminoglycosides. *Nucleic Acids Res* **44**: 6583–6598.

Bengtson MH, Joazeiro CAP. 2010. Role of a ribosome-associated E3 ubiquitin ligase in protein quality control. *Nature* **467**: 470–473.

Benyelles M, Episkopou H, O'Donohue M, Kermasson L, Frange P, Poulain F, Burcu Belen F, Polat M, Bole-Feysot C, Langa-Vives F, Gleizes P-E, Villartay J, Callebaut I, Decottignies A, Revy P. 2019. Impaired telomere integrity and rRNA biogenesis in PARN-deficient patients and knock-out models. *EMBO Mol Med* **11**.

Berndt H, Harnisch C, Rammelt C, Stöhr N, Zirkel A, Dohm JC, Himmelbauer H, Tavanez JP, Hüttelmaier S, Wahle E. 2012. Maturation of mammalian H/ACA box snoRNAs: PAPD5-dependent adenylation and PARN-dependent trimming. *Rna* **18**: 958–972.

Boczonadi V, Müller JS, Pyle A, Munkley J, Dor T, Quartararo J, Ferrero I, Karcagi V,

Giunta M, Polvikoski T, Birchall D, Princzinger A, Cinnamon Y, Lützkendorf S, Piko H, Reza M, Florez L, Santibanez-Koref M, Griffin H, Schuelke M, Elpeleg O, Kalaydjieva L, Lochmüller H, Elliott DJ, Chinnery PF, Edvardson S, Horvath R. 2014. EXOSC8 mutations alter mRNA metabolism and cause hypomyelination with spinal muscular atrophy and cerebellar hypoplasia. *Nat Commun* **5**.

Brunet MA, Leblanc S, Roucou X. 2020. Reconsidering proteomic diversity with functional investigation of small ORFs and alternative ORFs. *Exp Cell Res* **393**: 112057.

Burns DT, Donkervoort S, Müller JS, Knierim E, Bharucha-Goebel D, Faqeih EA, Bell SK, AlFaifi AY, Monies D, Millan F, Retterer K, Dyack S, MacKay S, Morales-Gonzalez S, Giunta M, Munro B, Hudson G, Scavina M, Baker L, Massini TC, Lek M, Hu Y, Ezzo D, AlKuraya FS, Kang PB, Griffin H, Foley AR, Schuelke M, Horvath R, Bönnemann CG. 2018. Variants in EXOSC9 Disrupt the RNA Exosome and Result in Cerebellar Atrophy with Spinal Motor Neuronopathy. *Am J Hum Genet* **102**: 858–873.

Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg EC, Akeson M, Vollmers C. 2017. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat Commun* **8**.

Cai W, Wei Y, Jarnik M, Reich J, Lilly MA. 2016. The GATOR2 Component Wdr24 Regulates TORC1 Activity and Lysosome Function ed. H. Kramer. *PLoS Genet* **12**: e1006036.

Chan PP, Lowe TM. 2016. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res* **44**: D184–D189.

Chandrasekaran V, Juszkiewicz S, Choi J, Puglisi JD, Brown A, Shao S, Ramakrishnan V, Hegde RS. 2019. Mechanism of ribosome stalling during translation of a poly(A) tail. *Nat Struct Mol Biol* **26**: 1132–1140.

Chang W, Cheng J, Allaire JJ, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B. 2021. shiny: Web Application Framework for R.

Chen B, Retzlaff M, Roos T, Frydman J. 2011. Cellular strategies of protein quality control. *Cold Spring Harb Perspect Biol* **3**: 1–14.

Chu J, Hong NA, Masuda CA, Jenkins B V., Nelms KA, Goodnow CC, Glynne RJ, Wu H, Masliah E, Joazeiro CAPP, Kay SA. 2009. A mouse forward genetics screen identifies LISTERIN as an E3 ubiquitin ligase involved in neurodegeneration. *Proc Natl Acad Sci* **106**: 2097–2103.

Chyzyńska K, Labun K, Jones C, Grellscheid SN, Valen E. 2021. Deep conservation of ribosome stall sites across RNA processing genes. *NAR Genomics Bioinforma* **3**: 1–13.

Clement SL, Lykke-Andersen J. 2008. A Tethering Approach to Study Proteins that Activate mRNA Turnover in Human Cells. *Methods Mol Biol* **419**: 121–133.

D'Orazio KN, Green R. 2021. Ribosome states signal RNA quality control. *Mol Cell* **81**: 1372–1383.

D'Orazio KN, Wu CC-C, Sinha N, Loll-Krippleber R, Brown GW, Green R. 2019. The endonuclease Cue2 cleaves mRNAs at stalled ribosomes during No Go Decay. *Elife* **8**.

Damgaard CK, Lykke-Andersen J. 2011. Translational coregulation of 5'TOP mRNAs by TIA-1 and TIAR. *Genes Dev* **25**: 2057–2068.

Darnell JE, Philipson L, Wall R, Adesnik M. 1971. Polyadenylic acid sequences: Role in conversion of nuclear RNA into messenger RNA. *Science (80-)* **174**: 507–510.

Deutscher MP. 1973. Synthesis and Functions of the -C-C-A Terminus of Transfer RNA. *Prog Nucleic Acid Res Mol Biol* **13**: 51–92.

Di Donato N, Neuhann T, Kahlert AK, Klink B, Hackmann K, Neuhann I, Novotna B, Schallner J, Krause C, Glass IA, Parnell SE, Benet-Pages A, Nissen AM, Berger W, Altmüller J, Thiele H, Weber BHF, Schrock E, Dobyns WB, Bier A, Rump A. 2016. Mutations in EXOSC2 are associated with a novel syndrome characterised by retinitis pigmentosa, progressive hearing loss, premature ageing, short stature, mild intellectual disability and distinctive gestalt. *J Med Genet* **53**: 419–425.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Doma MK, Parker R. 2006. Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. *Nature* **440**: 561–564.

Doma MK, Parker R. 2007. RNA Quality Control in Eukaryotes. *Cell* **131**: 660–668.

Dupasquier M, Kim S, Halkidis K, Gamper H, Hou YM. 2008. tRNA Integrity Is a Prerequisite for Rapid CCA Addition: Implication for Quality Control. *J Mol Biol* **379**: 579–588.

Durand S, Franks TM, Lykke-Andersen J. 2016. Hyperphosphorylation amplifies

UPF1 activity to resolve stalls in nonsense-mediated mRNA decay. *Nat Commun* **7**: 1–12.

Durand S, Lykke-Andersen J. 2013. Nonsense-mediated mRNA decay occurs during eIF4F-dependent translation in human cells. *Nat Struct Mol Biol* **20**: 702–709.

Eddy SR. 2001. Non–coding RNA genes and the modern RNA world. *Nat Rev Genet 2001 212* **2**: 919–929.

Edmonds M, Vaughan MH, Nakazato H. 1971. Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly-labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship. *Proc Natl Acad Sci U S A* **68**: 1336–1340.

Eggens VRC, Barth PG, Niermeijer JMF, Berg JN, Darin N, Dixit A, Fluss J, Foulds N, Fowler D, Hortobágyi T, Jacques T, King MD, Makrythanasis P, Máté A, Nicoll JAR, O'Rourke D, Price S, Williams AN, Wilson L, Suri M, Sztriha L, Dijns-De Wissel MB, Van Meegen MT, Van Ruissen F, Aronica E, Troost D, Majoie CBLM, Marquering HA, Poll-Thé BT, Baas F. 2014. EXOSC3 mutations in pontocerebellar hypoplasia type 1: novel mutations and genotype-phenotype correlations. *Orphanet J Rare Dis* **9**.

Endoh T, Kawasaki Y, Sugimoto N. 2013. Stability of RNA quadruplex in open reading frame determines proteolysis of human estrogen receptor α. *Nucleic Acids Res* **41**: 6222–6231.

Endoh T, Sugimoto N. 2016. Mechanical insights into ribosomal progression overcoming RNA G-quadruplex from periodical translation suppression in cells. *Sci Rep* **6**: 22719.

Ermolaeva MA, Dakhovnik A, Schumacher B. 2015. Quality control mechanisms in cellular and systemic DNA damage responses. *Ageing Res Rev* **23**: 3–11.

Forster L, Ardakani RM, Qadah T, Finlayson J, Ghassemifar R. 2015. The Effect of Nonsense Mediated Decay on Transcriptional Activity Within the Novel β-Thalassemia Mutation HBB: c.129delT. *Hemoglobin* **39**: 334–339.

Franks TM, Singh G, Lykke-Andersen J. 2010. Upf1 ATPase-Dependent mRNP Disassembly Is Required for Completion of Nonsense- Mediated mRNA Decay. *Cell* **143**: 938–950.

Frischmeyer PA, Dietz HC. 1999. Nonsense-Mediated mRNA Decay in Health and Disease. *Hum Mol Genet* **8**: 1893–1900.

Frischmeyer PA, Van Hoof A, O'Donnell K, Guerrerio AL, Parker R, Dietz HC. 2002. An mRNA surveillance mechanism that eliminates transcripts lacking termination

codons. *Science (80-)* **295**: 2258–2261.

Frohman MA, Dush MK, Martin GR. 1988. Rapid production of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. *Proc Natl Acad Sci U S A* **85**: 8998–9002.

Fuchs RT, Sun Z, Zhuang F, Robb GB. 2015. Bias in Ligation-Based Small RNA Sequencing Library Construction Is Determined by Adaptor and RNA Structure. *PLoS One* **10**: e0126049.

Gaidatzis D, Burger L, Florescu M, Stadler MB. 2015. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat Biotechnol 2015 337* **33**: 722–729.

Glover ML, Burroughs AM, Monem PC, Egelhofer TA, Pule MN, Aravind L, Arribere JA. 2020. NONU-1 Encodes a Conserved Endonuclease Required for mRNA Translation Surveillance. *Cell Rep* **30**: 4321-4331.e4.

Goldman DH, Livingston NM, Movsik J, Wu B, Green R. 2021. Live-cell imaging reveals kinetic determinants of quality control triggered by ribosome stalling. *Mol Cell* **81**: 1830-1840.e8.

Goldstrohm AC, Wickens M. 2008. Multifunctional deadenylase complexes diversify mRNA control. *Nat Rev Mol Cell Biol* **9**: 337–344.

Gosselin P, Martineau Y, Morales J, Czjzek M, Glippa V, Gauffeny I, Morin E, Le Corguillé G, Pyronnet S, Cormier P, Cosson B, Le Corguill G, Pyronnet S, Cormier P, Cosson B. 2013. Tracking a refined eIF4E-binding motif reveals Angel1 as a new partner of eIF4E. *Nucleic Acids Res* **41**: 7783–7792.

Gu J, Shumyatsky G, Makan N, Reddy R. 1997. Formation of 2',3'-cyclic phosphates at the 3' end of human U6 small nuclear RNA in vitro. *J Biol Chem* **272**: 21989–21993.

Han P, Shichino Y, Schneider-Poetsch T, Mito M, Hashimoto S, Udagawa T, Kohno K, Yoshida M, Mishima Y, Inada T, Iwasaki S. 2020. Genome-wide Survey of Ribosome Collision. *Cell Rep* **31**: 107610.

Heo I, Ha M, Lim J, Yoon MJ, Park JE, Kwon SC, Chang H, Kim VN. 2012. Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs. *Cell* **151**: 521–532.

Honda S, Morichika K, Kirino Y. 2016. Selective amplification and sequencing of cyclic phosphate–containing RNAs by the cP-RNA-seq method. *Nat Protoc* **11**: 476–489.

Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Ridwan Amode M, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Charkhchi M, Cummins C, da Rin Fioretto L, Davidson C, Dodiya K, El Houdaigui B, Fatima R, Gall A, Giron CG, Grego T, Guijarro-Clarke C, Haggerty L, Hemrom A, Hourlier T, Izuogu OG, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Martinez JG, Marugán JC, Maurel T, McMahon AC, Mohanan S, Moore B, Muffato M, Oheh DN, Paraschas D, Parker A, Parton A, Prosovetskaia I, Sakthivel MP, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Steed E, Szpak M, Szuba M, Taylor K, Thormann A, Threadgold G, Walts B, Winterbottom A, Chakiachvili M, Chaubal A, de Silva N, Flint B, Frankish A, Hunt SE, IIsley GR, Langridge N, Loveland JE, Martin FJ, Mudge JM, Morales J, Perry E, Ruffier M, Tate J, Thybert D, Trevanion SJ, Cunningham F, Yates AD, Zerbino DR, Flicek P. 2021. Ensembl 2021. *Nucleic Acids Res* **49**: D884–D891.

Huter P, Arenz S, Bock L V., Graf M, Frister JO, Heuer A, Peil L, Starosta AL, Wohlgemuth I, Peske F, Nováček J, Berninghausen O, Grubmüller H, Tenson T, Beckmann R, Rodnina M V., Vaiana AC, Wilson DN. 2017. Structural Basis for Polyproline-Mediated Ribosome Stalling and Rescue by the Translation Elongation Factor EF-P. *Mol Cell* **68**: 515-527.e6.

Ibrahim F, Maragkakis M, Alexiou P, Mourelatos Z. 2018. Ribothrypsis, a novel process of canonical mRNA decay, mediates ribosome-phased mRNA endonucleolysis. *Nat Struct Mol Biol* **25**: 302–310.

Ikeuchi K, Tesina P, Matsuo Y, Sugiyama T, Cheng J, Saeki Y, Tanaka K, Becker T, Beckmann R, Inada T. 2019. Collided ribosomes form a unique structural interface to induce Hel2-driven quality control pathways. *EMBO J* **38**: e100276.

Joazeiro CAP. 2017. Ribosomal stalling during translation: Providing substrates for ribosome-associated protein quality control. *Annu Rev Cell Dev Biol* **33**: 343–368.

Jurkin J, Henkel T, Nielsen AF, Minnich M, Popow J, Kaufmann T, Heindl K, Hoffmann T, Busslinger M, Martinez J. 2014. The mammalian tRNA ligase complex mediates splicing of XBP1 mRNA and controls antibody secretion in plasma cells. *EMBO J* **33**: 2922.

Juszkiewicz S, Chandrasekaran V, Lin Z, Kraatz S, Ramakrishnan V, Hegde RS. 2018. ZNF598 Is a Quality Control Sensor of Collided Ribosomes. *Mol Cell* **72**: 469–481.

Kalisiak K, Kuliński TM, Tomecki R, Cysewski D, Pietras Z, Chlebowski A, Kowalska K, Dziembowski A. 2017. A short splicing isoform of HBS1L links the cytoplasmic exosome and SKI complexes in humans. *Nucleic Acids Res* **45**: 2068–2080.

Katoh T, Sakaguchi Y, Miyauchi K, Suzuki T, Suzuki T, Kashiwabara SI, Baba T.

2009. Selective stabilization of mammalian microRNAs by 3' adenylation mediated by the cytoplasmic poly(A) polymerase GLD-2. *Genes Dev* **23**: 433–438.

Kikin O, D'Antonio L, Bagga PS. 2006. QGRS Mapper: A web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res* **34**: W676–W682.

Kim YJ, Sivetz N, Layne J, Voss DM, Yang L, Zhang Q, Krainer AR. 2022. Exon-skipping antisense oligonucleotides for cystic fibrosis therapy. *Proc Natl Acad Sci U S A* **119**.

Kishor A, Fritz SE, Hogg JR. 2019. Nonsense-mediated mRNA decay: the challenge of telling right from wrong in a complex transcriptome. *Wiley Interdiscip Rev RNA* **10**: e1548.

Kowalsky AH, Namkoong S, Mettetal E, Park HW, Kazyken D, Fingar DC, Lee JH. 2020. The GATOR2–mTORC2 axis mediates Sestrin2-induced AKT Ser/Thr kinase activation. *J Biol Chem* **295**: 1769.

Kurzik-Dumke U, Zengerle A. 1996. Identification of a novel Drosophila melanogaster gene, angel, a member of a nested gene cluster at locus 59F4,5. *Biochim Biophys Acta* **1308**: 177–181.

Łabno A, Warkocki Z, Kulínski T, Krawczyk PS, Bijata K, Tomecki R, Dziembowski A. 2016. Perlman syndrome nuclease DIS3L2 controls cytoplasmic non-coding RNAs and provides surveillance pathway for maturing snRNAs. *Nucleic Acids Res* **44**: 10437–10453.

LaCava J, Houseley J, Saveanu C, Petfalski E, Thompson E, Jacquier A, Tollervey D. 2005. RNA Degradation by the Exosome Is Promoted by a Nuclear Polyadenylation Complex. *Cell* **121**: 713–724.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, Fitzhugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, Levine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T,

Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Hong ML, Dubois J, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, De La Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JGR, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin E V., Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran J V., Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AFA, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson M V., Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Patrinos A, Morgan MJ. 2001. Initial sequencing and analysis of the human genome. *Nat 2001 4096822* **409**: 860–921.

Laothamatas I, Gao P, Wickramaratne A, Quintanilla CG, Dino A, Khan CA, Liou J, Green CB. 2020. Spatiotemporal regulation of NADP(H) phosphatase Nocturnin and its role in oxidative stress response. *Proc Natl Acad Sci U S A* **117**: 993–999.

Lardelli RM, Lykke-Andersen J. 2020. Competition between maturation and degradation drives human snRNA 3′ end quality control. *Genes Dev* **34**.

Lardelli RM, Schaffer AE, C Eggens VR, Zaki MS, Grainger S, Sathe S, Van Nostrand EL, Schlachetzki Z, Rosti B, Akizu N, Scott E, Silhavy JL, Dean Heckman L, Ozgur Rosti R, Dikoglu E, Gregor A, Guemez-Gamboa A, Musaev D, Mande R, Widjaja A, Shaw TL, Markmiller S, Marin-Valencia I, Davies JH, de Meirleir L, Kayserili H, Altunoglu U, Louise Freckmann M, Warwick L, Chitayat D, Blaser S, Okay Çağlayan A, Bilguvar K, Per H, Fagerberg C, Christesen HT, Kibaek M, Aldinger KA, Manchester D, Matsumoto N, Muramatsu K, Saitsu H, Shiina M, Ogata K, Foulds N, Dobyns WB, Chi NC, Traver D, Spaccini L, Maria Bova S, Gabriel SB, Gunel M, Maria Valente E, Nassogne M-C, Bennett EJ, Yeo GW, Baas F, Lykke-Andersen J, Gleeson JG. 2017. Biallelic mutations in the 3′ exonuclease TOE1 cause pontocerebellar hypoplasia and uncover a role in

snRNA processing. *Nat Genet* **49**: 457–464.

Laumonnier F, Shoubridge C, Antar C, Nguyen LS, Van Esch H, Kleefstra T, Briault S, Fryns JP, Hamel B, Chelly J, Ropers HH, Ronce N, Blesson S, Moraine C, Gécz J, Raynaud M. 2009. Mutations of the UPF3B gene, which encodes a protein widely expressed in neurons, are associated with nonspecific mental retardation with or without autism. *Mol Psychiatry 2010 157* **15**: 767–776.

Lee M, Choi Y, Kim K, Jin H, Lim J, Nguyen TA, Yang J, Jeong M, Giraldez AJ, Yang H, Patel DJ, Kim VN. 2014. Adenylation of maternally inherited MicroRNAs by wispy. *Mol Cell* **56**: 696–707.

Lee SY, Mendecki J, Brawerman G. 1971. A polynucleotide segment rich in adenylic acid in the rapidly-labeled polyribosomal RNA component of mouse sarcoma 180 ascites cells. *Proc Natl Acad Sci U S A* **68**: 1331–1335.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923–930.

Liu X, Zheng Q, Vrettos N, Maragkakis M, Alexiou P, Gregory BD, Mourelatos Z. 2014. A MicroRNA Precursor Surveillance System in Quality Control of MicroRNA Synthesis. *Mol Cell* **55**: 868–879.

Liudkovska V, Dziembowski A. 2021. Functions and mechanisms of RNA tailing by metazoan terminal nucleotidyltransferases. *Wiley Interdiscip Rev RNA* **12**.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Lund E, Dahlberg JE, Nuclear RNA S, Lund E, Dahlberg JE. 1992. Cyclic 2′,3′-phosphates and nontemplated nucleotides at the 3′ end of spliceosomal U6 small nuclear RNA's. *Science (80-)* **255**: 327–330.

Lykke-Andersen J, Shu M-D, Steitz JA. 2000. Human Upf Proteins Target an mRNA for Nonsense-Mediated Decay When Bound Downstream of a Termination Codon. *Cell* **103**: 1121–1131.

Mah JK, Korngut L, Dykeman J, Day L, Pringsheim T, Jette N. 2014. A systematic review and meta-analysis on the epidemiology of Duchenne and Becker muscular dystrophy. *Neuromuscul Disord* **24**: 482–491.

Maquat LE. 2004. Nonsense-mediated mRNA decay: splicing, translation and mRNP

dynamics. *Nat Rev Mol Cell Biol 2004 52* **5**: 89–99.

Miller JN, Pearce DA. 2014. Nonsense-mediated decay in genetic disease: Friend or foe? *Mutat Res* **762**: 52–64.

Mroczek S, Dziembowski A. 2013. U6 RNA biogenesis and disease association. *Wiley Interdiscip Rev RNA* **4**: 581–592.

Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Pääbo S, Mann M. 2011. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* **7**.

Navickas A, Chamois S, Saint-Fort R, Henri J, Torchet C, Benard L. 2020. No-Go Decay mRNA cleavage in the ribosome exit tunnel produces 5′-OH ends phosphorylated by Trl1. *Nat Commun* **11**: 122.

Newman MA, Mani V, Hammond SM. 2011. Deep sequencing of microRNA precursors reveals extensive 3′ end modification. *RNA* **17**: 1795–1803.

Nguyen D, Grenier St-Sauveur V, Bergeron D, Dupuis-Sandoval F, Scott MSS, Bachand F. 2015. A Polyadenylation-Dependent 3' End Maturation Pathway Is Required for the Synthesis of the Human Telomerase RNA. *Cell Rep* **13**: 2244–2257.

Nguyen LS, Kim HG, Rosenfeld JA, Shen Y, Gusella JF, Lacassie Y, Layman LC, Shaffer LG, Gécz J. 2013. Contribution of copy number variants involving nonsense-mediated mRNA decay pathway genes to neuro-developmental disorders. *Hum Mol Genet* **22**: 1816–1825.

Nguyen LS, Wilkinson MF, Gecz J. 2014. Nonsense-mediated mRNA decay: inter-individual variability and human disease. *Neurosci Biobehav Rev* **46 Pt 2**: 175.

Nicholson AL, Pasquinelli AE. 2019. Tales of Detailed Poly(A) Tails. *Trends Cell Biol* **29**: 191–200.

Pastor F, Kolonias D, Giangrande PH, Gilboa E. 2010. Induction of tumour immunity by targeted inhibition of nonsense-mediated mRNA decay. *Nat 2010 4657295* **465**: 227–230.

Perumal K, Reddy R. 2002. The 3′ end formation in small RNAs. *Gene Expr* **10**: 59–78.

Pinto PH, Kroupova A, Schleiffer A, Mechtler K, Jinek M, Weitzer S, Martinez J. 2020. ANGEL2 is a member of the CCR4 family of deadenylases with 2',3'-cyclic phosphatase activity. *Science* **369**: 524–530.

Pirouz M, Ebrahimi AG, Gregory RI. 2019. Unraveling 3'-end RNA uridylation at

nucleotide resolution. *Methods* **155**: 10–19.

Pisareva VP, Skabkin MA, Hellen CUT, Pestova T V, Pisarev A V. 2011. Dissociation by Pelota, Hbs1 and ABCE1 of mammalian vacant 80S ribosomes and stalled elongation complexes. *EMBO J* **30**: 1804–17.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

R Core Team. 2021. R: A Language and Environment for Statistical Computing.

Rinke J, Steitz JA. 1982. Precursor molecules of both human 5S ribosomal RNA and transfer RNAs are bound by a cellular protein reactive with anti-La Lupus antibodies. *Cell* **29**: 149–159.

Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink JJ, Lopez G, Valencia A, Tress ML. 2013. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res* **41**: D110.

Roy KR, Chanfreau GF. 2020. Robust mapping of polyadenylated and non-polyadenylated RNA 3′ ends at nucleotide resolution by 3′-end sequencing. *Methods* **176**: 4–13.

Rudnik-Schöneborn S, Senderek J, Jen JC, Houge G, Seeman P, Puchmajerová A, Graul-Neumann L, Seidel U, Korinthenberg R, Kirschner J, Seeger J, Ryan MM, Muntoni F, Steinlin M, Sztriha L, Colomer J, Hübner C, Brockmann K, Van Maldergem L, Schiff M, Holzinger A, Barth P, Reardon W, Yourshaw M, Nelson SF, Eggermann T, Zerres K. 2013. Pontocerebellar hypoplasia type 1: clinical spectrum and relevance of EXOSC3 mutations. *Neurology* **80**: 438–446.

Saito S, Hosoda N, Hoshino SI. 2013. The Hbs1-Dom34 protein complex functions in non-stop mRNA decay in mammalian cells. *J Biol Chem* **288**: 17832–17843.

Schaffer AE, Eggens VRC, Caglayan AO, Reuter MS, Scott E, Coufal NG, Silhavy JL, Xue Y, Kayserili H, Yasuno K, Rosti RO, Abdellateef M, Caglar C, Kasher PR, Cazemier JL, Weterman MA, Cantagrel V, Cai N, Zweier C, Altunoglu U, Satkin NB, Aktar F, Tuysuz B, Yalcinkaya C, Caksen H, Bilguvar K, Fu XD, Trotta CR, Gabriel S, Reis A, Gunel M, Baas F, Gleeson JG. 2014. CLP1 founder mutation links tRNA splicing and maturation to cerebellar development and neurodegeneration. *Cell* **157**: 651–663.

Scott DD, Norbury CJ. 2013. RNA decay via 3’ uridylation. *Biochim Biophys Acta - Gene Regul Mech* **1829**: 654–665.

Shcherbik N, Wang M, Lapik YR, Srivastava L, Pestov DG. 2010. Polyadenylation and degradation of incomplete RNA polymerase I transcripts in mammalian cells.

EMBO Rep **11**: 106–11.

Shigematsu M, Kawamura T, Kirino Y. 2018. Generation of 2′,3′-Cyclic Phosphate-Containing RNAs as a Hidden Layer of the Transcriptome. *Front Genet* **9**: 562.

Shigematsu M, Kirino Y. 2020. Oxidative stress enhances the expression of 2′,3′-cyclic phosphate-containing RNAs. *RNA Biol* **17**: 1060–1069.

Shukla S, Jeong HC, Sturgeon CM, Parker R, Batista LFZ. 2020. Chemical inhibition of PAPD5/7 rescues telomerase function and hematopoiesis in dyskeratosis congenita. *Blood Adv* **4**: 2717–2722.

Shukla S, Parker R. 2017. PARN Modulates Y RNA Stability and Its 3′-End Formation. *Mol Cell Biol* **37**.

Shukla S, Schmidt JC, Goldfarb KC, Cech TR, Parker R. 2016. Inhibition of telomerase RNA decay rescues telomerase deficiency caused by dyskerin or PARN defects. *Nat Struct Mol Biol* **23**: 286–292.

Simms CL, Hudson BH, Mosior JW, Rangwala AS, Zaher HS. 2014. An active role for the ribosome in determining the fate of oxidized mRNA. *Cell Rep* **9**: 1256–64.

Simms CL, Yan LL, Zaher HS. 2017. Ribosome Collision Is Critical for Quality Control during No-Go Decay. *Mol Cell* **68**: 361-373.e5.

Singh G, Rebbapragada I, Lykke-Andersen J. 2008. A competition between stimulators and antagonists of Upf complex recruitment governs human nonsense-mediated mRNA decay. *PLoS Biol* **6**: 860–871.

Sinha KM, Gu J, Chen Y, Reddy R. 1998. Adenylation of small RNAs in human cells. Development of a cell-free system for accurate adenylation on the 3’-end of human signal recognition particle RNA. *J Biol Chem* **273**: 6853–6859.

Sisu C. 2021. GENCODE Pseudogenes. *Methods Mol Biol* **2324**: 67–82.

Son A, Park JE, Kim VN. 2018. PARN and TOE1 Constitute a 3′ End Maturation Module for Nuclear Non-coding RNAs. *Cell Rep* **23**: 888–898.

Sundaramoorthy E, Leonard M, Mak R, Liao J, Fulzele A, Bennett EJ. 2017a. ZNF598 and RACK1 Regulate Mammalian Ribosome-Associated Quality Control Function by Mediating Regulatory 40S Ribosomal Ubiquitylation. *Mol Cell* **65**: 751-760.e4.

Suzuki S, Yasuda T, Shiraishi Y, Miyano S, Nagasaki M. 2011. ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics* **12**: S7.

Terrey M, Adamson SI, Gibson AL, Deng T, Ishimura R, Chuang JH, Ackerman SL. 2020. GTPBP1 resolves paused ribosomes to maintain neuronal homeostasis. *Elife* **9**: 1–22.

Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi H. 2022. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci* **31**: 8–22.

Tropea D, Hardingham N, Millar K, Fox K. 2018. Mechanisms underlying the role of DISC1 in synaptic plasticity. *J Physiol* **596**: 2747–2771.

Tsuboi T, Kuroha K, Kudo K, Makino S, Inoue E, Kashima I, Inada T. 2012. Dom34:Hbs1 Plays a General Role in Quality-Control Systems by Dissociation of a Stalled Ribosome at the 3′ End of Aberrant mRNA. *Mol Cell* **46**: 518–529.

Tuck AC, Rankova A, Arpat AB, Liechti LA, Hess D, Iesmantavicius V, Castelo-Szekely V, Gatfield D, Bühler M. 2020. Mammalian RNA Decay Pathways Are Highly Specialized and Widely Linked to Translation. *Mol Cell* **77**: 1222-1236.e13.

Tummala H, Walne A, Collopy L, Cardoso S, De La Fuente J, Lawson S, Powell J, Cooper N, Foster A, Mohammed S, Plagnol V, Vulliamy T, Dokal I. 2015. Poly(A)-specific ribonuclease deficiency impacts telomere biology and causes dyskeratosis congenita. *J Clin Invest* **125**: 2151.

Van Hoof A, Frischmeyer PA, Dietz HC, Parker R. 2002. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science (80-)* **295**: 2262–2264.

Van Nostrand EL, Nguyen TB, Gelboin-Burkhart C, Wang R, Blue SM, Pratt GA, Louie AL, Yeo GW. 2017. Robust, Cost-Effective Profiling of RNA Binding Protein Targets with Single-end Enhanced Crosslinking and Immunoprecipitation (seCLIP). pp. 177–200, Humana Press, New York, NY.

Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, Stanton R, Rigo F, Guttman M, Yeo GW. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**: 508–514.

Van Rossum G, Drake FL. 2019. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian

G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov G V., Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Yuan Wang Z, Wang A, Wang X, Wang J, Wei MH, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu SC, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Lai Cheng M, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Ni Tint N, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander K V., Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Deslattes Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. 2001. The sequence of the human genome. *Science (80-)* **291**: 1304–1351.

Wagih O. 2017. ggseqlogo: A “ggplot2” Extension for Drawing Publication-Ready Sequence Logos.

Wagner E, Clement SL, Lykke-Andersen J. 2007. An unconventional human Ccr4-Caf1 deadenylase complex in nuclear cajal bodies. *Mol Cell Biol* **27**: 1686–1695.

Wallace BD, Berman Z, Mueller GA, Lin Y, Chang T, Andres SN, Wojtaszek JL, Derose EF, Appel CD, London RE, Yan S, Williams RS. APE2 Zf-GRF facilitates

3′-5′ resection of DNA damage following oxidative stress.

Wan J, Yourshaw M, Mamsa H, Rudnik-Schöneborn S, Menezes MP, Hong JE, Leong DW, Senderek J, Salman MS, Chitayat D, Seeman P, Von Moers A, Graul-Neumann L, Kornberg AJ, Castro-Gago M, Sobrido MJ, Sanefuji M, Shieh PB, Salamon N, Kim RC, Vinters H V., Chen Z, Zerres K, Ryan MM, Nelson SF, Jen JC. 2012. Mutations in the RNA exosome component gene EXOSC3 cause pontocerebellar hypoplasia and spinal motor neuron degeneration. *Nat Genet* **44**: 704–708.

Watson JD. 1963. Involvement of RNA in the synthesis of proteins. *Science* **140**: 17–26.

Weber R, Chung MY, Keskeny C, Zinnall U, Landthaler M, Valkov E, Izaurralde E, Igreja C. 2020. 4EHP and GIGYF1/2 Mediate Translation-Coupled Messenger RNA Decay. *Cell Rep* **33**: 108262.

Weißbach S, Langer C, Puppe B, Nedeva T, Bach E, Kull M, Bargou R, Einsele H, Rosenwald A, Knop S, Leich E. 2015. The molecular spectrum and clinical impact of DIS3 mutations in multiple myeloma. *Br J Haematol* **169**: 57–70.

Welch JD, Slevin MK, Tatomer DC, Duronio RJ, Prins J a NF, Marzluff WF. 2015. EnD-Seq and AppEnD : sequencing 3 ′ ends to identify nontemplated tails and degradation intermediates. *Rna* **21**: 1375–1389.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Winkler GS, Balacco DL. 2013. Heterogeneity and complexity within the nuclease module of the Ccr4-Not complex. *Front Genet* **4**: 296.

Witherell G. 2001. In vitro translation using HeLa extract. *Curr Protoc cell Biol* **Chapter 11**.

Wolin SL, Maquat LE. 2019. Cellular RNA surveillance in health and disease. *Science (80-)* **366**: 822–827.

Xu X, Zhang L, Tong P, Xun G, Su W, Xiong Z, Zhu T, Zheng Y, Luo S, Pan Y, Xia K, Hu Z. 2013. Exome sequencing identifies UPF3B as the causative gene for a Chinese non-syndrome mental retardation pedigree. *Clin Genet* **83**: 560–564.

Yang Q, Yu CH, Zhao F, Dang Y, Wu C, Xie P, Sachs MS, Liu Y. 2019. eRF1 mediates codon usage effects on mRNA translation efficiency through premature termination at rare codons. *Nucleic Acids Res* **47**: 9243.

Yang W. 2011. Nucleases: Diversity of structure, function and mechanism. *Q Rev*

Biophys **44**: 1–93.

Yip MCJ, Keszei AFA, Feng Q, Chu V, McKenna MJ, Shao S. 2019. Mechanism for recycling tRNAs on stalled ribosomes. *Nat Struct Mol Biol* **26**: 343–349.

Yu S, Kim VN. 2020. A tale of non-canonical tails: gene regulation by post-transcriptional RNA tailing. *Nat Rev Mol Cell Biol 2020 219* **21**: 542–556.

Zillmann M, Gorovsky MA, Phizicky EM. 1991. Conserved mechanism of tRNA splicing in eukaryotes. *Mol Cell Biol* **11**: 5410.

Zinoviev A, Ayupov RK, Abaeva IS, Hellen CUT, Pestova T V. 2020. Extraction of mRNA from Stalled Ribosomes by the Ski Complex. *Mol Cell* **77**: 1340-1349.e6.