Regulation and Functional Inference of Alternative Polyadenylation

Hsin Wei Tseng

Department of Biochemistry

McGill University

Montreal, Quebec, Canada

October 2023

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Table of Contents

ABSTRACT **6**

RÉSUMÉ **8**

LIST OF FIGURES **10**

LIST OF TABLES **11**

LIST OF ABBREVIATIONS **12**

PREFACE **14**

CONTRIBUTION OF AUTHORS **15**

ORIGINAL CONTRIBUTIONS TO KNOWLEDGE **16**

ACKNOWLEDGEMENTS **17**

CHAPTER 1: INTRODUCTION **19**

1.1 3'-END PROCESSING OF EUKARYOTIC MRNA 20
1.2 POST-TRANSCRIPTIONAL REGULATION BY 3' UNTRANSLATED REGIONS 24
1.2.1 3'UTRs regulate mRNA stability and translation *25*
1.2.2 3'UTRs regulate mRNA localization *28*
1.2.3 3'UTRs mediate protein complex assembly *30*
1.3 INTRODUCTION TO ALTERNATIVE POLYADENYLATION 31
1.3.1 3'UTR-APA isoform characteristics and functional relevance *32*
1.3.2 Upstream region APA characteristics and relevant functions *37*
1.4 REGULATION OF APA 39
1.4.1 APA regulation by CPA factors *39*
1.4.2 APA regulation by transcription *44*
1.4.3 Other regulators of APA *47*
1.5 APA PROFILES AND PHYSIOLOGICAL RELEVANCE 51
1.5.1 Global 3'UTR changes in proliferating and activated cells *51*
1.5.2 Global 3'UTR changes in differentiation and development *52*
1.5.3 APA dysregulation in diseases *54*
1.6 CURRENT METHODS FOR STUDYING APA 57
1.7 THESIS RATIONALE AND OBJECTIVES 61

CHAPTER 2: DISTINCT, OPPOSITE FUNCTIONS FOR CFIM59 AND CFIM68 IN MRNA ALTERNATIVE POLYADENYLATION OF PTEN AND IN THE PI3K/AKT SIGNALLING CASCADE **62**

2.1 ABSTRACT 63
2.2 INTRODUCTION 64
2.3 RESULTS 67
2.3.1 CFIm regulates Pten mRNA and protein expression *67*
2.3.2 CFIm controls Pten mRNA 3'UTR isoform expression *69*
2.3.3 Conservation and species-specific functions of CFIm on human PTEN mRNAs *73*
2.3.4 CFIm subunits and UGUA RNA elements in Pten APA regulation *74*

2.3.5 Distinct and opposing roles for CFIm59 and CFIm68 on global APA *78*
2.3.6 CFIm APA regulation in the PI3K/Akt pathway *82*
2.3.7 Distinct associations of PTEN APA, CFIm59 and CFIm68 across onco-transcriptomes .. *84*
2.4 Discussion 87
2.4.1 Distinct selectivity for CFIm59 and CFIm68 in APA regulation *88*
2.4.2 Breadth of APA regulation in the PI3K/AKT/PTEN axis and cancer *92*
2.5 Data availability 93
2.6 Acknowledgements 93
2.7 Funding 94
2.8 Materials and methods 94
2.8.1 Cell culture and transfection *94*
2.8.2 RNA isolation and Northern blots *95*
2.8.3 Pten 3'UTR isoform-specific RT-qPCR *95*
2.8.4 Reporter constructs *96*
2.8.5 Western blotting *96*
2.8.6 3'UTR-seq and data processing *96*
2.8.7 Polysome profiling *97*
2.8.8 mRNA stability assay *98*
2.8.9 Generation of KO cells *98*
2.8.10 Analysis of PAR-CLIP and motif enrichment *99*
2.8.11 Cancer APA analysis *99*
2.8.12 Statistical analyses *99*

CHAPTER 3: TRANSCRIPTOME-WIDE ASSESSMENT FOR THE IMPACT OF 3'UTR SHORTENING ON MRNA STABILITY 100

3.1 Abstract 101
3.2 Introduction 102
3.3 Results 105
3.3.1 3'UTR Decay-seq recapitulates transcriptome-wide mRNA stability *105*
3.3.2 CFIm68 depletion induces a global 3'UTR shortening that does not correlate with the average changes in TU stability *107*
3.3.3 Condition-dependent correlation between 3'UTR length and transcript stability *109*
3.3.4 The impact of CFIm68 knockout on TU stability depends on 3'UTR-specific features . *111*
3.3.5 A possible role for CFIm68 in the regulation of mRNA stability through EJC/NMD *115*
3.4 Discussion 118
3.4.1 Unique sequence features of 3'UTRs govern transcript stability *119*
3.4.2 Role of CFIm68 in mRNA turnover *121*
3.5 Acknowledgements 122
3.6 Funding 122
3.7 Materials and methods 122
3.7.1 3'UTR-seq sample preparation and sequencing *122*
3.7.2 3'UTR-seq data processing *123*
3.7.2.1 Pre-processing, read alignment, and counting features 123
3.7.2.3 Differential transcript stability 124

3.7.2.4 Differential gene expression 124
3.7.3 Characterization of aUTR 125
3.7.4 Statistical analysis and data visualization 125

CHAPTER 4: TRACKING APA THROUGH METASTASIS AT SINGLE-CELL RESOLUTION 126

4.1 Abstract 127
4.2 Introduction 128
4.3 Results 131
4.3.1 APA better resolves tumor origin than gene expression 131
4.3.2 3'UTRs are shortened from primary to metastatic tumors 134
4.3.3 3'UTR length positively correlates with proliferation 136
4.3.4 A common super-cluster characterizes a putative quiescent cancer stem cell population 137
4.4 Discussion 142
4.4.1 Single-cell APA analyses provide unique functional insights 142
4.4.2 A unique APA profile in a putative quiescent cancer stem cell population 144
4.5 Acknowledgements 147
4.6 Funding 147
4.7 Materials and methods 148
4.7.1 Patient-derived xenograft model of ER+ breast cancer 148
4.7.2 Single cell dissociation and sequencing 148
4.7.3 Single-cell RNA-seq data processing and gene expression analyses 149
4.7.4 APA analyses in single-cell RNA-seq 150

CHAPTER 5: GENERAL DISCUSSION 151

5.1 How are context-specific global APA patterns achieved? 152
5.2 Impact of APA on expression beyond regulation of mRNA stability 154
5.3 Genomic approaches in APA studies 156
5.4 Cell adaptation in experimental and cancer contexts 159
5.5 Conclusion 162

REFERENCES 163

APPENDIX 1: SUPPLEMENTAL INFORMATION TO CHAPTER 2 197

APPENDIX 2: SUPPLEMENTAL INFORMATION TO CHAPTER 3 208

APPENDIX 3: SUPPLEMENTAL INFORMATION TO CHAPTER 4 212

Abstract

More than half of mammalian genes express alternative polyadenylation (APA) transcript isoforms that differ solely in the 3' untranslated regions (3'UTRs) by utilizing alternative poly(A) signals (PASs). This thesis investigates the regulation of APA, as well as its impact on transcript stability and expression across mammalian cell lines and mouse cancer model systems. The functions of 3'UTRs arise from the dynamic interplay between their encoded *cis*-regulatory elements and a diverse network of RNA-binding proteins, microRNAs, and their effectors. Inclusion and exclusion of *cis*-regulatory elements through APA can thus have a broad impact on post-transcriptional regulation of mRNA stability, translatability, and localization. APA patterns, or the relative expression of 3'UTR isoforms, are highly specific to cell types and physiological states. Disruption of critical APA regulators, such as the CFIm complex, can skew these expression patterns and cause developmental defects as well as lead to increased tumorigenicity. While several APA regulators have been identified in depletion and mutation screens, most of the associated targets, specific functions, as well as their regulatory impact on global transcript stability remain elusive. Moreover, although APA dysregulation is widespread in cancer, how it contributes to cancer cell identities and the disease progression is still unclear.

With a combination of biochemical assays and bioinformatics analyses, in Chapter 2, we identified CFIm as a direct APA regulator of the dosage-sensitive tumor suppressive gene Pten. We uncovered the differential regulation of Pten by the two subunits of CFIm, CFIm59 (also known as CPSF7) and CFIm68 (also known as CPSF6), and further revealed their widespread and opposing functions in APA regulation transcriptome wide. CFIm depletion also resulted in distinct APA patterns for numerous genes in the PI3K/Akt pathway that Pten antagonizes. Analyses of APA in PTEN-driven cancers showed that APA dysregulation is recurrent in this pathway. This chapter

reveals the distinct target selectivity of CFIm59 and CFIm68, and the breadth and complexity of APA regulation by CFIm.

In Chapter 3, we leveraged CFIm68 knockout-induced global 3'UTR shortening to study the impact of APA dysregulation on transcript stability. Surprisingly, 3'UTR shortening significantly affected the stability and expression for a limited subset of transcripts, while at the transcriptome level it showed no correlation with the average change in transcript stability, highlighting the gene-specific impact of APA. We identified different 3'UTR sequences and structural features that distinguish this subset of mRNAs in their response to 3'UTR shortening. Stability changes upon CFIm68 knockout were also not limited to transcripts expressing multiple 3'UTR isoforms. Lastly, we discovered evidence for a novel function of CFIm68 as a regulator of transcript stability through the exon junction complex (EJC)/nonsense-mediated decay (NMD) axis.

While overall shortening of the 3'UTR in tumor tissues is well documented, few studies reported on the nature and extent of APA shifts in metastasis. In Chapter 4, we characterized APA dynamics in heterogeneous cell populations that emerge during the metastatic process. For this, we performed the first longitudinal study of APA from primary to metastatic tumors at single cell resolution using a well-established breast cancer patient-derived xenograft model. We revealed that APA signatures better distinguish tumor origins than gene expression signatures. 3'UTRs were overall shorter in the metastases compared to the primary tumors. Surprisingly, this shortening correlated with a decreased proliferation signature. Lastly, we identified a cell population expressing an atypical APA pattern with quiescent cancer stem cell-like signatures, further demonstrating the potential for cell type discovery and stratification through APA analyses.

The results presented in this thesis expand our understanding for the regulation of APA and provide a new perspective on APA functions in cellular signaling, cancer, and metastasis.

Résumé

Plus de la moitié des gènes de mammifères expriment des isoformes de polyadénylation alternative (APA) qui diffèrent uniquement dans les régions 3' non traduites (3'UTR), et qui sont dérivés en utilisant des signaux poly(A) (PAS) alternatifs. Cette thèse étudie la régulation de l'APA, ainsi que son impact sur la stabilité et l'expression des transcrits dans les lignées cellulaires de mammifères et les systèmes de modèles de cancer chez la souris. Les fonctions des 3'UTR découlent de l'interaction dynamique entre les éléments cis-régulateurs codés et un réseau divers de protéines de liaison à l'ARN, de microARN et de leurs effecteurs. L'inclusion ou l'exclusion d'éléments cis-régulateurs par le biais de l'APA peut donc avoir un impact important sur la régulation post-transcriptionnelle de la stabilité, de la traductibilité et de la localisation des ARNm. Les profils d'APA, ou l'expression relative des isoformes 3'UTR, sont très sensibles et spécifiques aux types de cellules et à leurs états physiologiques. La perturbation de régulateurs de l'APA, tels que le complexe CFIm, peut transformer ces profils d'expression et entraîner des défauts de développement ainsi qu'une augmentation de la tumorigénicité. Bien que plusieurs régulateurs de l'APA aient été identifiés lors de criblages génétiques et de mutation, la plupart des ARNm cibles associés, les fonctions spécifiques de ces régulateurs, ainsi que l'impact sur la stabilité globale des transcrits restent sous-explorés. En outre, bien que la dérégulation de l'APA soit très répandue dans le cancer, la manière dont elle contribue à l'identité et au comportement des cellules cancéreuses et donc à la progression de la maladie n'est toujours pas claire.

Grâce à une combinaison d'essais biochimiques et d'analyses bio-informatiques, nous avons identifié, au chapitre 2, CFIm comme un régulateur direct de l'APA du gène suppresseur de tumeur Pten, dont le dosage est précisément régulé. Nous avons découvert la régulation différentielle de Pten par les deux sous-unités de CFIm, CFIm59 et CFIm68, et avons révélé leurs fonctions

étendues et opposées dans la régulation de l'APA à l'échelle du transcriptome. La déplétion de CFIm a également entraîné des profils d'APA distincts pour de nombreux gènes de la voie PI3K/Akt, que Pten antagonise. Les analyses de l'APA dans les cancers induits par PTEN ont montré que la dysrégulation de l'APA est récurrente dans cette voie. Ce chapitre révèle la sélectivité et les fonctions distincte de CFIm59 et CFIm68, ainsi que l'étendue et la complexité de la régulation de l'APA par CFIm. Enfin, nous avons découvert des preuves d'une nouvelle fonction de CFIm68 en tant que régulateur de la stabilité d'ARNm par l'intermédiaire du complexe de jonction d'exon (EJC)/*Non-sens mediated decay* (NMD).

Bien que le raccourcissement global des 3'UTR dans les tissus tumoraux soit bien documenté, peu d'études ont rapporté la nature et l'étendue des changements de l'APA dans les métastases. Dans le chapitre 4, nous avons caractérisé la dynamique de l'APA dans des populations cellulaires hétérogènes qui émergent au cours du processus métastatique. Pour ce faire, nous avons réalisé la première étude longitudinale de l'APA des tumeurs primaires et tumeurs métastatiques à la résolution de la cellule unique, en utilisant un modèle de xénogreffe bien établi dérivé de patientes atteintes de cancer du sein. Nous avons révélé que les profils de l'APA distinguent mieux les origines des tumeurs que les signatures d'expression génique. En général, les 3'UTR étaient globalement plus courts dans les métastases que dans les tumeurs primaires. Or, de manière surprenante, ce raccourcissement est corrélé à une signature de prolifération réduite. Enfin, nous avons identifié une population de cellules exprimant un profil APA atypique avec des signatures de type cellules souches cancéreuses quiescentes, démontrant ainsi le potentiel de découverte et de stratification des types cellulaires grâce aux analyses APA.

Les résultats présentés dans cette thèse élargissent notre compréhension de la régulation de l'APA et offrent une nouvelle perspective sur les fonctions de l'APA dans la signalisation cellulaire,

le cancer et les métastases.

List of figures

Figure 1.1: The cleavage and polyadenylation machinery. 22

Figure 1.2: Types of APA. 36

Figure 2.1: CFIm regulates *Pten* mRNA and protein expression.....68

Figure 2.2: CFIm controls *Pten* mRNA 3'UTR isoform expression. 73

Figure 2.3: CFIm subunits and UGUA RNA elements in *Pten* APA regulation.....77

Figure 2.4: Distinct and opposing roles for CFIm59 and CFIm68 on global APA. 82

Figure 2.5: CFIm APA regulation in the PI3K/Akt pathway. 83

Figure 2.6: PTEN APA in cancers. 86

Figure 2.7: Model of *Pten* APA regulation by CFIm. 91

Figure 3.1: 3'UTR Decay-seq recapitulates transcriptome-wide mRNA stability. 106

Figure 3.2: CFIm68 depletion induces a global 3'UTR shortening that does not correlate with the average changes in TU stability. 108

Figure 3.3: CFIm68 KO changes mRNA expression for a limited subset of TUs depending on its impact on TU stability. 109

Figure 3.4: Condition-dependent correlation between 3'UTR length and transcript stability. 110

Figure 3.5: The impact of CFIm68 knockout on TU stability depends on 3'UTR-specific features. 114

Figure 3.6: A possible role for CFIm68 in the regulation of mRNA stability through EJC/NMD. 118

Figure 3.7: Model for the regulation of mRNAs by CFIm68. 118

Figure 4.1: APA better resolves tumor origin than gene expression. 133

Figure 4.2: 3'UTRs are shortened from primary to metastatic tumors. 135

Figure 4.3: 3'UTR length positively correlates with proliferation.....137

Figure 4.4: A common super-cluster characterizes a putative quiescent cancer stem cell population. 141

Figure A1.1: Total RNA-seq analyses of NIH3T3 CFIm KD. 198

Figure A1.2: Pten mRNA translatability and stability upon CFIm59 and -68 depletion. 200

Figure A1.3: Human PTEN APA upon CFIm59 and CFIm68 KO. 200

Figure A1.4: Pten swap constructs expression in CFIm59 and CFIm68 KO cells. 201
Figure A1.5: PDUI analyses of PI3K subunits in cancer. 202
Figure A1.6: CFIm59 and -68 PAR-CLIP on APA genes and motif enrichment surrounding responsive PAS. 203
Figure A2.1: CFIm68 KO results in global 3'UTR shortening. 209
Figure A2.2: Changes in TU stability are not correlated with the magnitude nor the direction of APA change in response to CFIm68 KO. 209
Figure A2.3: The distribution of 3'UTR lengths in wildtype cells. 210
Figure A2.4: aUTR lengths correlate with the magnitude of APA shift for lengthened and shortened TUs in response to CFIm68 KO. 210
Figure A2.5: CFIm68 KO affects mRNA expression of genes in different mRNA decay pathways. 211
Figure A3.1: Cells passing quality control. 213
Figure A3.2: Gene set enrichment analysis for all gene expression clusters. 214
Figure A3.3: Gene set enrichment analysis for mRNAs that are significantly shortened in the met-specific super-cluster compared to the pri-specific super-cluster. 215
Figure A3.4: Proliferation-correlated signature scores significantly correlate with meta-PCNA scores. 216
Figure A3.5: Overlaps between gene expression clusters and PDUI (super-)clusters. 217
Figure A3.6: mRNA expression of cleavage and polyadenylation machinery components for cells in each gene expression cluster. 218
Figure A3.7: mRNA expression of stem cell markers for cells in each gene expression cluster. 219

List of tables

Table 1.1: Oligos used. 205

List of abbreviations

A ..... adenosine
APA ..... alternative polyadenylation
ARE ..... AU-rich element
bp ..... base pair
C ..... cytosine
CDS ..... coding sequence
CFI (CFIm) ..... (mammalian) cleavage factor I
CFII (CFIIm) ..... (mammalian) cleavage factor II
CPA ..... cleavage and polyadenylation
CPSF ..... cleavage and polyadenylation stimulation factor
CRISPR ..... clustered regularly interspersed short palindromic repeats
CSTF ..... cleavage stimulation factor
CTD ..... C-terminal domain
DNA ..... deoxyribonucleic acid
G ..... guanosine
GSEA ..... gene set enrichment analysis
KD ..... knockdown
KO ..... knock out
lncRNA ..... long non-coding RNA
mRNA ..... messenger RNA
mRNP ..... mRNA nucleoprotein complex
miRNA ..... microRNA
miRISC ..... miRNA-induced silencing complex
nt ..... nucleotide
ORF ..... open reading frame
PAP ..... poly(A) polymerase
PAS ..... polyadenylation signal
PCR ..... polymerase chain reaction
PDUI ..... percentage of distal poly(A) site usage index
PDX ..... patient derived xenograft

PI3K ..... phosphoinositide 3-kinase
Pol II..... RNA polymerase II
Pre-mRNA ..... precursor mRNA
PTEN..... phosphatase and tensin homolog
RT-qPCR ..... real-time quantitative PCR
RED..... relative expression difference
RISC..... RNA-induced silencing complex
RBP ..... RNA binding protein
RNA ..... ribonucleic acid
RNA-seq ..... RNA sequencing
RNP..... ribonucleoprotein
scRNA-seq ..... single-cell RNA sequencing
siRNA ..... small interfering RNA
snRNP ..... small nuclear RNP
T ..... thymine
TU ..... transcription unit
U..... uracil
UR-APA ..... upstream region APA
UTR..... untranslated region
WT ..... wildtype

Preface

In compliance with the guidelines for preparing a doctoral thesis at McGill University, this thesis is manuscript-based. It consists of one published article and two manuscripts in preparation.

Chapter 1: General introduction and literature review

Chapter 2: Manuscript published

Hsin-Wei Tseng, Anthony Mota-Sydor, Rania Leventis, Predrag Jovanovic, Ivan Topisirovic, Thomas F. Duchaine. Distinct, opposite functions for CFIm59 and CFIm68 in mRNA alternative polyadenylation of Pten and in the PI3K/Akt signalling cascade. *Nucleic Acids Res*. Sep 9, 2022; doi: 10.1093/nar/gkac704

Chapter 3: Manuscript in preparation

Hsin-Wei Tseng, Thomas F. Duchaine. Transcriptome-wide assessment for the impact of 3'UTR shortening on mRNA stability.

Chapter 4: Manuscript in preparation

Hsin-Wei Tseng, Phuong U. Le, Rima Ezzeddine, Charlotte Girondel, Marco Biondini, Matthew Dankner, Kevin Petrecca, Peter M. Siegel, Thomas F. Duchaine. Tracking APA through metastasis at single-cell resolution.

Chapter 5: General discussion

Contribution of authors

Chapter 2: I performed all the experiments and analyses except; Antony Mota-Sydor performed CFIm knockdown experiments and cloning of Pten UGUA mutation constructs; Rania Leventis developed the CFIm knockout NIH3T3 cells; Predrag Jovanovic from the Ivan Topisirovic lab performed the polysome profiling experiments; Thomas F. Duchaine and I designed the study and wrote the manuscript.

Chapter 3: I performed all the experiments and analyses. Thomas F. Duchaine and I designed the study and wrote the manuscript.

Chapter 4: I performed tumor tissue dissociation and all the data analyses; Phuong U. Le from the Kevin Petrecca lab performed tumor tissue dissociation; Rima Ezzeddine, Charlotte Girondel, and Marco Biondini handled the PDX mice and performed tumor resection; Matthew Dankner and Peter M. Siegel selected the PDX model of choice and contributed to the experimental design; Thomas F. Duchaine and I designed the study and wrote the manuscript.

Original contributions to knowledge

Chapter 2: Distinct, opposite functions for CFIm59 and CFIm68 in mRNA alternative polyadenylation of Pten and in the PI3K/Akt signalling cascade.

- Identified CFIm as a direct APA regulator of Pten.
- Revealed the opposing functions for CFIm59 and CFIm68 in APA regulation of Pten and transcriptome-wide.
- Uncovered widespread regulation of APA in the PI3K/Akt signalling pathway by CFIm.

Chapter 3: Transcriptome-wide assessment for the impact of 3'UTR shortening on mRNA stability.

- Challenged common assumptions on the relationship between 3'UTR length and mRNA stability.
- Revealed a lack of correlation between 3'UTR shortening and changes in transcript stability when averaged across the transcriptome.
- Uncovered a limited subset of transcription units expressing 3'UTR isoforms of significantly different stability.
- Uncovered a novel putative role of CFIm68 as a regulator of mRNA stability through the EJC/NMD pathway.

Chapter 4: Tracking APA through metastasis at single-cell resolution.

- Performed the first single-cell APA analysis on matched primary and metastatic tumors.
- Revealed that APA profiles resolve tumor origins better than gene expression profiles.
- Identified 3'UTR shortening from primary to metastatic tumors.
- Uncovered a surprising positive correlation between 3'UTR length and proliferation signatures through metastasis.
- Identified a quiescent stem-like cell population with a unique APA expression profile.

Acknowledgements

To my supervisor, Dr. Thomas Duchaine, thank you for guiding and supporting me through all these years ever since I was just an undergraduate student. You have always led by example, as an enthusiastic and ambitious scientist and a patient mentor (who is also an excellent martini mixologist). I am incredibly grateful and consider myself fortunate to have the opportunity to work with and learn from you.

To my research advisory committee members, Dr. Ian Watson and Dr. Luke McCaffrey, thank you for the continuous support and invaluable feedback on the projects through these years.

I am grateful for the funding sources that have supported me throughout my graduate studies: Canadian Institutes of Health Research, the Fonds de recherche du Québec – Santé, the Canderel studentship, and Ms. Georgette Duchaine.

I am grateful to have had the opportunity to work with many collaborators. To Dr. Walter Gotlieb, Dr. Anne-Marie Mes-Masson, Dr. Luke McCaffrey, Dr. Hamed Najafabadi, Dr. Peter Siegel, Dr. Ivan Topisirovic, Dr. Keven Petrecca, and your students and staff, thank you for your generous commitment of time and resources through these collaborations over the years.

To all the past and present lab members, thank you for making the lab a fun place to work. Caroline Thivierge, thank you for teaching me the fundamental scientific techniques and principles that I relied on throughout my graduate studies. Rania, thank you for your efficient takeover of the genotyping and other experiments. You are a lifesaver. Elva, Maxime, Vinay, and Chehrona, thank you for your friendship, all the stimulating scientific discussions, and commiserations during frustrating times. You are awesome. Anthony, thank you for your genuine friendship and for being the greatest student I will ever have. Alexandra, Damian, Sylvain, Armin, thank you for all the support and fond memories.

To my past mentor, Dr. Asmahan Abu-Arish, thank you for your patient guidance and for showing me your genuine passion for science which has inspired me to pursue research.

To all my friends outside of the lab, thank you for being the greatest support system one can ask for. Camile, Stephanie, Brittany, Shane, Jeff, and Brian, I cherish all the memories and fun times we share. Elva and Maxime, thank you for all the love you have shown me, especially for supporting me through the last stretch of my graduate studies. I could not have done it without you.

Lastly, I need to thank my loving parents, Kuei-Fang Shen and Ming-Tang Tseng, who supported and continue to support me in their own ways. I am grateful for the sacrifices you have made to allow me to pursue the path I have chosen. This thesis is dedicated to you.

Chapter 1:

Introduction

1.1 3’-end processing of eukaryotic mRNA

Genetic information is stored in DNA and transcribed to messenger RNAs (mRNAs) to be translated into proteins (Crick, 1970). Newly synthesized precursor mRNA (pre-mRNA) by eukaryotic RNA polymerase II (Pol II) undergoes maturation processes of 5’ capping with a 7-methylguanosine cap, intron removal by the splicing machinery, and 3’-end processing (Hocine *et al.*, 2010). All three steps must be completed for proper nuclear export and subsequent translation of the mature mRNA in the cytoplasm. Pre-mRNA processing is thus crucial for the production and function of protein-coding transcripts. Differential regulation of pre-mRNA processing further expands the diversity of mRNA transcripts and protein products generated from a gene, and shapes the post-transcriptional control of gene expression (Tian and Manley, 2017).

Pre-mRNA 3’-end processing involves a two-step process: co-transcriptional endonucleolytic cleavage of the nascent transcript at a specific site, followed by the synthesis of a poly(A) tail, a process known as polyadenylation, at the 3’-terminus of the cleaved transcript. The architecture of a mature mRNA is composed of a protein-coding sequence (CDS) flanked by untranslated regions (UTRs) at both 5’ and 3’ ends. The cleavage site, also known as the polyadenylation site, thus defines the 3’-end of a mature mRNA transcript, as well as the sequence and length of the 3’UTR. Sequences in the 3’UTRs are critical for regulating mRNA stability, translation efficiency, and localization, as well as the corresponding protein localization and functions (Tian and Manley, 2017). The process of cleavage and polyadenylation (CPA) is also closely associated with different co-transcriptional activities including splicing and transcription termination (Richard and Manley, 2009).

The general mechanism of CPA is highly conserved across eukaryotes (Boreikaitė and Passmore, 2023). In mammals, it is accomplished by ~20 core proteins of the CPA machinery

comprised of several multiprotein complexes and associated factors (Shi *et al.*, 2009) (Figure 1.1). Many of these proteins are also brought together through physical interactions with the C-terminal domain (CTD) of Pol II (Boreikaitė and Passmore, 2023). The specificity and efficiency of 3'-end processing are then determined by the binding of CPA factors to sequence elements encoded at the 3'-end of mRNAs. The core sequence element marking the location for 3'-end processing is the polyadenylation signal (PAS), which is canonically AAUAAA but has many other less efficiently utilized non-canonical variants (Beaudoing *et al.*, 2000; Chen *et al.*, 1995; Proudfoot and Brownlee, 1976). The PAS is recognized by one of the core complexes of 3'-end processing, the CPA specificity factor (CPSF), comprised of CPSF160, CPSF100, CPSF73, CPSF30, FIP1, and WDR33. More specifically, the PAS is directly bound by CPSF30 and WDR33 (Chan *et al.*, 2014). Recognition of the PAS is further facilitated by FIP1 binding to the U-rich sequences upstream of PAS (Kaufmann *et al.*, 2004; Lackford *et al.*, 2014). Upon PAS recognition, the transcript 3'-end is subsequently cleaved by the endonuclease CPSF73 typically ~15-30-nt downstream of the PAS, preferentially immediately 3' to a CA dinucleotide (Mandel *et al.*, 2006; Proudfoot, 2011; Proudfoot and Brownlee, 1976).

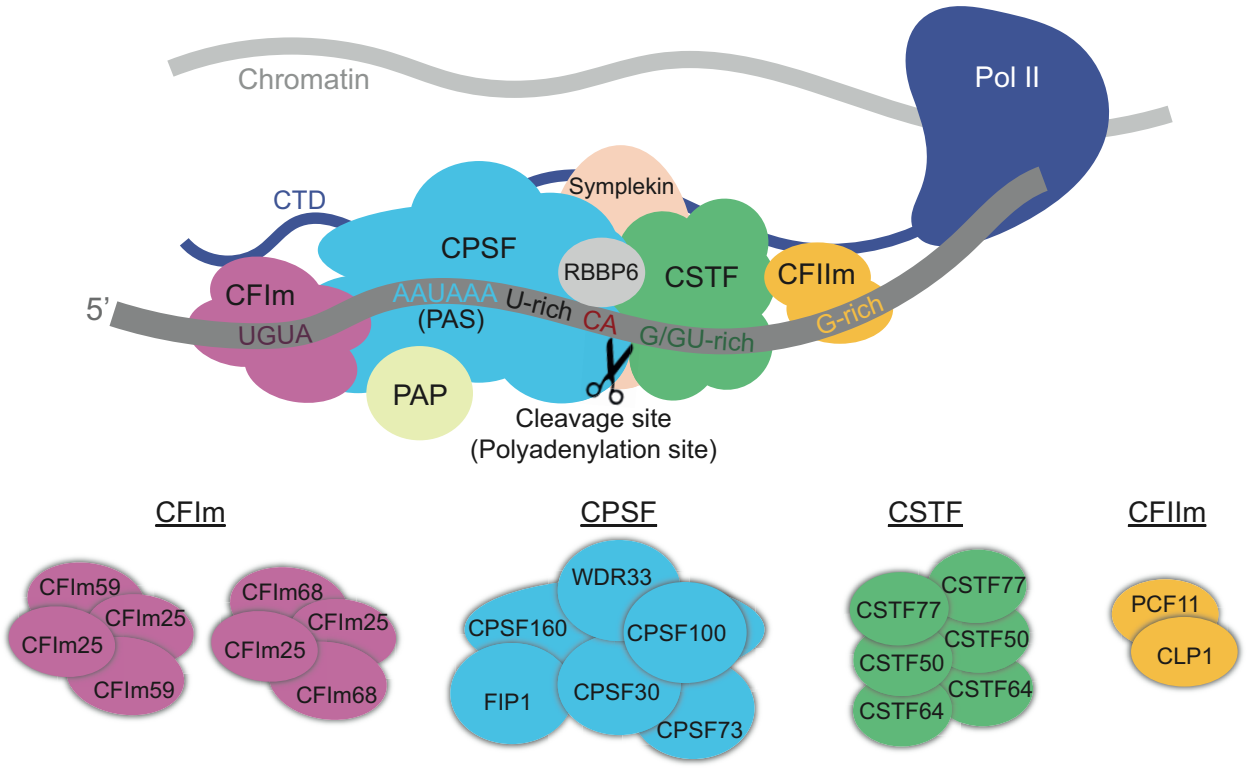


Figure 1.1: The cleavage and polyadenylation machinery. During transcription by RNA polymerase II (Pol II), the cleavage and polyadenylation (CPA) machinery assemble on the 3'-end of nascent transcripts. The sequence elements bound by different components of the CPA machinery are centered around the AAUAAA polyadenylation signal (PAS). The different CPA protein complexes are color-coded, and individual members of each complex are listed at the bottom. The sequence elements share the same color-coding as the CPA complex that binds to them. CPSF, cleavage and polyadenylation specificity factor; CSTF, cleavage stimulation factor; CFIm, mammalian cleavage factor I; CFIIm, mammalian cleavage factor II; RBBP6, RB-binding protein 6; PAP, poly(A) polymerase; WDR33, WD repeat domain 33; FIP1, factor interacting with PAPOLA and CPSF1; CLP1, cleavage factor polyribonucleotide kinase subunit 1.

Once the transcript is cleaved, polyadenylation is carried out by the poly(A) polymerase (PAP), whose activity is also promoted by FIP1 through direct physical interaction (Kaufmann *et*

al., 2004; Kumar *et al.*, 2021). Interestingly, *in vitro* reconstitution of a minimal 3'-end cleavage complex suggests that PAP may also be essential for the cleavage of some transcripts, but not for all (Boreikaite *et al.*, 2022; Schmidt *et al.*, 2022; Takagaki *et al.*, 1989; Terns and Jacob, 1989). The processivity of PAP and the length of the poly(A) tail, typically ~250 adenosines for humans, are further regulated by the nuclear poly(A)-binding protein (PABPN1, also known as PABP2) (Kühn *et al.*, 2009). The addition of a poly(A) tail and proper control of its length confer the nascent transcript resistance to degradation, as well as ensure nuclear export and efficient translation of the mRNA (Passmore and Coller, 2022).

Besides CPSF and PAP, 3'-end cleavage requires additional components of the CPA machinery. One of the required factors is the cleavage specificity factor (CSTF) complex, which forms a heterohexamer consisting of dimers of CSTF77, CSTF64 and/or its paralog CSTF64τ, and CSTF50 (Yang *et al.*, 2018). CSTF50 acts to stabilize the complex while CSTF64 specifically binds U- and GU-rich sequence elements downstream of the cleavage site (Chen and Wilusz, 1998; Takagaki and Manley, 1997). CSTF also interacts with components of CPSF through CSTF77 as well as through a scaffold protein symplekin, likely serving as an additional anchor for the CPA machinery (Takagaki and Manley, 2000; Zhang *et al.*, 2020).

The mammalian cleavage factor II (CFIIm, also known as CFII) is also essential for the process of 3'-end cleavage. It is composed of a heterodimer of CLP1 and PCF11 and co-purifies with CPSF in an RNA-independent manner (de Vries *et al.*, 2000). It may also increase the sequence specificity of the CPA machinery through binding to G-rich sequences further downstream of the G/GU-rich motifs in some pre-mRNAs (Schäfer *et al.*, 2018). Another important function of CFIIm is to mediate transcription termination through the direct interaction of PCF11 with Pol II CTD (Kamieniarz-Gdula *et al.*, 2019; West and Proudfoot, 2008).

More recently, an additional CPA factor RBBP6 was found to be essential for the mammalian CPSF endonuclease activity *in vitro*. While it is not a constitutive subunit of human CPSF, RBBP6 can bind near the active site of the CPSF73 endonuclease and was postulated to induce conformational changes to promote CPSF73 activity (Schmidt *et al.*, 2022). However, this interaction might be weak or indirect as a separate study was able to co-purify them only in the presence of a PAS-containing pre-mRNA (Boreikaite *et al.*, 2022). This discovery highlights the need for additional work to better understand the mechanisms underlying CPSF73 activation.

Another important component of the core CPA machinery is the mammalian cleavage factor I (CFIm, also known as CFI), a heterotetramer consisting of a dimer of CFIm25 (also known as NUDT21 or CPSF5) and a dimer of CFIm59 (also known as CPSF7) or CFIm68 (also known as CPSF6) (Yang *et al.*, 2011). CFIm directly interacts with the UGUA motif in the proximity of PAS through the nudix domain of CFIm25 (Brown and Gilmartin, 2003; Yang *et al.*, 2010). This binding allows recruitment of CPSF via an interaction between CFIm68 and FIP1, and subsequently promotes CPA at this site (Zhu *et al.*, 2018). While CFIm is not required for 3'-end cleavage *in vitro* (Schmidt *et al.*, 2022) *per se*, it acts as an enhancer of CPA and is heavily involved in the selection of alternative polyadenylation sites (see section 1.4.1). The CFIm complex is an important focus for this thesis work.

1.2 Post-transcriptional regulation by 3' untranslated regions

While the same DNA blueprint is encoded in every cell of a human body, coordinated spatial and temporal control of gene expression is crucial for proper cellular functions and homeostasis that give rise to the diverse cellular makeup of an organism. Coincidentally, while protein-coding regions of mRNAs are remarkably conserved, and the number of protein-coding genes is similar between humans and simpler eukaryotes, the sequence space of 3'UTRs has expanded

considerably through evolution to higher organisms (Chen *et al.*, 2012; Mayr, 2017). Additionally, beyond the CDS, 3'UTR-encoded sequence information can also be transmitted to proteins via 3'UTR-mediated protein-protein interactions (Berkovits and Mayr, 2015). Altogether, growing evidence suggests the important roles 3'UTRs may play in regulating the cellular complexity of an organism.

The functions of 3'UTRs are mediated by the interactions between the encoded *cis*-regulatory elements, a network of *trans*-acting factors comprising RNA-binding proteins (RBPs), and the microRNA-Induced Silencing Complex (miRISC). A 3'UTR *cis*-element interacts with one or more *trans*-acting factors and carries multiple functions depending on the given cellular state and setting (Dominguez *et al.*, 2018). Some RBPs further cooperate with one another to enable context-specific functions (Broderick *et al.*, 2011; Campbell *et al.*, 2012). Efficient RBP binding is also limited by the RNA secondary and tertiary structures which are often present in the 3'UTRs (Agarwal *et al.*, 2015; Taliaferro *et al.*, 2016). Together, these 3'UTR-mediated complex interactions lead to the recruitment of effector proteins and facilitate post-transcriptional gene regulation that converges on pathways controlling the stability, translational efficiency, and localization of mRNAs (Mayr, 2019). Importantly, the diversity of 3'UTRs can be further expanded through alternative polyadenylation, generating multiple 3'UTR transcript isoforms from one gene and adding yet another layer of complexity to 3'UTR-mediated post-transcriptional gene regulation.

1.2.1 3'UTRs regulate mRNA stability and translation

Precisely regulated mRNA stability and translation directly impact on the abundance of corresponding protein products. Perhaps the best-studied 3'UTR-mediated mRNA expression regulation is through microRNAs (miRNAs). miRNAs are ~22 nucleotide (nt)-long small non-

coding RNAs that primarily target the 3'UTRs through a partial sequence complementarity (Gebert and MacRae, 2019). They are estimated to control the expression of over 30% of all protein-coding genes in mammals across almost all known cellular processes (Filipowicz *et al.*, 2008; Rana, 2007). Efficient recruitment of miRNAs and their associated miRISC can repress translation as well as initiate deadenylation and the eventual decay of the targeted transcripts (Filipowicz *et al.*, 2008). While sequence complementarity is the primary requirement for miRNA targeting, the efficacy also depends on a variety of other parameters, including site accessibility and the sequence context surrounding the target sites. For example, miRNA target sites situated near both ends of the 3'UTR or within an AU-rich context tend to be more efficient than sites closer to the center of the 3'UTR, presumably due to the presence of RNA folding structures (Grimson *et al.*, 2007). Similarly, sites within the 5'UTR, CDS, or in the immediate vicinity of the stop codon (within ~15-nt) are inefficiently targeted, likely because of interference from the translation machinery (Grimson *et al.*, 2007). Furthermore, multiple miRNA target sites in proximity can cooperate to trigger a more potent silencing output partly through enhancing miRISC affinity for the 3'UTRs and recruitment of effector machineries (Broderick *et al.*, 2011; Flamand *et al.*, 2017). The stoichiometry between a miRNA, the miRISC, and its targets, as well as the relative expression between the competing targets, also significantly contribute to the factors affecting the silencing output (Mayya and Duchaine, 2015; Salmena *et al.*, 2011; Tay *et al.*, 2011). As such, the silencing efficacy of miRNA targeting is often dependent on the individual 3'UTRs and the specific physiological contexts.

3'UTRs are also hubs for another well-studied class of *cis*-elements estimated to be present in 5-8% of all human genes, the AU-rich elements (AREs) (Bakheet *et al.*, 2006). In addition to favoring the destabilization of RNA secondary structures, depending on the repeats, patterns, and surrounding sequence contexts, AREs are recognized by a host of RBPs that can often compete

with one another and at times exert opposing effects on the target mRNAs. Of these ARE-binding RBPs, the AU-rich binding factor-1 (AUF1), Tristetraprolin (TTP), and KH-type splicing regulatory protein (KHSRP) promote degradation of ARE-containing mRNA by recruiting the exosome complex (Lykke-Andersen and Wagner, 2005). In contrast, Human antigen R (HuR, also known as ELAVL1) binding to the AREs primarily stabilizes target mRNAs, likely owing to its inability to recruit the exosome (Chen *et al.*, 2001). In some cases, HuR also acts as a positive translation regulator of mRNAs, including for the tumor suppressive gene *TP53*, *KHSRP*, and *ELAVL1* itself, by mitigating miRNA-mediated translation repression and/or enhancing target mRNA recruitment to polysomes (Mazan-Mamczarz *et al.*, 2003; Pullmann *et al.*, 2007; Srikantan *et al.*, 2012). The binding specificity and capacity of these RBPs can further be modulated by post-translational modifications to confer additional functional flexibility of these AREs (Briata *et al.*, 2005; Shen *et al.*, 2005).

The eukaryotic PUF (Pumilio and FBF) protein-binding elements, typically in the 5'-UGUR-3' pattern (where R represents purine), are another class of *cis*-elements residing primarily in the 3'UTRs (Wickens *et al.*, 2002). PUF family proteins regulate mRNAs with diverse functions, including embryonic development, stem cell maintenance, and neuronal functions, as studied extensively in *C. elegans*, *Drosophila,* and mammalian models alike (Bernstein et al., 2005; Goldstrohm et al., 2018; Parisi and Lin, 1999; Quenault et al., 2011). The best characterized roles of PUFs are as post-transcriptional repressors. This is likely achieved through recruitment of the CCR4-NOT deadenylase complex to accelerate target mRNA degradation (Goldstrohm *et al.*, 2006), as well as recruitment of the translation initiation inhibitor eIF4E-BP to repress translation of the target mRNAs (Blewett and Goldstrohm, 2012). Although the examples are less characterized, PUFs can also enhance the expression of certain mRNAs. Early mutagenesis studies

suggested a direct regulation of these mRNAs by PUFs, which is mediated in part by recruitment of the cytoplasmic poly(A) polymerase, and/or cooperation with the cytoplasmic polyadenylation element binding protein (CPEB) to enhance mRNA stability and promote translation (Kaye *et al.*, 2009; Piqué *et al.*, 2008; Suh *et al.*, 2009).

Local RNA structures in the 3'UTRs not only can dictate the accessibility of RBPs for their *cis*-elements, but also act as *cis*-elements themselves for RBPs that recognize specific stem-loops and double-stranded RNA structures. These interactions can regulate target mRNA stability and translation, in some cases independently of any specific single-stranded sequence element. For instance, in eukaryotes, RNA structures bound by RBPs include the iron regulatory proteins (IRP1 and IRP2) critical for iron metabolism (Binder *et al.*, 1994), Staufen1 of the STAU1-mediated mRNA decay (Kim *et al.*, 2005), and UPF1 of the nonsense-mediated mRNA decay pathway (Fischer *et al.*, 2020). Although considered more specialized and transcript-specific, these structural *cis*-elements further expand the dimension of 3'UTR-mediated mRNA stability and translational control.

1.2.2 3'UTRs regulate mRNA localization

Asymmetric subcellular localization of mRNAs is an evolutionarily conserved mechanism to spatially restrict protein synthesis, which is crucial for proper functions of polarized cell types such as fibroblasts, neurons, as well as oocytes and developing embryos (Martin and Ephrussi, 2009). Prior reporter *in situ* hybridization studies have identified diverse and transcript-specific localization *cis*-elements or "zip codes", primarily situated in the 3'UTRs, which are necessary and sufficient for targeting these mRNAs to specific subcellular locations (Bergalet and Lécuyer, 2014; Kislauskis and Singer, 1992; Lécuyer *et al.*, 2007). Currently, the identified mRNA localization elements range from five or six to several hundred nucleotides-long without a clear

consensus, making large-scale identification of additional localization elements challenging (Engel *et al.*, 2020). Nonetheless, some localization elements function across multiple eukaryotes and cell types, indicating that conserved machineries and RBPs are likely involved (Bullock and Ish-Horowicz, 2001).

One of the primary mechanisms for mRNA localization promoted by *cis*-elements is through active and directed transport, which is well-illustrated in the case of *Drosophila bicoid* mRNA. The presence of several ~50-nt *bicoid* localization elements (BLEs) in the 3'UTR is necessary and sufficient for the localization and anchorage of the *bicoid* mRNA to the anterior pole of *Drosophila* oocytes, and the resulting events are essential for proper embryonic development (Macdonald *et al.*, 1993; Macdonald and Struhl, 1988). Following studies mutated these BLEs to further reveal that rather than relying on the primary sequences, BLEs functions rely on their formation of stem-loop structures that permit intermolecular dimerization and binding of the RBP Staufen (Ferrandon *et al.*, 1997; Macdonald and Kerr, 1997). These interactions enable the formation of RNA-RBP complexes known as ribonucleoproteins (RNPs) that can be packaged into RNA transport granules to travel along microtubules (Ferrandon *et al.*, 1994). Mechanistically similar to *bicoid* localization, the 54-nt-long "zip code" of *β-actin* mRNA first identified in chicken embryos is recognized by the RBP zipcode-binding protein 1 (ZBP1) (Kislauskis *et al.*, 1994; Ross *et al.*, 1997). ZBP1 binding is necessary and sufficient to target *β-actin* transcripts to the lamellipodia of fibroblasts (Latham *et al.*, 2001; Oleynikov and Singer, 2003), as well as to the neuronal dendrites (Eom *et al.*, 2003; Zhang *et al.*, 2001) in a cytoskeleton- and motor protein-dependent manner.

mRNA localization can also be achieved by regulating the local stability of the targeted transcripts. This "degradation/protection" mechanism is demonstrated in the *Drosophila hsp83* mRNA, which is degraded everywhere in the embryo cytoplasm except for the pole plasm

(Bashirullah *et al.*, 2001). Selective degradation of *hsp83* has been mapped to sequence elements in its open reading frame (ORF) and the 3'UTR, which facilitate RBP Smaug binding and subsequent recruitment of the deadenylase complex (Bashirullah *et al.*, 2001; Bashirullah *et al.*, 1999; Semotok *et al.*, 2005; Semotok *et al.*, 2008). Whereas, the *Drosophila nanos* mRNA is localized to the posterior pole of the embryos partly through local stabilization by Oskar, which prevents Smaug from binding the 3'UTR of *nanos* (Zaessinger *et al.*, 2006). In later stages of oogenesis, however, "diffusion/entrapment" is another mechanism contributing to selective *nanos* localization. During these stages, a strong cytoplasmic flow moves *nanos* mRNA throughout the oocyte such that it readily encounters actin-based anchors at the posterior pole, where it becomes entrapped in RNP particles (Forrest and Gavis, 2003). This mode of localization is likely mediated by multiple RBPs, including Rumpelstiltskin and Aubergine, which bind to *cis*-elements at the 3'UTR of *nanos* (Becalska *et al.*, 2011; Jain and Gavis, 2008).

1.2.3 3'UTRs mediate protein complex assembly

While local translation of mRNAs spatially regulated by 3'UTRs is a well-known mechanism of protein localization, more recent findings support a new role of 3'UTRs in mediating protein localization post-translationally, independently of mRNA localization. This was first shown in the 3'UTR of mRNA encoding the transmembrane protein CD47, which normally acts as a phagocytosis-suppression signal (Berkovits and Mayr, 2015). *CD47* generates two mRNA isoforms that only differ in the 3'UTRs: one isoform has a long 3'UTR while the other has a short 3'UTR. Only the long 3'UTR isoform contains AU-rich elements capable of recruiting SET proteins to the site of *CD47* translation at the endoplasmic reticulum (ER) in a HuR-dependent manner. This enables the interaction between SET and the newly translated C-terminus of the CD47 protein and allows subsequent assembly of a protein complex with RAC1, which facilitates

CD47 translocation to the plasma membrane. Whereas, CD47 translated from the short 3'UTR isoform lacking these AU-rich elements is unable to translocate to the plasma membrane, thus failing to protect cells from phagocytosis by macrophages (Berkovits and Mayr, 2015).

Analogous models of 3'UTR-mediated protein complex assembly impacting on functions of the encoded proteins were subsequently demonstrated in additional plasma membrane proteins in human and yeast cells (Berkovits and Mayr, 2015; Chartron *et al.*, 2016; Ma and Mayr, 2018), as well as expanded to nuclear and cytosolic proteins. For instance, the processing bodies (P-bodies) are membraneless cytoplasmic granules composed primarily of translationally repressed mRNAs and proteins related to mRNA decay (Luo *et al.*, 2018). In yeast, the interaction between 40S ribosomal protein S28-B (Rps28b) and enhancer of mRNA-decapping protein 3 (Edc3) facilitates P-body assembly (Fernandes and Buchan, 2020). A recent study uncovered that Edc3 also binds the *Rps28B* 3'UTR directly to facilitate robust Edc3 interaction with the Rps28b protein translated *in cis*. Deletion of the *Rps28B* 3'UTR greatly reduces this interaction and impairs P-body formation (Fernandes and Buchan, 2020). Together, the recently uncovered functions of 3'UTR-mediated protein complex assembly exemplify the transmission of genetic information encoded in the 3'UTRs to proteins with an impact on the protein function, further highlighting the importance of precise 3'UTR regulation.

1.3 Introduction to alternative polyadenylation

The phenomenon of alternative polyadenylation (APA) was first described more than three decades ago (Edwalds-Gilbert *et al.*, 1997), as some genes were found to encode multiple PASs and generate distinct transcript isoforms when a different PAS was utilized during pre-mRNA 3'-end processing. Genome-wide studies estimated that at least 70% of mammalian mRNA-coding genes express APA isoforms (Derti *et al.*, 2012; Hoque *et al.*, 2013). The most frequently utilized

and expressed alternative PASs are situated in the pre-mRNA 3'UTRs, and effectively generate APA isoforms that differ only in the 3'UTRs. These APA events are commonly known as 3'UTR-APA. While much less frequent, and accounting for less than 3% of expressed mRNAs in mouse (Hoque *et al.*, 2013), alternative PASs situated upstream of 3'UTRs, primarily within the intronic regions, may also be utilized. Upstream region APA (UR-APA) events such as these produce APA isoforms differing in both the CDS and 3'UTRs and contribute to proteome diversification by generating both functional and non-functional truncated proteins.

Transcriptomic profiling of APA across various cell types and tissues, biological processes, and diseases indicates that APA patterns and their consequences depend on physiological conditions and gene-specific properties (Derti *et al.*, 2012; Lianoglou *et al.*, 2013; Xia *et al.*, 2014).

1.3.1 3'UTR-APA isoform characteristics and functional relevance

APA occurs most frequently in the 3'UTR sequences of pre-mRNAs and generate 3'UTR isoforms through the use of multiple PASs present in the 3'-terminal exon of a transcription unit (Figure 1.2A). The functional relevance of 3'UTR-APA can partly be inferred through certain characteristic differences between transcripts produced by multi-3'UTR genes and single-3'UTR genes. For multi-3'UTR genes, the 3'UTR isoforms generated can differ significantly in length. In mice, the median length of the longest 3'UTR isoform expressed by each 3'UTR-APA gene (generated by the PAS most distal to the promoter, or distal PAS) is ~1800 nt and the shortest (generated by proximal PAS) is ~250 nt. Whereas, the median 3'UTR length of mRNAs expressed by single-3'UTR genes is ~300 nt (Hoque *et al.*, 2013). The extended 3'UTRs produced by multi-3'UTR genes are thus likely to harbour additional *cis*-regulatory elements to fine-tune the expression of these mRNAs. Interestingly, the modes of transcriptional regulation between single-3'UTR and multi-3'UTR genes also differ. While transcription for ~80% of single-3'UTR genes

is regulated developmentally through controlled chromatin accessibility and expression of transcription factors, most of the multi-3'UTR genes appear to be transcriptionally "on" across cell types (Lianoglou *et al.*, 2013). However, these multi-3'UTR genes express tissue- and cell-specific APA patterns, suggesting an important role of 3'UTR-APA in the functional regulation of these genes, more so than through transcriptional control (Lianoglou *et al.*, 2013).

Changes in miRNA-mediated gene silencing, through both mRNA destabilization and translational repression, are perhaps the most studied consequences of 3'UTR-APA as most conserved miRNA target sites in mammals reside in the 3'UTRs (Ji *et al.*, 2009; Sandberg *et al.*, 2008). When comparing any two tissues in humans, APA explains approximately 10% of differential target mRNA repression by miRNAs, and thus incorporating cellular APA profiles significantly improves miRNA target prediction accuracy (Nam *et al.*, 2014). In this context, the altered expression of 3'UTR isoforms results in differential targeting by miRNAs. This was first demonstrated in activated T cells and cancer cells, both of which express globally shortened 3'UTRs when compared to naïve T cells and non-cancer cells, respectively (Mayr and Bartel, 2009; Sandberg *et al.*, 2008). Expressing shorter 3'UTR isoforms of certain proto-oncogenes in non-cancer cells can lead to more oncogenic transformation than the expression of their longer isoforms due to loss of miRNA-mediated repression (Mayr and Bartel, 2009). Other gene-specific consequences of miRNA de-repression resulting from 3'UTR shortening have also been demonstrated (To *et al.*, 2009; Tranter *et al.*, 2011). Nonetheless, miRNA targeting efficiency is highly dependent on the context of the sequence surrounding the target sites in the 3'UTR. For instance, target sites situated near both extremities of the 3'UTR are more susceptible to targeting (Grimson *et al.*, 2007). Consequently, 3'UTR shortening can also potentiate miRNA binding sites originally buried within the longer 3'UTR isoform (Flamand *et al.*, 2017). In line with this, some

conserved miRNA binding sites are enriched immediately 5' of the proximal PAS of pro-differentiation and anti-proliferation mRNAs. Thus 3'UTR shortening often seen during cell proliferation and transformation can enhance miRNA targeting of these mRNAs to support further proliferative abilities (Hoffman *et al.*, 2016). As such, the interpretation of the effect of 3'UTR-APA on miRNA-mediated gene silencing is dependent on the specific 3'UTRs under study.

3'UTR-APA can similarly modulate the inclusion and exclusion of other *cis*-elements regulating transcript stability and translation, including AU-rich elements (AREs) and binding sites for RNA-binding proteins (RBPs) like PUFs and STAU1 (Tian and Manley, 2017). For instance, a genetic polymorphism in the 3'UTR of the human gene IFN-regulatory factor 5 (*IRF5*) is strongly associated with systemic lupus erythematosus. This variant falls within the proximal PAS of *IRF5*, altering it from the canonical AAUAAA to AAUGAA and causing a shift towards usage of the distal PAS. Consequently, expression of the long 3'UTR isoform is favored. This isoform harbours extra copies of destabilizing AREs and is less stable than the short isoform (Graham *et al.*, 2007).

The length of 3'UTR itself has also been associated with transcript stability and translational control. A longer 3'UTR is considered a feature subjected to nonsense-mediated decay (NMD) (Hogg and Goff, 2010). Long 3'UTR isoforms also have more space to potentially harbour additional destabilizing elements and thus are often assumed less stable than the corresponding short 3'UTR isoforms. However, this view has been challenged previously in a transcriptomic study, where under steady-state the short isoforms were only marginally more stable than the long (Spies *et al.*, 2013). On the other hand, and at least for translational control, previous studies employing polysome fractionation followed by deep sequencing yielded conflicting results regarding the translatability of longer 3'UTR isoforms as they overall seem to associate more with

heavy polysomes and thus are more efficiently translated (Floor and Doudna, 2016; Fu *et al.*, 2018; Spies *et al.*, 2013). Therefore, the extent to which the length of 3'UTR itself can explain transcript stability and translatability remains to be determined. In Chapter 3, I further examine this assumption using transcriptome-wide stability assessment under global APA perturbation.

As the field of 3'UTRs expands, the importance of 3'UTR-APA functions beyond the regulation of transcript stability and translatability is increasingly recognized. For instance, diverse *cis*-elements residing within the 3'UTRs also include localization signals that regulate asymmetric spatial distribution of transcripts to enable localized translation. This is particularly evident for polarized cell types and crucial for early developmental processes. For example, in neuronal cells localized translation in dendrites and axons is common. In some cases, the long 3'UTR isoforms are preferentially localized to the neurites where they are translated, while the short isoforms are restricted to the cell body (An *et al.*, 2008; Harrison *et al.*, 2014; Yudin *et al.*, 2008). Whereas, in another transcriptome-wide survey, although a significant number of 3'UTR transcript isoforms were differentially localized in neuronal cell lines, the number of long isoforms enriched in the neurites compared to the cell body was similar to that of the short isoforms (Taliaferro *et al.*, 2016).

3'UTR-APA serves another important function through mediating differential assembly of protein complexes. This is exemplified by the human BIRC3 ubiquitin ligase known for its role in regulating cell death and immune functions (Beug *et al.*, 2012). While BIRC3 functions in cell death regulation is 3'UTR-independent, only the long *BIRC3* 3'UTR isoform encodes *cis*-elements capable of facilitating STAU1 (Staufen homolog 1)- and HuR-dependent protein complex assembly of BIRC3, IQGAP1, and RALA. This complex in turn associates with CXCR4 to control cell migration (Lee and Mayr, 2019). Such a mechanism contributes to the gain of migratory function in malignant B cells derived from leukemia, where the long *BIRC3* 3'UTR isoform is

upregulated. Importantly, upregulation of the long *BIRC3* isoform is also independent of the overall *BIRC3* mRNA and protein abundance and localization (Lee and Mayr, 2019).

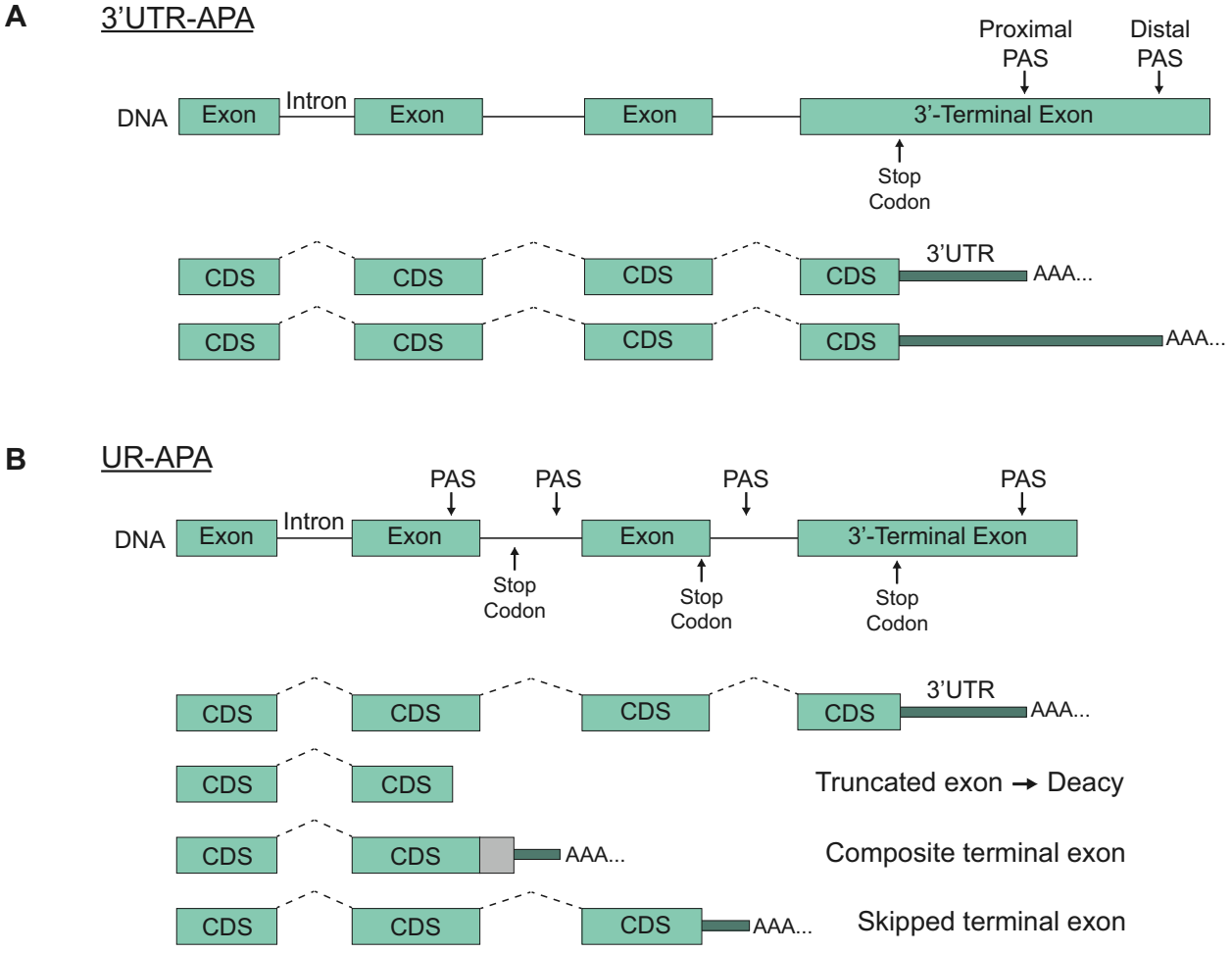


Figure 1.2: Types of APA. (A) 3'UTR-APA are generated by the usage of an alternative polyadenylation signal (PAS) situated in the pre-mRNA 3'UTR. Usage of the proximal PAS closer to the stop codon produces a shorter 3'UTR isoform, whereas usage of the distal PAS further from the stop codon generates a longer 3'UTR isoform. (B) Upstream region (UR)-APA isoforms are produced through the usage of PASs situated upstream of the pre-mRNA 3'UTR. UR-APA can result in transcripts with a truncated exon, which are subject to decay. It can also generate transcripts with a composite terminal exon when the coding sequence (CDS) extends into the neighboring intron, or with a skipped terminal exon when an alternative 3'-terminal exon is used.

1.3.2 Upstream region APA characteristics and relevant functions

While utilized much less frequently and less efficiently than PASs situated in the 3'UTRs, approximately 20-40% of all utilized PASs detected in humans and mice are located upstream of the 3'-terminal exon, primarily within the introns, and can lead to mRNA variants under certain circumstances (Hoque *et al.*, 2013; Tian *et al.*, 2007). These variants generated from upstream region APA (UR-APA) can affect both the coding sequence and 3'UTR. The biogenesis of UR-APA transcripts depends on the dynamic interplay between the polyadenylation and splicing machineries. Depending on several factors including the strength of 5' splice sites and polyadenylation signals, the size of the introns, and relative expression of APA and splicing machineries, the resulting UR-APA transcripts can be divided into three main types (Figure 1.2B) (Edwalds-Gilbert *et al.*, 1997; Tian *et al.*, 2007; Zhang *et al.*, 2005): *truncated exons*, which are generated from internal exonic APA sites; *skipped terminal exons*, which are generated when an exon upstream of the usual 3'-terminal exon is alternatively selected as the terminal exon through splicing; and *composite terminal exons*, which are generated when an exon extends into the downstream intronic region due to inhibition of the 5' splice site. Transcripts generated with truncated exons typically lack in-frame stop codons and are quickly degraded via the non-stop decay pathway (Arribere and Fire, 2018). Meanwhile, the two other different types of UR-APA events, skipped terminal exons and composite exons, can more often generate truncated proteins either without apparent functions or with alternative functions.

UR-APA, like 3'UTR-APA, is expressed in a tissue- and cell type-specific manner (Gruber *et al.*, 2018; Singh *et al.*, 2018). Proliferating cells often exhibit increased UR-APA events, while differentiating cells downregulate UR-APA (Elkon *et al.*, 2012; Hoque *et al.*, 2013; Taliaferro *et al.*, 2016). In a study of APA across multiple tissue types, UR-APA also appeared to be particularly

prevalent among immune cell populations (Lianoglou *et al.*, 2013; Singh *et al.*, 2018). In general, UR-APA contributes to the regulation and diversification of protein functions and abundance as the following examples illustrate.

One well-known example of UR-APA involves the transcript variants produced from the calcitonin-related polypeptide-α gene (*CALCA*). The first transcript variant containing a skipped terminal exon generated through alternative splicing and usage of a promoter-proximal PAS encodes the protein calcitonin. However, when the 3'-terminal exon is included through the usage of the distal PAS, the transcript generated encodes calcitonin gene-related peptide 1 (CGRP) (Amara *et al.*, 1982). The expression of these APA isoforms is also controlled by the splicing factor SRp20 in a tissue-specific manner: the calcitonin-encoding isoform is highly expressed in the thyroid C cells, while the CGRP-encoding isoform prevails in neuronal cells (Lou *et al.*, 1998).

Another extensively studied case is the immunoglobulin M (*IgM*) heavy chain gene. During B cell activation, the transcript generated switches from usage of the distal PAS to a proximal intronic PAS, which results in the inclusion of a composite terminal exon. This switch eliminates the transmembrane domain of the heavy chain and shifts protein production towards a secreted form of antibody from the membrane-bound form (Alt *et al.*, 1980; Takagaki *et al.*, 1996).

UR-APA can also be used by cells to repress gene expression through the generation of truncated transcripts and proteins without apparent functions. For instance, cleavage and polyadenylation factors PCF11 and CSTF77 are found to autoregulate their expression through intronic APA. Each of these genes encodes a highly conserved intronic PAS that would produce severely truncated and non-functional proteins when used. Upregulation of PCF11 and CSTF77 proteins promotes the usage of their own intronic PAS, thus forming a negative feedback loop to control their activity (Pan *et al.*, 2006; Wang *et al.*, 2019). Suppression of gene expression driven

by intronic APA is also observed on a wider scale, such as in B cell leukemia. While patients with B cell leukemia generally have few DNA mutations, the generation of truncated proteins through intronic APA is prevalent, particularly for tumor-suppressive genes (Lee *et al.*, 2018).

1.4 Regulation of APA

As the prevalence and consequences of APA become clearer, it is crucial to understand its regulation and underlying mechanisms. A growing number of cleavage and polyadenylation (CPA) factors and other RNA-binding proteins have been involved in the regulation of APA through their specific activity and abundance. As CPA takes place co-transcriptionally, APA is also regulated by related events and factors including splicing, Pol II transcriptional elongation dynamics, and transcriptional termination. Together, these mechanisms contribute to both global and gene-specific regulation of APA and give rise to the cell type- and physiological condition-specific expression of APA isoforms.

1.4.1 APA regulation by CPA factors

A fundamental mechanism of APA regulation involves modulating the expression of core components in the CPA machinery. This mechanism can preferentially promote the usage of one PAS while attenuating the other, often depending on the *cis*-elements surrounding the cleavage sites. An example is first demonstrated by CSTF64, which is the RNA-binding subunit of the CSTF complex essential for the cleavage reaction (Takagaki and Manley, 1997; Takagaki *et al.*, 1989). CSTF64 is strongly upregulated during B cell differentiation, which increases the availability of the CSTF complexes. This is accompanied by an upregulated usage of the intronic PAS for the *IgM* heavy chain transcript, eliminating the transmembrane domain (Takagaki and Manley, 1998; Takagaki *et al.*, 1996). A similar mechanism is proposed to control the expression of the transcription factor NF-ATc during T cell activation. NF-ATc expresses three prominent transcript

isoforms leading to three protein isoforms: the short isoform is expressed in activated effector T cells, while the two longer isoforms are expressed in naïve T cells. Analogous to B cells, CSTF64 is upregulated in effector T cells where the proximal PAS is used, compared to naïve T cells wherein the distal PAS is used (Chuvpilo *et al.*, 1999). Consistent with these differences, concurrent depletion of CSTF64 and its paralog CSTF64τ in HeLa cells resulted in global 3'UTR lengthening where the distal PAS usage is promoted (Yao *et al.*, 2013). Knockdown of the other two CSTF subunits, CSTF77 and CSTF50, similarly results in global 3'UTR lengthening (Li *et al.*, 2015). These APA regulation events are also sequence-dependent, as U- and GU-rich elements are enriched near the PASs regulated by CSTF64 and CSTF64τ (Yao *et al.*, 2013). Interestingly however, the depletion of CSTF64 or CSTF64τ alone is not sufficient to elicit such an effect, suggesting at least partial redundancy in the regulation of APA by these highly similar proteins (Yao *et al.*, 2012; Yao *et al.*, 2013).

Another CPA factor that promotes distal PAS usage in a sequence-dependent manner is the RBP FIP1, which is a part of the CPSF complex and interacts with poly(A) polymerase (PAP) to facilitate polyadenylation (Kaufmann *et al.*, 2004). Knockdown of FIP1 induces overall 3'UTR lengthening and hampers the capacity of embryonic stem cell (ESC) self-renewal as well as somatic cell reprogramming (Lackford *et al.*, 2014; Li *et al.*, 2015). APA regulation by FIP1 is sequence-dependent as it specifically binds U-rich sequences, which are consistently enriched upstream of PASs regulated by FIP1 (Lackford *et al.*, 2014).

Knockdown of another RBP in the CPA machinery, RBBP6, also causes 3'UTR lengthening. RBBP6 itself expresses multiple protein isoforms. Among these, the full-length isoform 1 (iso1) and the truncated iso3 compete to bind CSTF64. However, only iso1 can facilitate 3'-end cleavage whereas dominant iso3 expression is inhibitory for cleavage. The total protein expression of

RBBP6 as well as the relative expression of iso1 and iso3 therefore can both modulate APA (Di Giammartino *et al.*, 2014). Consistent with this, iso1 and iso3 expression are often inversely correlated, including in cancers where iso1 is upregulated while iso3 is downregulated, and in the differentiation process of mouse myoblasts where the opposite trend is seen (Ji and Tian, 2009; Mbita *et al.*, 2012; Motadi *et al.*, 2011). Interestingly, the knockdown of RBBP6 also preferentially reduces the expression of transcripts enriched with AU-rich elements in the 3'UTR, despite the apparent lack of binding sequence specificity of RBBP6 (Di Giammartino *et al.*, 2014). Nonetheless, this observation suggests a link between 3'-end processing efficiency and the regulation of transcript stability.

PCF11, a component of the CFIIm complex, has also been implicated in APA regulation. Knockdown of PCF11 results in general 3'UTR lengthening (Li *et al.*, 2015; Ogorodnikov *et al.*, 2018). Consistently, lengthening of the 3'UTR during neuronal differentiation correlates with a decrease in the expression of PCF11 (Ogorodnikov *et al.*, 2018). While mechanistically how PCF11 depletion favors the usage of distal PASs is not fully known, it is in line with the role of PCF11 in modulating Pol II transcription termination (Kamieniarz-Gdula *et al.*, 2019; West and Proudfoot, 2008).

Another regulator of APA in the core CPA machinery, and perhaps the best-studied, is the CFIm complex, which forms a heterotetramer consisted of a CFIm25 dimer with a dimer of CFIm59 or CFIm68 (Kim *et al.*, 2010; Yang *et al.*, 2011). Dysregulation of APA through CFIm is associated with defects in cell fate determination as well as disorders including cancer and neuropsychiatric diseases (Brumbaugh *et al.*, 2018; Gennarino *et al.*, 2015; Masamha *et al.*, 2014). Structural studies indicated that CFIm25 specifically interacts with the UGUA element through its nudix domain, whereas CFIm68 enhances RNA binding and facilitates RNA looping through its

RNA recognition motif (RRM) (Brown and Gilmartin, 2003; Yang *et al.*, 2011; Yang *et al.*, 2010). In contrast to most other core CPA factors known to influence APA, depletion of CFIm25 or CFIm68 leads to global 3'UTR shortening (Gruber *et al.*, 2012; Li *et al.*, 2015; Martin *et al.*, 2012; Zhu *et al.*, 2018). Earlier studies have proposed at least three different models for this function for CFIm in APA regulation, and at least conceptually, they are not mutually exclusive. The first model suggests that CFIm functions as a repressor of CPA, possibly through binding to sub-optimal PASs, often proximal, to sterically block CPSF recruitment (Gruber *et al.*, 2012; Martin *et al.*, 2012). A second model suggests that the CFIm25 dimer binds two copies of UGUA, one located upstream of the proximal PAS and the other upstream of a distal PAS, such that the proximal PAS is looped out and thus repressed (Yang *et al.*, 2011). Neither of those two models has been directly tested. A later study proposed a third model suggesting that the CFIm25/68 complex is a UGUA sequence-dependent enhancer of CPA (Zhu *et al.*, 2018). This model is in line with the enrichment of UGUA elements upstream of distal cleavage sites over proximal cleavage sites for genes affected by CFIm25 and CFIm68 deficiency (Li *et al.*, 2015; Zhu *et al.*, 2018). At the molecular level, CFIm68 interacts with FIP1, the aforementioned component of the CPSF complex, through its RS-like domain and subsequently recruits other components of the CPA machinery to the adjacent PAS to promote its usage. This interaction is further modulated by post-translational modifications of CFIm68 as hyper-phosphorylation of its RS-like domain abolishes the interaction with FIP1 in human cells (Zhu *et al.*, 2018). Phosphorylation of CFIm68 RS-like domain was detected in flies where it regulates its nuclear localization and the complex's RNA binding efficiency. In this context, the depletion of two kinases responsible for CFIm68 phosphorylation affects APA of certain transcripts under starvation stress (Tang *et al.*, 2018). Curiously, despite the overall similarity in the protein structure and domains between CFIm59 and CFIm68, CFIm59 only

weakly interacts with FIP1 and partially rescues CFIm68 depletion. Previous studies have also reported a general lack of effect on APA when CFIm59 is depleted in HeLa and HEK293 cells (Gruber *et al.*, 2012; Kim *et al.*, 2010; Li *et al.*, 2015; Zhu *et al.*, 2018). In Chapter 2, we delineate these findings and investigate the contrast in APA regulation by CFIm59 and CFIm68.

PABPN1 is another CPA factor known to control APA. PABPN1 is essential only for polyadenylation but not cleavage, and functions to limit the length of poly(A) tail (~250 nt in humans) by regulating the interaction between CPSF and PAP (Kerwitz *et al.*, 2003; Kühn *et al.*, 2009). Knockdown of PABPN1 induces global 3'UTR shortening (de Klerk *et al.*, 2012; Jenal *et al.*, 2012; Li *et al.*, 2015). Consistently, ectopic expression of trePABPN1, which is a polyalanine-stretch expansion mutant of PABPN1 found in patients with oculopharyngeal muscular dystrophy (OPMD), similarly leads to 3'UTR shortening in cultured cells and mouse muscle tissues (Jenal *et al.*, 2012). This suggests a link between APA dysregulation and OPMD pathology. A recent study suggests that pathogenic trePABPN1 is prone to forming nuclear aggregates that can also sequester CFIm25 to effectively deplete the pool of functional CFIm25 (Guan *et al.*, 2023). Early studies in normal physiological conditions had suggested that PABPN1 could bind directly to proximal PAS regions and suppress their usage by competing with CPSF (Jenal *et al.*, 2012). Curiously, there is no apparent correlation, at least in mouse myoblasts, between the extent of APA regulation (based on the expression shift between 3'UTR isoforms) and 3'UTR size changes upon knockdown of PABPN1. Incidentally, such a correlation is a common feature for other APA-regulating CPA factors (Li *et al.*, 2015). This suggests that APA regulation by PABPN1 may be distinct from that of other CPA factors, but the detailed mechanism remains unclear.

Although some CPA factors are categorized as promoting either proximal or distal PAS usage, this is likely an oversimplification. There are other CPA factors that do not fall into either category

and their depletion can nonetheless affect APA for a restricted subset of transcripts (Li *et al.*, 2015). Aside from APA regulation enacted by individual CPA factors, the mere expression of core CPA factors has been correlated negatively with overall 3'UTR length during cell differentiation and de-differentiation processes (Ji and Tian, 2009). This is consistent with the "first come, first serve" model of polyadenylation site choice, in which the proximal PAS usage is upregulated when the core CPA machinery is not limiting.

1.4.2 APA regulation by transcription

The process of cleavage and polyadenylation happens predominantly co-transcriptionally and is tightly coupled with transcription termination (Proudfoot, 2011). It is thus not surprising that several aspects of the transcription process can influence APA activity. These regulatory influences broadly fall into two categories: Pol II elongation dynamics, and recruitment of factors bridging Pol II and the CPA machinery.

Pausing and backtracking of Pol II effectively reduce the rate of transcription elongation (Gromak *et al.*, 2006; Sheridan *et al.*, 2019). Pausing of Pol II downstream of a functional PAS due to the presence of DNA and RNA sequence elements and structures, such as G-rich sequences, facilitates CPA at this site and contributes to APA (Beaudoin and Perreault, 2013). Increasing Pol II pausing by mutating transcription elongation factors such as TFIIS and SPT5, and the Pol II subunit RPB2 similarly enhances usage of proximal PASs (Cui and Denis, 2003). SPT5 regulation of Pol II elongation and termination is further controlled by its phosphorylation status. Disruption of SPT5 kinase CDK9 or phosphatase PNUT1-PP1 thus also affects Pol II speed and the choice of PAS (Cortazar *et al.*, 2019; Tellier *et al.*, 2022).

Pol II elongation dynamics are also intimately associated with changes in the chromatin structures. Analyses of nucleosome distribution indicate a depletion of nucleosomes immediately

surrounding the region encoding the PASs, likely in part due to the presence of AT-rich sequences (Jiang and Pugh, 2009; Spies *et al.*, 2009). Following this nucleosome-depleted region, sequences downstream of highly utilized alternative PASs exhibit higher affinity for nucleosomes than sequences downstream of less used PASs (Spies *et al.*, 2009). Such an increase in nucleosome density after the PASs is also correlated with Pol II accumulation, suggesting an increased Pol II pausing at these sites (Grosso *et al.*, 2012). In a similar manner, heterochromatin formation surrounding the proximal PASs of certain genes also increases Pol II pausing and promotes CPA at these sites (Neve *et al.*, 2016).

Chromatin structures are also closely connected with DNA methylation through regulated recruitment of chromatin remodeling factors. Global depletion of DNA methylation through genetic deletion of DNA methyltransferase (DNMT) in human HCT116 cells led to APA changes of nearly 500 genes without having a significant effect on their total expression. Importantly, DNMT knockout in this case also did not affect the expression of most core CPA factors (Nanavaty *et al.*, 2020). One of the proposed models for such DNA methylation-dependent APA is through the insulator protein CTCF, whose recruitment to DNA is methylation-dependent. Depletion of DNA methylation in the sequence region between the alternative PASs of a transcription unit allows CTCF binding and enhances the usage of the proximal PAS. Mechanistically, CTCF binding subsequently recruits the cohesin complex to facilitate chromatin looping, and likely forms a physical roadblock for Pol II to prevent further elongation into the distal PAS (Nanavaty *et al.*, 2020). Interestingly, the formation of new chromatin loops through CTCF and cohesin recruitment to these unmethylated regions is concurrent with an altered phosphorylation status of the Pol II CTD (Nanavaty *et al.*, 2020). As CTD is crucial for Pol II-mediated mRNA processing (see below),

changes in CTD modifications may also contribute to the observed increase in proximal PAS usage upon a reduction in DNA methylation.

Pol II CTD and the array of post-translational modifications on its heptapeptide repeat act as a platform for the recruitment of dozens of proteins including elongation factors that can regulate the rate of Pol II elongation (Zaborowska *et al.*, 2016). Among the modifiers of CTD residues is the cyclin-dependent kinase 12 (CDK12). The role of CDK12 as a regulator of APA through CTD phosphorylation was first described in the context of cancer, where CDK12 loss-of-function mutations appeared to specifically increase intronic APA of DNA repair genes, which in turn led to an impaired DNA damage response (Dubbury *et al.*, 2018). Mechanistically, loss of CDK12 impedes Pol II elongation and promotes premature termination in a gene length-dependent manner. Longer transcriptional units are disproportionally affected, including a substantial number of DNA repair genes (Krajewska *et al.*, 2019). DNA repair genes are also more susceptible to altered APA resulting from CDK12 depletion due to lower ratios of U1 small nuclear RNP (snRNP) binding to the number of intronic PASs (more in section 1.4.3) (Krajewska *et al.*, 2019).

Recruitment of CDK12 to Pol II early in the transcription process is mediated by the Pol II-associated factor 1 complex (PAF1C) (Yu *et al.*, 2015). The CDC73 subunit of PAF1C, commonly mutated in hereditary and sporadic parathyroid tumors, also physically bridges Pol II with CPA factors CPSF and CSTF (Rozenblatt-Rosen *et al.*, 2009). These interactions tie together transcription elongation, termination, and CPA. Depletion of CDC73 or another PAF1C subunit PAF1 in mouse myoblast cells leads to Pol II termination defects and increased usage of proximal PASs in both introns and the 3'UTR (Yang *et al.*, 2016). Of note, knockdown of the other PAF1C subunit SKI8, which interacts with the exosome, similarly increases proximal 3'UTR PAS usage but exerts little effect on the usage of upstream intronic PASs (Yang *et al.*, 2016). How different

components of PAF1C can regulate different positions for CPA of a transcript remains an open question.

Additional proteins that bridge Pol II and CPA factors include the SR-related and CTD-associated factor 4 (SCAF4) and SCAF8. These proteins interact with Pol II CTD, PAF1C, as well as components of the CPSF complex, and perform partially redundant functions as "anti-terminators" (Gregersen *et al.*, 2019). Specifically, SCAF4 and SCAF8 bind upstream of proximal PASs in a sequence-dependent manner to suppress transcription termination and CPA at these sites. Through this mechanism, concurrent depletion of SCAF4 and SCAF8 in human cells leads to premature termination at proximal alternative PASs for over 1300 genes (Gregersen *et al.*, 2019).

1.4.3 Other regulators of APA

Additional RBPs that function in other aspects of mRNA co- and post-transcriptional regulation have also been involved in the regulation of APA. For instance, APA is closely interconnected with the process of splicing. Indeed, several protein-protein interactions occur between the core splicing factors and the core CPA factors. At the 3' splice site, immediately upstream of the 3'-terminal exon where the U2 auxiliary factor (U2AF) complex binds, the subunit U2AF65 recruits CFIm59 through direct interaction to stimulate CPA at the PAS immediately downstream (Millevoi *et al.*, 2006). Additional interactions are also seen between U2AF65 and PAP, as well as U2 snRNP and the CPSF complex (Kyburz *et al.*, 2006; Vagner *et al.*, 2000). These interactions, along with early studies indicating that upstream introns can activate PASs, were suggested as a mechanism for the molecular identification of 3'-terminal exons (Martinson, 2011; Movassat *et al.*, 2016; Niwa *et al.*, 1990).

Knockdown of U2AF65 and U1 snRNP components induces overall 3'UTR shortening through alternative 3'UTR PAS usage (Berg *et al.*, 2012; Li *et al.*, 2015). Additionally, the

depletion of these core splicing factors increases intronic PAS usage (Li *et al.*, 2015) This is the case of U1 snRNP, which recognizes 5' splice sites on pre-mRNAs. Inhibition of U1 snRNP increases the usage of cryptic PASs near transcription start sites, suggesting that these sites are normally repressed by U1 snRNP (Kaida *et al.*, 2010). Early studies indicate that U1 snRNP interacts with and inhibits PAP (Gunderson *et al.*, 1994; Gunderson *et al.*, 1998). More recent findings further support a model wherein U1 snRNP suppresses PAS through the formation of a distinct complex with core CPA factors (So *et al.*, 2019). Base-pairing of U1 snRNA to 5' splice sites is essential to maintain the suppressive function of this complex, as the expression of U1 antisense oligonucleotides induced 3'UTR shortening and activation of intronic PASs (Berg *et al.*, 2012; Kaida *et al.*, 2010; So *et al.*, 2019). This process of U1 snRNP binding to 5' splice sites on nascent transcripts to inhibit nearby PASs, often in an intron and gene size-dependent manner, is known as telescripting (Berg *et al.*, 2012; Oh *et al.*, 2017). The phenomenon of telescripting also provides an explanation as to why U1 snRNP is present at higher abundance than other snRNPs in cells despite its 1:1 stoichiometry with other snRNPs in the spliceosome (Baserga and Steitz, 1993; Wahl *et al.*, 2009). Consistently, rapid upregulation of transcription during neuronal activation creates a shortage of U1 snRNPs relative to nascent transcripts, which coincides with prevalent mRNA shortening (Berg *et al.*, 2012).

The family of SR proteins are well-conserved RBPs that harbor RS domains rich in Ser-Arg dipeptide repeats. These proteins were initially discovered as essential regulators for splice-site selection, but are now known to regulate different stages of mRNA biogenesis and metabolism (Howard and Sanford, 2015). Out of twelve SR proteins, ablation of SRSF3 or SRSF7 in mouse P19 cells elicits strong global and opposing changes in APA. SRSF3 depletion induces 3'UTR shortening whereas SRSF7 depletion induces 3'UTR lengthening (Müller-McNicoll *et al.*, 2016).

Mechanistically, both proteins preferentially bind upstream of proximal 3'UTR PASs, however only SRSF7 can recruit the CPA cofactor FIP1 through the unique residues in its RS domain, which is absent from SRSF3. Additionally, depletion of SRSF3 results in unproductive splicing of CFIm68, and subsequently reduces protein expression of both CFIm25 and CFIm68 to indirectly control APA (Schwich *et al.*, 2021). SRSF3 binding to the 3'UTR also promotes nuclear export through interaction with nuclear RNA export factor 1 (NXF1). This further suggests a role of 3'UTR-dependent differential nuclear export in shaping 3'UTR isoform ratios in the cytoplasm (Chen *et al.*, 2019; Müller-McNicoll *et al.*, 2016).

Heterogeneous nuclear RNPs (hnRNPs) are a class of ubiquitously expressed nuclear RBPs known to regulate several aspects of RNA metabolism (Geuens *et al.*, 2016). Of these proteins, hnRNP H1 (also known as hnRNP H), hnRNP H2 (also known as hnRNP H'), and hnRNP F are closely related and share high amino acid sequence identity (Mauger *et al.*, 2008). hnRNP H1/H2/F proteins preferentially bind G-rich sequences and are known to regulate both splicing and APA. hnRNP F competes with CSTF64 for binding to G-rich sequences downstream of PASs and inhibits CPA at these sites (Alkan *et al.*, 2006; Veraldi *et al.*, 2001). In contrast, hnRNP H1/H2 binding to the upstream region of PASs tends to promote CPA, likely through interaction with PAP and stabilization of downstream CSTF64 binding (Bagga *et al.*, 1998; Millevoi *et al.*, 2009). Consistently, during B cell differentiation, the plasma cell lineage expressing the secreted form of antibodies exhibits a higher hnRNP H1/H2 to hnRNP F ratio, while this ratio is reduced in memory B cells that predominantly express membrane-bound form of antibodies (Veraldi *et al.*, 2001). This is because the expression level of CSTF64 is similar between plasma cells and memory B cells, but its activity needed for intronic PAS usage of IgM heavy chain mRNA is attenuated in memory B cells due to upregulated hnRNP F. In line with this, ectopic expression of hnRNP F in plasma

cells impedes CSTF64 activity and reduces expression of the secreted form of antibodies (Veraldi *et al.*, 2001).

The ELAV/Hu family proteins, initially studied in *Drosophila*, are another group of RBPs that regulate different aspects of RNA biogenesis and metabolism that functionally intersect with APA (Wei and Lai, 2022). *Drosophila* Elav (embryonic-lethal abnormal visual system) is known to mediate 3'UTR lengthening in neurons through binding to U-rich sequences typically enriched downstream of proximal 3'UTR PASs and suppressing CPA (Hilgers *et al.*, 2012; Wei *et al.*, 2020). Different lines of evidence support this model. For example, ectopic expression of Elav in non-neuronal S2 cells confers global 3'UTR lengthening, often with shifts to extended 3'UTR isoforms exclusively expressed in neurons (Wei *et al.*, 2020). Tethering of Elav to the vicinity of a PAS also suppresses CPA at this site (Hilgers *et al.*, 2012). Additionally, Elav is recruited to the paused Pol II at GAGA-bearing promoters, which sensitizes these select genes to proximal PAS bypass (Oktaba *et al.*, 2015). This mechanism also reinforces the link between APA and transcriptional control.

Similar APA regulatory functions are also seen in mammalian homologs of Elav, which are Elav-like (ELAVL) 1-4 proteins, also known as HuR, HuB, HuC, and HuD, respectively. Mechanistically, mammalian Hu proteins interact with CPA factors CSTF64 and CPSF160 and interfere with CSTF64 binding to GU-rich *cis*-elements downstream of PASs (Zhu *et al.*, 2007). HuR is expressed ubiquitously whereas the other Hu proteins are largely restricted to neuronal tissues (Akamatsu *et al.*, 1999; King *et al.*, 1994). However, during neurodifferentiation, the expression of HuR is repressed through altered 3'UTR-APA. Specifically, the longer HuR mRNA isoform is translationally repressed and less stable than the shorter isoform. All Hu proteins regulate HuR APA, such that during neurodifferentiation the short HuR isoform expression is

inhibited while expression of the long HuR isoform is enhanced (Dai *et al.*, 2012; Mansfield and Keene, 2012). This regulation serves to balance the pro-differentiation activity of HuB/C/D and the pro-proliferation role of HuR.

1.5 APA profiles and physiological relevance

Advances in next-generation sequencing have allowed transcriptome-wide profiling of APA, namely the sum of expression of all transcript isoforms generated through APA. These large-scale analyses revealed that the dynamic changes in APA profiles are highly dependent on the specific cellular contexts studied. While the functional consequences of APA often remain challenging to interpret, global shifts in 3'UTR isoform expression are consistently observed in different biological processes such as proliferation and cell activation, as well as differentiation and development (Tian and Manley, 2017). Under pathological contexts, global dysregulation in 3'UTR-APA is also seen in diseases including cancer.

1.5.1 Global 3'UTR changes in proliferating and activated cells

General shortening of the 3'UTR frequently accompanies the process of cell proliferation (Neve *et al.*, 2017). For instance, this is clearly the case during T lymphocyte activation. T lymphocyte activation is an important part of the immune response in which T cells increase proliferation and expand clonally. A seminal study demonstrated a global shift towards usage of the proximal PASs, such that 86% of changes in 3'UTR-APA occur in the direction of 3'UTR shortening. This process is conserved across human and murine T cells (Sandberg *et al.*, 2008). Analysis of cell proliferation based on associated gene signatures similarly revealed a negative correlation between the overall 3'UTR length and proliferation across 135 human tissues and cell lines (Sandberg *et al.*, 2008). Such negative correlation was also seen in later studies on mouse embryonic development and through the generation of induced pluripotent stem cells (iPSC) (Ji *et*

al., 2009; Ji and Tian, 2009). While the mechanism driving 3'UTR shortening in proliferating cells is likely manifold, core CPA factors are highly expressed in proliferative cells, compared to their levels in differentiated and less proliferative cells (Ji *et al.*, 2009; Ji and Tian, 2009). This supports a possible mechanism for a dynamically regulated expression of the CPA machinery during the coordinated processes of proliferation and differentiation. Consistently, binding sites for transcription factors related to proliferation and differentiation such as E2F, c-myc, and p53 are shared in the promoter regions of RNA processing genes, including for CPA factors (Elkon *et al.*, 2012; Ji and Tian, 2009).

Neuronal cells activated by depolarizing agents similarly display a trend of mRNA shortening with increased usage of proximal intronic and 3'UTR PASs, specifically for a group of genes under regulation by the MEF2 family of transcription factors (Flavell *et al.*, 2008). Global shifts in APA profile in response to extracellular stimuli also occur in other cell types, including astrocytes, T cells, and B cells. However, the changes in APA are specific to each of the signalling pathways responding to the particular stimulus (Flavell *et al.*, 2008). In line with this, APA regulation for specific pre-mRNAs is controlled by specific signalling pathways. For instance, the mTOR (mammalian target of rapamycin) pathway is a crucial regulator of cell growth and proliferation (Saxton and Sabatini, 2017). In mouse embryonic fibroblast cells, activation of mTOR leads to overall 3'UTR shortening. Among the most affected 3'UTRs are the transcripts encoding proteins related to ubiquitin-dependent proteolysis, which likely functions to supply amino acids in order to sustain the anabolic activity upon mTOR activation (Chang *et al.*, 2015).

1.5.2 Global 3'UTR changes in differentiation and development

APA profiles can also vary as a function of cellular differentiation status. During the generation of iPSC from various sources of differentiated somatic cells, the overall 3'UTR length

is consistently shorter in the reprogrammed iPS cells, which is phenotypically closer to embryonic stem cells (Ji and Tian, 2009). A similar trend is seen in the process of mouse embryonic development and myoblast differentiation wherein increasing differentiation generally correlates with overall 3'UTR lengthening (Ji *et al.*, 2009). Consistently, suppression of the core CPA factor CFIm25 (leading to general 3'UTR shortening) enhances iPSC reprogramming efficiency by more than tenfold, while it impairs the differentiation of embryonic stem cells (Brumbaugh *et al.*, 2018). Interestingly, however, during haematopoietic stem cell differentiation the 3'UTRs are generally shortened (Sommerkamp *et al.*, 2020). This further suggests a cell type-specific regulation of APA dynamics.

Other than these general trends, changes in the 3'UTR length through developmental stages are also characteristic to different cell and tissue types. Brain tissues express progressively lengthened 3'UTRs throughout both embryonic and postnatal developmental stages. Whereas, in testes the 3'UTR length increases during embryonic development and drops sharply after birth when the spermatogenesis program initiates (Ji *et al.*, 2009; Li *et al.*, 2016). During spermatogenesis, genes expressing transcripts with significantly shortened 3'UTRs are also enriched in functions associated with sperm maturation, highlighting the functional relevance of APA in spermatogenesis (Li *et al.*, 2016). More recent analyses of APA at the single-cell scale similarly demonstrated an overall 3'UTR lengthening during mouse embryonic development. However, at each stage, cells of the neuronal developmental lineage consistently have the longest 3'UTRs while cells of the hematopoiesis lineage have the shortest 3'UTRs (Agarwal *et al.*, 2021).

Interestingly, for some tissues, the switch between 3'UTR isoforms goes beyond commonly expressed PASs. For example, the use of unique PASs has been detected in testis or brain tissues (Liu *et al.*, 2007; Miura *et al.*, 2013). In general, testis-specific PASs rely less on the canonical

AAUAAA signal and have unique upstream and downstream elements (Liu *et al.*, 2007). On the other hand, brain-specific PASs often generate unusually long (> 10 kb) 3'UTRs that harbor significantly more *cis*-elements (Miura *et al.*, 2013). This contributes to significantly expanding the potential for brain-specific post-transcriptional regulation. Supporting these tissue-specific APA patterns, analysis of APA at the single-cell scale further demonstrated that different cell types can be resolved from each other based solely on the relative expression of their 3'UTR isoforms (Velten *et al.*, 2015). While little is known about how tissue-specific APA is achieved, it may be partially attributed to the tissue-specific enrichment of expression for some APA regulators, such as CSTF64τ (paralog of CSTF64) in testes and NOVA in the brain (MacDonald, 2019).

1.5.3 APA dysregulation in diseases

Mutations of crucial CPA *cis*-elements in the 3'UTR can disrupt individual APA events and lead to both pathological gain-of-function and loss-of-function phenotypes. For instance, in α- and β-thalassemia, mutations in the PASs of *HBA* and *HBB* mRNAs, respectively, lead to transcription read-through and usage of the downstream PASs. This results in unstable transcripts and reduces the overall expression of *HBA* and *HBB* (Orkin *et al.*, 1985; Whitelaw and Proudfoot, 1986). Whereas, in thrombophilia, the mutation of sequences encoding prothrombin mRNA downstream of a polyadenylation site leads to its increased CPA efficiency and the abnormal accumulation of prothrombin transcripts (Danckwardt *et al.*, 2004; Gehring *et al.*, 2001).

Other than changes in APA for individual genes, global changes in APA profiles associated with different diseases can arise because of defects in CPA factors. For example, the poly-alanine stretch expansion in PABPN1 is linked to global 3'UTR shortening in OPMD patients (detailed in section 1.4.1) (Jenal *et al.*, 2012). Extensive APA defects have also been detected in neurodegenerative diseases like amyotrophic lateral sclerosis (ALS) and frontotemporal dementia

(FTD), which are most frequently caused by a repeat expansion in the *C9ORF72* gene (DeJesus-Hernandez *et al.*, 2011; Prudencio *et al.*, 2015). Notably, global 3'UTR shortening is observed in the cerebellum of ALS and FTD patients featuring this gene expansion (Prudencio *et al.*, 2015). While the exact mechanism for the pathogenicity of this expansion is not fully understood, multiple models have been suggested, including the sequestering of RBPs like hnRNP H1 into nuclear foci through interaction with the mutant C9ORF72 (Lee *et al.*, 2013).

Global changes in APA profiles are also prevalent across most, if not all, cancer types studied (Lin *et al.*, 2012; Mayr and Bartel, 2009; Xia *et al.*, 2014). One of the largest systems analyses across 358 sets of paired tumors and adjacent non-tumor tissues in 7 cancer types identified ~1350 genes with APA changes, with up to 98% being 3'UTR shortening events (Xia *et al.*, 2014). Certain APA changes in tumors are recurrent across cancer types, but the global APA patterns are also specific to cancer types (Xia *et al.*, 2014). Select subsets of 3'UTR-APA events can even be used to add strong prognostic power beyond regular clinical and molecular covariates (Wang *et al.*, 2016; Wang *et al.*, 2018; Xia *et al.*, 2014). These widespread 3'UTR shortening events are also associated with enhanced tumor growth and increased metastatic potential by promoting cellular migration and invasion (Andres *et al.*, 2018; Lai *et al.*, 2015; Masamha *et al.*, 2014). However, the nature and extent of changes in 3'UTR profiles that may accompany the adaptation process of cancer cells in the establishment of metastasis remain unclear. In Chapter 4, we leverage single-cell sequencing technologies and carefully investigate the alterations in the 3'UTR length between paired metastatic and primary tumors.

Functionally, it is thought that 3'UTR shortening eliminates destabilizing *cis*-elements like miRNA binding sites from the shortened 3'UTR isoforms, and thus promotes activation, or de-repression, of oncogenes in the context of cancer (Mayr and Bartel, 2009; Sandberg *et al.*, 2008).

Consistently, in highly proliferative or transformed cells, upregulated genes related to cell growth, such as *IMP1* and *CCND1*, often display 3'UTR shortening (Masamha and Wagner, 2017; Masamha *et al.*, 2014; Mayr and Bartel, 2009; Sandberg *et al.*, 2008). These lines of evidence strongly suggest the impairment of 3'UTR regulatory roles through APA in cancer. However, it is noteworthy that there are still a significant number of 3'UTR lengthening events even in cancer cells undergoing general 3'UTR shortening, including for both pro- and anti-tumor genes (Fu *et al.*, 2011; Mohanan *et al.*, 2022; Singh *et al.*, 2009). Furthermore, some cancer cells instead undergo general 3'UTR lengthening (Fu *et al.*, 2011; Shulman and Elkon, 2019). This highlights the complexity of APA dynamics in cancer and the difficulty in the interpretation of their physiological impacts.

While the mechanisms driving APA changes in cancer are largely unknown, altered expression in some CPA factors has been associated with global 3'UTR shortening. One of the best-known is CFIm25, whose depletion leads to global 3'UTR shortening (Brumbaugh *et al.*, 2018; Li *et al.*, 2015). CFIm25 mRNA is downregulated in several types of solid tumors and its reduced expression is correlated with poor prognosis (Chu *et al.*, 2019; Masamha *et al.*, 2014; Sun *et al.*, 2017). Among different pathways that may be regulated by CFIm25, loss of CFIm25 leads to 3'UTR shortening of genes enriched in the oncogenic RAS signalling pathway in glioblastoma cells (Chu *et al.*, 2019). In this way, CFIm25 depletion may remodel the 3'UTR landscape and alter the responsiveness of certain cells to specific oncogenes or additional signals. For instance, CFIm25 depletion cooperates with weakly activated RAS mutants to enhance RAS-activation phenotypes including proliferation and migration in cancer cells, and the multivulva phenotype in *C. elegans* (Subramanian *et al.*, 2021).

1.6 Current methods for studying APA

Accurate analysis of transcriptomic data often relies on the availability of a comprehensively annotated reference genome. In the case of polyadenylation site annotations, it is far from being complete, and analyses have thus far been limited to few species (Ye *et al.*, 2022). A significant portion of the annotations still cannot be matched across different databases, even within human datasets, which are currently the most abundant (Szkop and Nobeli, 2017). The annotation-building process is also challenging due to the inherent diversity of APA, such as its cell type specificity and the possibility of cryptic PAS activation in certain biological conditions. To overcome the lack of comprehensive polyadenylation site annotation, many current pipelines of APA analysis incorporate *de novo* identification of polyadenylation sites based on the data queried. At present, there are two main routes in achieving a comprehensive transcriptome-wide analysis of APA. One is to use computational methods that can be applied to standard RNA sequencing (RNA-seq) data to extract information on polyadenylation sites; the other method is to devise specific experimental methods aimed at probing the 3'-end of transcripts.

Since RNA-seq has become increasingly accessible in the last decade, the volume of available data and coverage of cell types have accumulated quickly. For this reason, approaches that can identify polyadenylation sites and determine the expression of 3’UTR isoforms from these data are highly promising and in demand. In principle, if RNA-seq reads could yield a uniform coverage along gene loci of individual transcript isoforms, polyadenylation sites would represent positions where the sequencing coverage drops sharply. Most recent computational tools, such as dynamic analysis of alternative polyadenylation in RNA-seq (DaPars) and APAtrap, were implemented based on this principle for *de novo* identification of polyadenylation sites, and have been applied to large-scale analysis of samples from The Cancer Genome Atlas (TCGA) (Feng *et*

al., 2018; Li *et al.*, 2021; Xia *et al.*, 2014; Ye *et al.*, 2018). However, RNA-seq coverage is rarely uniform along gene loci, making accurate identification of polyadenylation sites more difficult and is likely to increase the incidence of false positives. Technical biases, such as a reduced read coverage of the 5' or 3' transcript ends in standard RNA-seq protocols (Wang *et al.*, 2009), also present added challenges. The issue of false positives can be partially addressed with methods that focus on identifying previously annotated polyadenylation sites. Tools that operate on this basis include polyadenylation site usage quantification from RNA sequencing data (PAQR) and quantification of alternative polyadenylation (QAPA) (Gruber *et al.*, 2018; Ha *et al.*, 2018). Nonetheless, these tools are limited by the quality and coverage of the existing polyadenylation site annotations.

While many computational tools have been developed to analyze APA from standard RNA-seq datasets, the inherent biases in standard RNA-seq make it difficult to identify polyadenylation sites accurately and extensively. For this reason, various experimental approaches have been developed to directly enrich reads captured from the 3' transcript ends, generally known as 3'-enriched RNA-seq (Chen *et al.*, 2017). These sequencing methods are intrinsically more accurate for identifying polyadenylation sites. They are also necessary for proper benchmarking of the computational methods employed on standard RNA-seq datasets. In general, these methods enrich for mRNA 3'-end reads by utilizing the poly(A) tail feature and sequencing from the poly(A) tail end. The protocols fall into two main categories depending on the reverse transcription method involved: oligo(dT)-based, or RNA adapter-based. Most current 3'-enriched RNA-seq protocols employ oligo(dT)-based reverse transcription. These protocols are typically straightforward to implement and thus have been broadly used for transcriptome-wide APA studies (Derti *et al.*, 2012; Fu *et al.*, 2011; Jenal *et al.*, 2012; Shepard *et al.*, 2011). However, other than priming the poly(A)

tails of mRNAs, oligo(dT) primers also anneal to internal A-rich sequences, a phenomenon known as internal priming, which can lead to false identification of polyadenylation sites. Most false peaks can be removed computationally through cross-referencing with genomic sequences for the presence of downstream successive As and/or the absence of upstream PAS hexamers. These strategies nonetheless may lose real polyadenylation sites flanked by A-rich sequences or ones without an upstream PAS, which are estimated to account for ~8% of mouse polyadenylation sites (Tian *et al.*, 2005). To circumvent internal priming, different protocols are developed such that RNA adapters are first ligated to the mRNAs. Rather than using oligo(dT) primers, reverse transcription is then carried out with primers annealing to the adapter sequence (Hoque *et al.*, 2013; Jan *et al.*, 2011). However, these methods are more labor-intensive and often involve complex RNA manipulation steps. Some have also reported a subpar performance in expression quantification using these methods (Derti *et al.*, 2012).

More recently, advances in single-cell RNA-seq (scRNA-seq) technologies enable transcriptomic analysis for individual cells and interrogation of cell-to-cell variability. Particularly, 3'-tagged scRNA-seq methods including CEL-seq (Hashimshony *et al.*, 2012), Drop-seq (Macosko *et al.*, 2015), and 10X Chromium (Zheng *et al.*, 2017) produce 3'-end enriched reads suitable for the analysis of APA at the single-cell resolution. However, scRNA-seq datasets are extremely noisy due to high cell-to-cell variability and rate of dropout events, in which a gene is detected at a moderate or high expression level in one cell but cannot be detected in another cell of the same type (Kharchenko *et al.*, 2014). These dropout events can arise due to reasons including the low amounts of mRNA per cell, inefficient mRNA capture, sequencing detection limit, and the general stochastic nature of mRNA expression (Kharchenko *et al.*, 2014). As a result, computational tools developed for bulk RNA-seq are poorly accommodated by the sparse data

produced from scRNA-seq. Specifically for the detection of polyadenylation sites and analysis of APA, a wide range of computational approaches have thus emerged during the last few years (Ye *et al.*, 2022). Some of these tools, including Sierra (Patrick *et al.*, 2020), scAPA (Shulman and Elkon, 2019), and scAPAtrap (Wu *et al.*, 2020), employ the peak calling strategy, based on the principle that reads from 3'-enriched scRNA-seq methods accumulate as peaks at genomic locations upstream of polyadenylation sites. Other tools rely on prior annotations of polyadenylation sites, such as MAAPER (Li *et al.*, 2021) and scUTRquant (Fansler *et al.*, 2023). Lastly, there are tools that do not belong in either category, such as scDaPars (Gao *et al.*, 2021) and APA-seq (Levin *et al.*, 2020). APA-seq requires sequencing data with pair-end reads, while many publicly available scRNA-seq data are produced by single-end sequencing. scDaPars adopts the DaPars pipeline, which is developed for bulk RNA-seq, to first produce the raw relative PAS usage in each cell. The program then employs a regression model to impute missing values of the sparse scRNA-seq counts by borrowing information from the nearest neighboring (most similar) cells. While a comprehensive benchmarking across current scRNA-seq tools for APA analysis is not yet available, each of these recent computational tools has its specific strengths and weaknesses. Nonetheless, they have empowered the discovery of novel cell types and challenged prior notions of "global" APA patterns in the differentiation processes as well as in tumor development (Burri and Zavolan, 2021; Gao *et al.*, 2021; Shulman and Elkon, 2019). In Chapter 4, we leverage these recent breakthroughs to profile APA dynamics during cancer progression.

1.7 Thesis rationale and objectives

With increasing recognition of the critical role of APA in controlling gene regulatory networks closely associated with cell identities and states, significant progress has been made in understanding its regulation and physiological impacts. However, despite the identification of several APA regulators, including the CFIm complex, most of their relevant targets and influence on global post-transcriptional regulation of mRNA stability remain to be investigated. Moreover, while the pathological dysregulation of APA is recurrent and widespread in cancer, how it contributes to the heterogeneity within tumor cell populations and the disease progression is not fully understood. In this thesis, we reached for new insights and answered these important questions by leveraging biochemical assays and bioinformatics analyses across mammalian cell culture and mouse model systems.

In Chapter 2, we identified CFIm as a direct APA regulator of the dosage-sensitive tumor suppressor gene *Pten*. We uncovered the transcriptome-wide opposing functions between CFIm59 and CFIm68, and their broad APA regulatory impact on the oncogenic PI3K/Akt signalling pathway. In Chapter 3, we sought to study the consequences of APA dysregulation on transcript stability by levering global 3'UTR shortening induced in CFIm68-KO cells. We uncovered evidence for a novel role of CFIm68 as a regulator of mRNA stability through the exon-junction complex (EJC)/nonsense-mediated decay (NMD) axis. Lastly, in Chapter 4, we conducted the first longitudinal study of APA from matching primary to metastatic tumors in a breast cancer mouse model at single-cell resolution. We uncovered a surprising correlation between 3'UTR length and proliferation gene signatures, and identified a cell population expressing atypical APA patterns and quiescent stem cell-like gene signatures.

Chapter 2:

Distinct, opposite functions for CFIm59 and CFIm68 in mRNA alternative polyadenylation of Pten and in the PI3K/Akt signalling cascade

Hsin-Wei Tseng[1,2], Anthony Mota-Sydor[1,2], Rania Leventis[1,2], Predrag Jovanovic[2,3,4,5], Ivan Topisirovic[2,3,4,5], Thomas F. Duchaine[1,2,*]

[1] Rosalind and Morris Goodman Cancer Institute, McGill University, Montréal, H3G1Y6, Canada.

[2] Department of Biochemistry, McGill University, Montréal, H3G1Y6, Canada.

[3] Lady Davis Institute for Medical Research, Montréal, H3T1E2, Canada

[4] Gerald Bronfman Department of Oncology, McGill University, Montréal, H4A3T2, Canada

[5] Department of Medicine, Division of Experimental Medicine, McGill University, Montréal, H4A3J1

* Correspondence: thomas.duchaine@mcgill.ca

***Nucleic Acids Research**, 9 September 2022, doi: 10.1093/nar/gkac704*

2.1 Abstract

Precise maintenance of PTEN dosage is crucial for tumor suppression across a wide variety of cancers. Post-transcriptional regulation of *Pten* heavily relies on regulatory elements encoded by its 3'UTR. We previously reported the important diversity of 3'UTR isoforms of *Pten* mRNAs produced through alternative polyadenylation (APA). Here, we reveal the direct regulation of *Pten* APA by the mammalian cleavage factor I (CFIm) complex, which in turn contributes to PTEN protein dosage. CFIm consists of the UGUA-binding CFIm25 and APA regulatory subunits CFIm59 or CFIm68. Deep sequencing analyses of perturbed (KO and KD) cell lines uncovered the differential regulation of *Pten* APA by CFIm59 and CFIm68 and further revealed that their divergent functions have widespread impact for APA in transcriptomes. Differentially regulated genes include numerous factors within the phosphoinositide 3-kinase (PI3K)/protein kinase B (Akt) signalling pathway that PTEN counter-regulates. We further reveal a stratification of APA dysregulation among a subset of *PTEN*-driven cancers, with recurrent alterations among PI3K/Akt pathway genes regulated by CFIm. Our results refine the transcriptome selectivity of the CFIm complex in APA regulation, and the breadth of its impact in *PTEN*-driven cancers.

2.2 Introduction

Alternative polyadenylation (APA) of messenger RNAs (mRNAs) has emerged as a fundamental layer of gene regulation and is thought to contribute to the tuning of at least 70% of all mammalian mRNA-coding genes (Derti *et al.*, 2012; Hoque *et al.*, 2013). APA greatly expands the transcript diversity through utilization of multiple polyadenylation signals (PAS) that lead to expression of mRNA isoforms with varying 3' untranslated region (3'UTR) lengths. PASs proximal to the coding sequence (CDS) give rise to shorter 3'UTR isoforms and can exclude *cis*-regulatory elements that otherwise would be encoded in longer 3'UTR isoforms. This reorganization severs the 3'UTR from its connections with trans-acting factors such as RNA-binding proteins (RBPs), microRNAs (miRNAs), the associated RISC, and their effector machineries, and further affects 3'UTR folding structures. As such, APAs can have profound impact on mRNA translation, stability, localization, and functions that are often closely linked to the molecular and transcriptional makeup of cell identities (Tian and Manley, 2017). APA analysis of single cell RNA-seq datasets also robustly delineated cell subpopulations in tumors and through spermatogenesis (Shulman and Elkon, 2019). Aberrant regulation of APA has been linked with human disease, such as through oncomir targeting (Esquela-Kerscher and Slack, 2006), and is prominent in cancer wherein it is often associated with marked global shortening of 3'UTRs (Gruber and Zavolan, 2019; Mayr and Bartel, 2009; Sandberg *et al.*, 2008).

*P*hosphatase and *TEN*sin homolog (*PTEN*) is one of the most frequently inactivated tumor suppressor genes in cancer (Li *et al.*, 1997). Its activity antagonizes the PI3K/Akt signalling pathway, thereby protecting cells against unchecked cell proliferation, invasion, or evasion of programmed cell death (Song *et al.*, 2012). Its expression must be tightly regulated as even a partial loss of *PTEN* expression can increase the likelihood of tumor formation (Alimonti *et al.*, 2010;

Carracedo *et al.*, 2011). Incidentally, *PTEN* mRNAs undergo extensive APA regulation, with up to 6 distinct lengths of 3'UTR observable in mouse cell lines, and even more in human cells. We had previously reported that longer isoforms exhibit greater stability and contribute to the bulk of protein expression (Thivierge *et al.*, 2018). Resolving how APA itself is regulated and how PTEN dosage is controlled through APA are important to understand gene dysregulation in cancer.

The mammalian 3' end processing apparatus is composed of 16 'core' proteins associated into the CPSF, CstF, CFIm and CFIIm complexes and a poly(A) polymerase (Millevoi and Vagner, 2010; Proudfoot, 2011) and molecular architectures of some of the cleavage and polyadenylation machinery has been detailed (Kumar *et al.*, 2019). The location of pre-mRNA 3' end is mainly defined by consensus poly(A) signals defined by the canonical sequence AAUAAA (Wickens and Stephenson, 1984), and the cleavage site ~15 nt downstream, where polyadenylation occurs. Potency and selectivity of PAS are thought to rely on surrounding *cis*- regulatory elements and input from a number of global APA regulators that were only recently identified (Schwich *et al.*, 2021; Tian and Manley, 2017). Of particular interest for this function is the mammalian cleavage factor I (CFIm), which associates with the UGUA motif in the proximity of PAS. At the molecular level, functional UGUA sites can potentiate cleavage through nearby PAS, but is not essential for cleavage (Zhu *et al.*, 2018). Biochemical and structural studies described CFIm as a heterotetrameric complex consisting of a CFIm25 dimer, which directly recognizes the UGUA sequence, and a dimer of CFIm59 or CFIm68 (Kim *et al.*, 2010; Li *et al.*, 2011; Yang *et al.*, 2011). CFIm subunit dysregulations are expected to lead to broad transcriptome changes. They have been associated with glioblastoma (GBM) aggressiveness and with neuropsychiatric diseases (Gennarino *et al.*, 2015; Masamha *et al.*, 2014), and depletions of CFIm25 and CFIm68 disrupt APA in a broad range of protein-coding mRNAs, including those in the miRNA pathway (Ghosh

et al., 2022). Curiously, it has been reported that depletion of CFIm25 or CFIm68, but not CFIm59, leads to global shortening of 3'UTRs (Martin *et al.*, 2012; Zhu *et al.*, 2018). CFIm59 function is thought to be partly redundant with CFIm68 as it partially rescues APA shift caused by CFIm68 depletion (Kim *et al.*, 2010; Zhu *et al.*, 2018), but the specific functions of CFIm59 in APA regulation remain elusive.

Here we identified CFIm as a direct regulator of *PTEN* APA with significant influences over its mRNA and protein expression. We uncovered distinct and mostly opposing effects for CFIm68 and CFIm59 depletion on APA, not only for *PTEN* but transcriptome-wide, including for several cancer genes. Collectively, our results provide a view of the precise roles for CFIm subunits in APA regulation of PTEN dosage, a broad impact on APA in the PI3K/Akt cascade, and on the overall transcriptome.

2.3 Results

2.3.1 CFIm regulates *Pten* mRNA and protein expression

To determine whether PTEN protein dosage is regulated through alternative polyadenylation (APA) of its transcripts, we first investigated the role of CFIm, a known APA regulator complex. To this end, we knocked down (KD) each member of the CFIm complex (CFIm25, CFIm59, and CFIm68) in the mouse fibroblast cell line NIH3T3 with two different siRNAs, each of which achieving at least 70% KD, and probed for PTEN protein expression by western blot (WB). Depletion of individual CFIm members resulted in significant upregulation of PTEN protein, with comparable increases ranging from 1.3 to 1.8-fold across CFIm members (Figure 2.1A, B). KD of subunits also consistently led to the upregulation of *Pten* mRNAs. RT-qPCR on *Pten* open reading frame (ORF) detected a ~2.5-fold mRNA upregulation upon CFIm25 KD, between 1.5 to 2.7-fold for CFIm59 KD, and between 2.2 to 4-fold for CFIm68 KD (Figure 2.1C). Variation across replicates and the apparently more limited changes in protein expression are consistent with the known multi-layered mechanisms controlling PTEN dosage which may be partially buffering the impact of CFIm KD (see Discussion). Interestingly, we noticed an inter-dependence of CFIm subunits for their expression. Depletion of CFIm25 led to a reduction of both CFIm59 and CFIm68 proteins, whereas depletion of CFIm59 or CFIm68 led to a reduction of the CFIm25 subunit. However, KD of CFIm59 and CFIm68 expression did not affect one another (Figure 2.1D). Coordinated expression of the complex subunits is likely co- or post-translational as mRNA-seq analyses from libraries derived from KD (Figure A1.1, also see below) could not explain this down-regulation.

These results demonstrate the functional relevance of CFIm as an important regulator of PTEN dosage and suggest an impact for this complex on *Pten* mRNA APA.

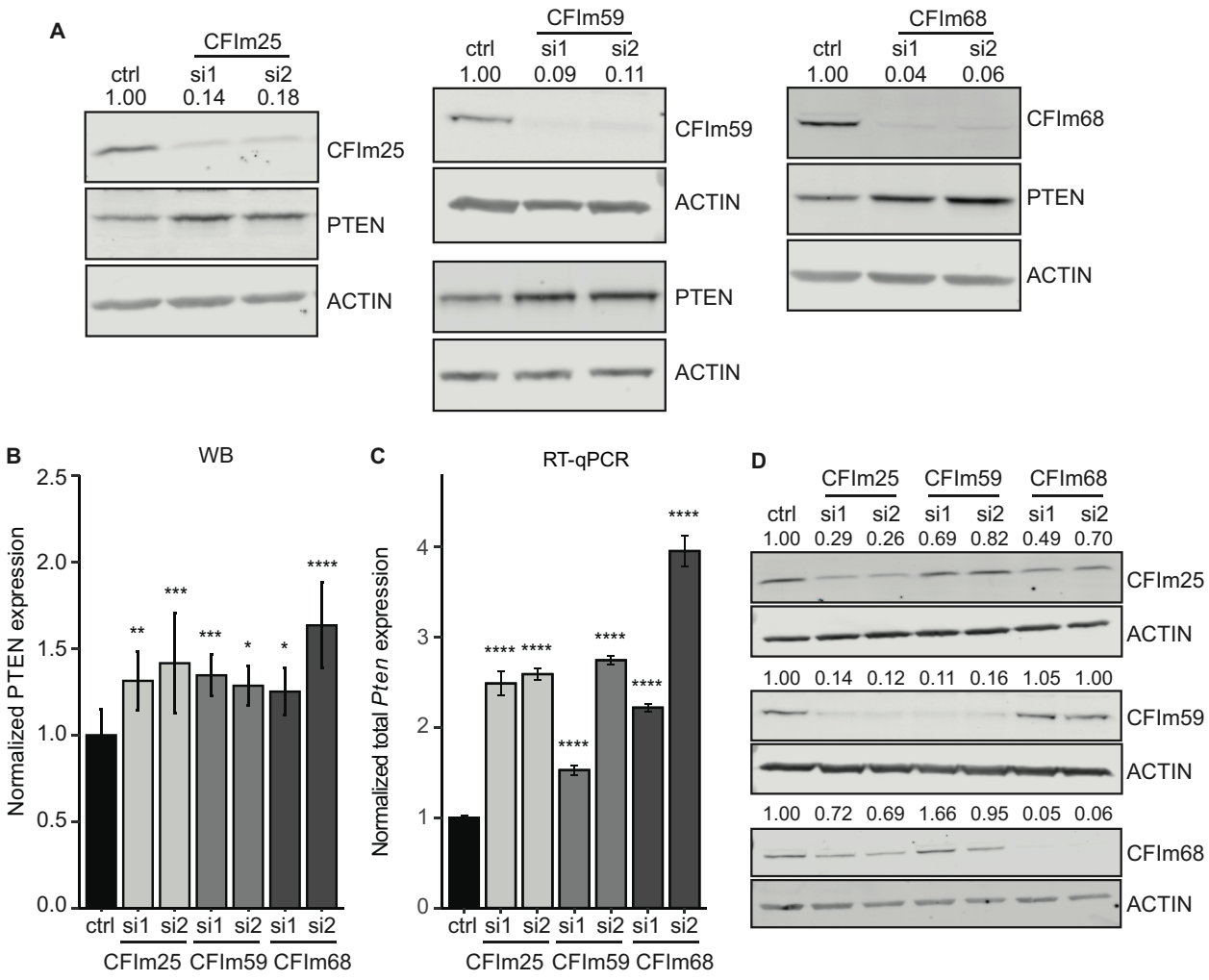


Figure 2.1: CFIm regulates *Pten* mRNA and protein expression.

(A) Representative western blots of CFIm25, -59, and -68 KD and the corresponding change in PTEN level. Two different siRNAs were used for each KD, and the efficiency was quantified. Actin was used as a loading control. (B) Normalized quantification of PTEN protein expression upon CFIm KD. (C) Quantification of normalized total *Pten* mRNA expression upon CFIm KD assayed through RT-qPCR and normalized to *Hprt*. (D) Western blotting of CFIm25, -59, and -68 showing co-dependent protein expression. All error bars indicate mean ± standard deviation across at least three biological replicates. P-values were calculated with one-way ANOVA followed by Dunnett's test (*$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$ and ****$P \leq 0.0001$).

2.3.2 CFIm controls *Pten* mRNA 3'UTR isoform expression

To determine how KD of CFIm subunits leads to *Pten* upregulation, we investigated the impact of their disruption on *Pten* mRNA APA. Using 3'UTR-seq (Lexogen), we observed multiple *Pten* 3'UTR isoforms that we had previously detected by 3' RACE and unambiguously mapped (Thivierge *et al.*, 2018). The two major *Pten* 3'UTR isoforms are defined by a proximal PAS located at 279 nt, and a cluster of 4 PAS located at 3255, 3276, 3312, and 3351 nt downstream of the stop codon, respectively. The 4 PAS near the 3.3k nt mark are clustered in a 102-nt stretch of the 3'UTR, their output is practically indistinguishable and will thus be considered together here. The predominant isoforms will be referred to as APA 300 nt and APA 3.3k. Lesser expressed isoforms between 5.4k and 6.1k nt were also consistently observed and will altogether be referred to as APA 5-6k nt.

KD of CFIm25 consistently shifted APA usage towards the proximal 300 nt isoform, at the expense of the longer 3.3k and 5-6k nt isoforms (Figure 2.2A). KD of CFIm68 resulted in an even more extensive APA shift towards the 300 nt isoform (Figure 2.2A). Intriguingly, knockdown of CFIm59 triggered a distinct and opposite shift, with longer isoforms (3.3k and 5-6k nt) becoming significantly favored relative to the proximal 300 nt isoform (Figure 2.2A). These results were further corroborated using northern blot on CFIm knockdown samples. The proximal (300 nt)-to-distal (3.3k and 5-6k nt) isoforms ratio increased in CFIm25 and CFIm68 KD and decreased in CFIm59 KD (Figure 2.2B). Lastly, we further quantified *Pten* 3'UTR APA upon CFIm depletion using isoform-specific RT-qPCR (Thivierge *et al.*, 2018). For CFIm25 and CFIm68 KD, the 300 nt 3'UTR isoform was increased 5-fold and 6-10–fold, respectively (Figure 2.2C), and this was accompanied by a significant reduction of distal isoforms. In contrast, CFIm59 KD led to the up-regulation of the 3.3k isoform by 1.8 to 3-fold, and up-regulation of the 5-6k nt isoforms by 2 to

4-fold, with no detected changes in the 300 nt isoform using this assay (Figure 2.2C). We further confirmed these observations by generating isogenic NIH3T3 clones where CFIm59 and 68 were knocked out (KO) and restored by transient transfection (Figure 2.2D). CFIm59 restoration (rescue) significantly increased the 300/3.3k isoform ratio in two independent clones, thus shifting Pten mRNA towards shorter isoforms, while this ratio decreased upon CFIm68 restoration (Figure 2.2E). These findings were corroborated by 3'UTR-seq results (Figure 2.2F), although CFIm68 functional restoration in 3'UTR lengthening was only partial in these conditions.

Altogether, these data provide compelling evidence of *Pten* APA regulation by all CFIm subunits and indicate that CFIm59 may exert a distinct and opposing function to CFIm68. We further noticed from both northern and RT-qPCR results, that APA redistribution was accompanied by an increase in overall *Pten* mRNA abundance (Figure 2.1C). In the case of CFIm68 KD, this was mainly driven by over-expression of the proximal 300 nt isoform, while CFIm59 KD instead led to a milder but significant increase in distal 3.3kb and 5-6kb isoforms.

The increase in PTEN protein expression observed upon CFIm59 and CFIm68 KD can at least partly be accounted for by differential expression of *Pten* mRNA 3'UTR isoforms. However, as one may expect perturbations in CFIm59 and -68 to affect overall and individual mRNA isoform translation and stability, we quantified *Pten* mRNA isoform translation and stability under WT, CFIm59 and -68 KD and KO conditions. For this, each *Pten* isoform was detected using 3'UTR isoform specific RT-qPCR across polysome gradient fractions from NIH3T3 cells under CFIm59 or -68 KD (Figure A1.2A). No significant change in the translatability of any individual isoform was detected upon CFIm59 or -68 depletion. Furthermore, no difference in mRNA stability for any individual *Pten* APA isoform was detected when comparing between isogenic NIH3T3 KO of CFIm59 and -68 in Actinomycin-D time-courses (Figure A1.2C, D). As observed before

(Thivierge *et al.*, 2018), longer APA 3.3k and APA 5-6k isoforms were enriched in polysomal fractions (Figure A1.2B), and were more stable in comparison with the short APA 300nt isoform (Figure A1.2D).

Together, these results indicate that the impact of CFIm perturbations on PTEN dosage is due to changes in the relative and absolute expression of individual *Pten* mRNA APA isoforms instead of changes in their translatability or stability.

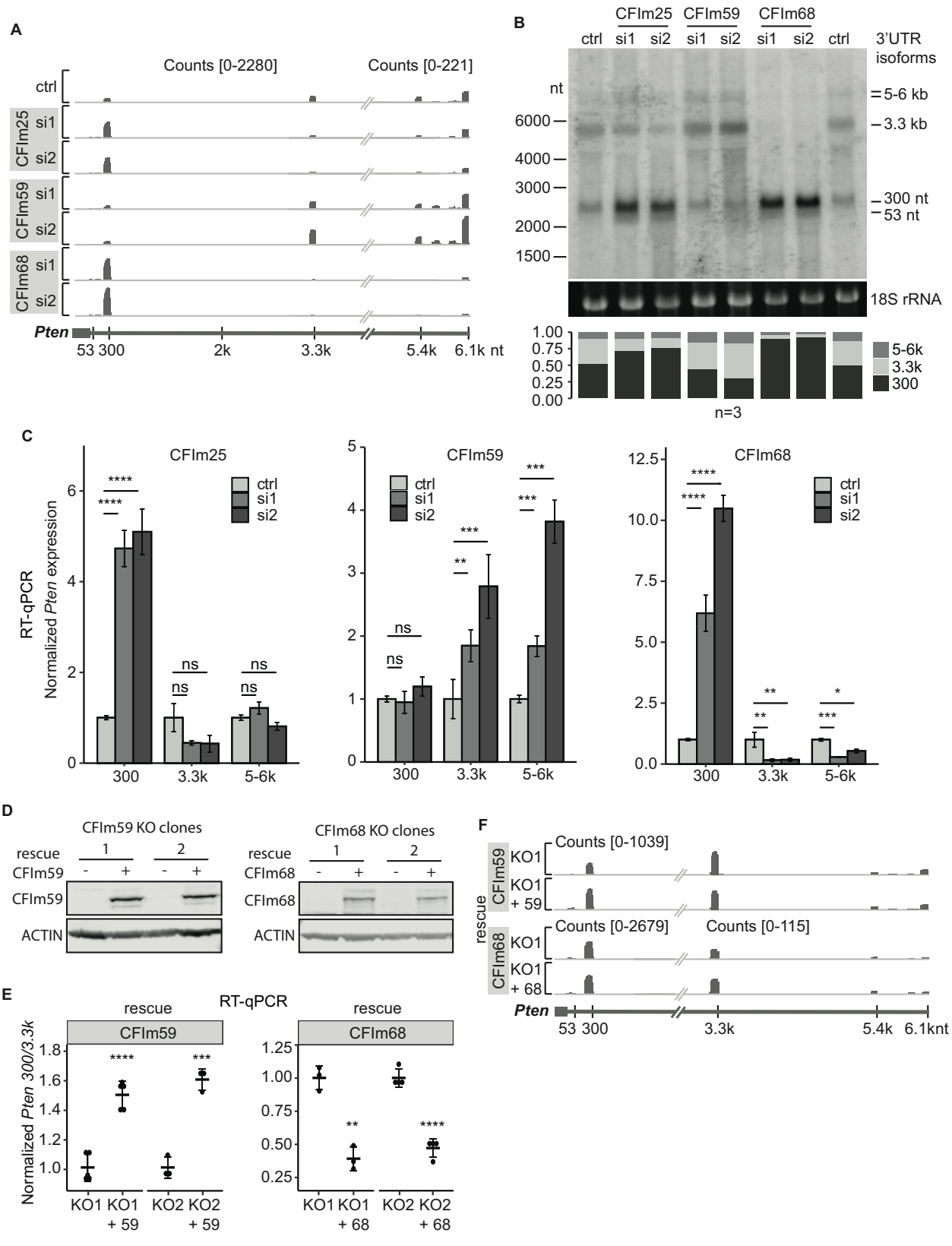

A
Counts [0-2280]
Counts [0-221]
ctrl
CFIm25 si1
si2
CFIm59 si1
si2
CFIm68 si1
si2
Pten
53 300
2k
3.3k
5.4k
6.1k nt
B
CFIm25
CFIm59
CFIm68
ctrl si1 si2 si1 si2 si1 si2 ctrl
3'UTR isoforms
nt
6000
4000
3000
2000
1500
5-6 kb
3.3 kb
300 nt
53 nt
18S rRNA
1.00
0.75
0.50
0.25
0.00
5-6k
3.3k
300
n=3
C
CFIm25
CFIm59
CFIm68
RT-qPCR
Normalized Pten expression
ctrl
si1
si2
300
3.3k
5-6k
ns
D
CFIm59 KO clones
CFIm68 KO clones
rescue
CFIm59
CFIm68
ACTIN
E
RT-qPCR
rescue
CFIm59
CFIm68
Normalized Pten 300/3.3k
KO1
KO1 + 59
KO2
KO2 + 59
KO1 + 68
KO2 + 68
F
Counts [0-1039]
Counts [0-2679]
Counts [0-115]
rescue
CFIm59
CFIm68
KO1
KO1 + 59
KO1 + 68
Pten
53 300
3.3k
5.4k
6.1knt

Figure 2.2: CFIm controls *Pten* mRNA 3'UTR isoform expression. (A) Representative *Pten* tracks of 3'UTR-seq upon CFIm25, -59, and -68 KD in NIH3T3. Schematic of *Pten* and 3'UTR isoforms previously captured by 3' RACE (Thivierge *et al.*, 2018) is shown on the bottom. Each track is separated into two parts with different scales indicated as count ranges on top to show all major isoforms. (B) Quantitative northern blotting of endogenous *Pten* upon CFIm KD in NIH3T3. Corresponding 3'UTR isoforms are indicated on the right with the size markers on the left. Bar graph on the bottom tabulates the average relative composition of the 300 nt, 3.3k and 5-6k isoforms across three biological replicates. (C) *Pten* 3'UTR isoform-specific RT-qPCR for the 300 nt, 3.3k and 5-6k isoforms upon CFIm KD in NIH3T3. (D) CFIm59 and CFIm68 expression rescue in respective KO cells. (E) *Pten* 3'UTR isoform ratio shift upon CFIm59 and CFIm68 rescue. APA 300 nt and APA 3.3k levels are measured by RT-qPCR and taken as ratios before normalization to the mock transfection control. (F) 3'UTR-seq tracks of *Pten* upon CFIm59 and CFIm68 rescue. All experiments were performed in at least three biological replicates with technical triplicates for RT-qPCR. Error bars indicated are mean ± standard deviation. P-values were calculated with one-way ANOVA followed by Dunnett's test in (C) and with Student's *t*-test in (E) (*$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$, ****$P \leq 0.0001$, and ns not significant).

2.3.3 Conservation and species-specific functions of CFIm on human *PTEN* mRNAs

The 3'UTR sequences of human and mouse *Pten* share 74% sequence identity. The predominant proximal (300 nt) and distal (3.3k) PAS as well as the nearby UGUA elements are conserved in human 3'UTR-encoding sequences. The human *PTEN* 3'UTRs further encode additional frequently used PAS, including an additional proximal PAS at 46 nt and another site at 880 nt (Thivierge *et al.*, 2018). We took advantage of the HEK293T isogenic cell line panel to compare the effect of CFIm on human *PTEN* 3'UTRs. Northern blot on endogenous *PTEN* mRNAs led to sensibly the same observations as with mouse sequences (Figure A1.3A). CFIm68

KO strongly favored use of proximal APAs, whereas CFIm59 KO significantly increased distal PAS 3'UTR isoforms in comparison with the parental cell line. These results were further corroborated by mining *PTEN* 3'UTRs from published PAS-seq datasets of CFIm59 KO and CFIm68 KO lines (Zhu *et al.*, 2018) (Figure A1.3B). Strikingly, APA 46nt *PTEN* 3'UTR was drastically reduced in CFIm59 KO, but its use was not significantly altered in CFIm68 KO, indicating that this human-favored proximal APA is exclusively and acutely sensitive to CFIm59 function.

Together, these results highlight the conservation of distinct, opposite roles of CFIm59 and CFIm68 in human and mouse in *PTEN* mRNA APA, and further reveal some species-specific differences in 3'UTR regulatory sequences.

2.3.4 CFIm subunits and UGUA RNA elements in *Pten* APA regulation

The CFIm complex recognizes the UGUA consensus motif near PAS through its CFIm25 RNA-binding subunit (Brown and Gilmartin, 2003; Yang *et al.*, 2011). To determine the direct contribution of CFIm to *Pten* APA regulation, we identified and mutagenized UGUA sites near the two major PAS in mouse 3'UTRs. Three candidate sites surround APA 300 nt, including UGUA1 and UGUA2 at 54 nt and 45 nt upstream, and UGUA3 at 55 nt downstream (Figure 2.3A). Two candidate CFIm binding sites are encoded near the 3.3k PAS (UGUA4 and UGUA5), at 127 nt and 94 nt upstream of the second and most used PAS in the cluster, as determined previously by 3'RACE (Thivierge *et al.*, 2018). We engineered an APA reporter by cloning the full-length mouse *Pten* ORF and 6.1 kb of its 3'UTR genomic sequence in a CMV-driven construct, and the five UGUA candidate sites were mutated to AGAA, individually or in combinations (Figure 2.3A). A similar mutation was previously used to incapacitate CFIm recognition and cleavage *in vitro* (Brown and Gilmartin, 2003). Constructs were transfected transiently in human HEK293T cells

and usage of *Pten* PAS was monitored and quantified using mouse-specific *Pten* mRNA northern blotting. The lack of signal in the empty-vector transfected cells demonstrated the specificity of the northern (Figure 2.3B, D, empty). WT reporter transfection led to expression of *Pten* 300 nt, 3.3k, and 5-6k 3'UTR isoforms, consistent with what was observed endogenously (Figure 2.3B). Relative use of proximal to distal PAS (300 nt/3.3k isoforms) was quantified for each UGUA mutation and normalized to WT (Figure 2.3C, E). Individually disrupting UGUA2 located upstream of APA 300 nt significantly shifted isoform ratio in favor of distal APA 3.3k, while mutation of the upstream UGUA1 or downstream UGUA3 alone had no significant effect. Individual mutation of either of the two UGUA that are proximal to APA 3.3k (UGUA 4 or UGUA5) did not significantly affect *Pten* APA. Double mutants with UGUA2 mutation decreased the 300/3.3k ratio to an extent comparable with the single UGUA2 mutation alone. However, triple UGUA1, 2, and 3 mutations near APA 300 nt further decreased the 300/3.3k ratio, favoring the 3.3k isoform, with contributions that appeared to be additive. In contrast, combining mutations in UGUA4 and 5 greatly increased the magnitude of 300/3.3k ratio, over either UGUA 4 or UGUA5 mutations, suggesting a possible compensation between the two sites (Figure 2.3B, C). Interestingly, when all five candidate UGUA sites were mutated, the 3.3k isoform prevailed over APA 300 nt, suggesting that distal isoforms are intrinsically favored, independently of UGUA and/or CFIm. This may be influenced by cell culture conditions such as density (Thivierge *et al.*, 2018) and/or input from additional, yet unidentified APA regulatory elements in *Pten* 3'UTR.

To investigate the impact of CFIm subunits CFIm59 and CFIm68 on UGUA sites, we profiled reporter PAS use by transfecting isogenic HEK293T lines bearing the CFIm59 or CFIm68 KO (Sowd *et al.*, 2016; Zhu *et al.*, 2018). As in the parental cell line, KO of either CFIm59 or CFIm68 significantly decreased the 300/3.3k APA ratio with combined mutations in proximal

UGUA sites 1, 2, and 3 (Figure 2.3D, E). The combined effect of mutation in the 3 proximal UGUA elements (1, 2, 3) in CFIm59 KO cells was indistinguishable from the parental line (Figure 2.3C, E). This suggests that the shift towards distal PAS upon CFIm59 depletion does not require the proximal UGUA sites. In contrast, CFIm68 KO exacerbated the impact of mutations in proximal UGUA1-3, further enhancing the use of distal PAS over their effect in the parental line (Figure 2.3D, E right panels). Furthermore, whereas mutation of distal UGUA4+5 in CFIm59 KO led to greater enhancement of proximal PAS use over the parental line, the same mutations did not alter proximal-to-distal PAS ratio in CFIm68 KO cells. Normalized 300 nt/3.3k PAS ratio of mutated UGUA4+5 in *Pten* increased from 1.4- to 1.7-fold in CFIm59 KO over the parental line, while this ratio decreased near or at the level (~1.1-fold) of the WT reporter in CFIm68 KO (Figure 2.3C, E). Lastly, to investigate the contribution of 300 nt and 3.3k PAS flanking sequences on PAS usage, we swapped the sequence region encompassing UGUA1 to UGUA3, which includes PAS 300, with the sequence spanning UGUA4 and the 3.3k PAS cluster (Figure 2.3A). Strikingly, when expressed this construct led to a ~3-fold increase in the relative abundance of the 300 nt isoform and completely ablated APA 3.3k expression (Figure 2.3F). Consistent results were observed by RT-qPCR and northern blotting (Figure 2.3F), and similar results were observed in CFIm59 or CFIm68 KO cells (Figure A1.4). These results highlight the importance of mRNA sequence context for the fine-tuning and regulated use of 300 nt and 3.3k PAS, and further suggest the intrinsic strength of 3.3k PAS regulatory elements (Figure 2.3B, C).

Overall, these data demonstrate the competitive regulation of proximal and distal PAS in *Pten* mRNA through UGUA sites. They are consistent with CFIm68 being the main direct regulator of distal (3.3k) UGUA elements 4 and 5, and with a partial compensation on proximal UGUAs 1-

3 by CFIm59 upon CFIm68 loss. Lastly, these experiments further support distinct specificities for CFIm59 and CFIm68 on UGUA elements (See Discussion and Model in Figure 2.7).

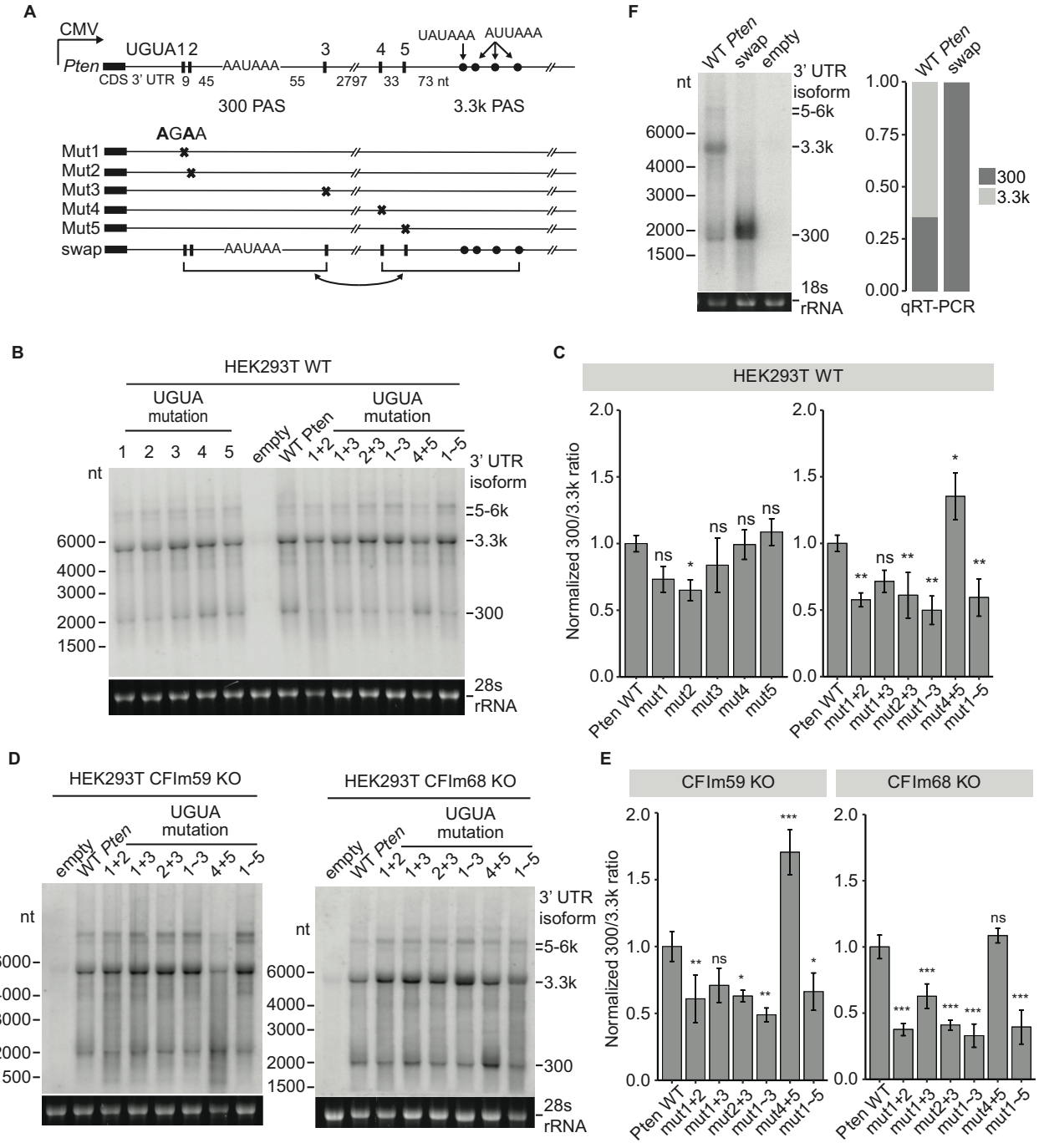


Figure 2.3: CFIm subunits and UGUA RNA elements in *Pten* APA regulation. (A) Schematic of five UGUA motifs surrounding mouse *Pten* PAS 300 nt and PAS 3.3k on a CMV-driven construct.

Distance between the PAS and each UGUA is indicated. Bottom shows individual constructs with mutations introduced at each UGUA site, and the regions swapped in in the experiments featured in (F). (B, D) Mouse *Pten*-specific northern blotting of HEK293T wildtype (B), CFIm59 KO and CFIm68 KO cells (D) transfected with *Pten* UGUA mutant constructs. Size markers are indicated on the left and the *Pten* 3'UTR isoforms on the right. Empty: mock transfection with construct backbone only. (C, E) Quantification of (B) and (D) across three biological replicates. Ratio of 300 nt and 3.3k isoforms was calculated for each transfected sample and normalized to the ratio of wildtype (WT) *Pten*-transfected sample. (F) Mouse *Pten*-specific northern blotting (left) and RT-qPCR (right) on *Pten* wildtype and swap constructs transfected in HEK293T. RT-qPCR quantification was averaged over biological triplicates. Error bars are mean ± standard deviation and P-values were calculated with one-way ANOVA followed by Dunnett's test (*$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$, ****$P \leq 0.0001$, and ns not significant).

2.3.5 Distinct and opposing roles for CFIm59 and CFIm68 on global APA

We next investigated the transcriptome-wide changes in APA upon impairment of CFIm complex subunits. 3'UTR-seq was performed on NIH3T3 cells following CFIm25, CFIm59, or CFIm68 KD, and the usage of distal and proximal PAS was quantified. Consistent with previous observations for *Pten* mRNA, CFIm25 and CFIm68 KD led to 10 to 12 times more downregulated distal PAS than upregulated, while there was 9 to 10 times more upregulated proximal PAS than downregulated (Figure 2.4A, top), thus supporting a transcriptome-wide shortening of 3'UTRs through APA. In contrast, CFIm59 KD resulted in 3.9 times more upregulated distal PAS than downregulated, and 2.6 times more downregulated proximal PAS than upregulated. CFIm59 KD resulted in a global lengthening of 3'UTR through APA (Figure 2.4A top). Similar analyses performed on the published PAS-seq of HEK293T CFIm59 KO and CFIm68 KO isogenic lines (Zhu *et al.*, 2018) led to comparable conclusions, although with fewer captured APA events (Figure

2.4A bottom), possibly due to clonal adaptations. Further analyzing these transcriptome datasets, a score for the Relative Expression Difference (RED) (Li *et al.*, 2015) of distal and proximal PAS usage was allocated for all CFIm-responsive gene passing quality control (see Materials and Methods). A positive RED score reflects an overall increase in distal over proximal PAS usage, whereas a negative RED score indicates the opposite, with an overall increase in proximal over distal PAS. KD of CFIm25 and CFIm68 in NIH3T3 resulted in negative RED scores for 77% and 75% of all responsive genes, respectively (Figure 2.4B, left). Similarly, in HEK293T, 85% of the responsive APA genes gave negative RED scores upon CFIm68 KO (Figure 2.4B, right). In comparison, 66% and 68% of CFIm-responsive genes produced positive RED scores upon CFIm59 KD in NIH3T3 and KO in HEK293T, respectively (Figure 2.4B). These results further support opposing functions for CFIm59 and CFIm68 in APA regulation at the transcriptome level. The current view is that CFIm recognition of UGUA elements is mediated through CFIm25 binding to RNA, and this subunit is a partner to both CFIm59 and CFIm68 subunits. Of the 5,934 genes that exhibit alternative 3'UTR isoforms in NIH3T3 cells, 80% (4,732) responded to the KD of at least one of the CFIm subunits. When we queried the overlap of the 4,420 CFIm59 and CFIm68 responsive genes, only 32% (1,420) overlapped. Moreover, among the 3,939 genes that were oppositely responsive to CFIm59 or CFIm68 KD, only 22% (855) were shared, wherein CFIm59 KD favored distal PAS and CFIm68 KD promoted proximal PAS (Figure 2.4C). We noted that beside *PTEN*, many other well-known genes involved in cancer followed a similar APA behavior. Among top hits along with *PTEN* were *BMPR2*, *ELAVL1*, and *NOTCH1* in HEK293T cells, as well as *Nras*, *Rictor*, and *Notch1* in NIH3T3 cells (Figure 2.4D, E). We further noticed several genes of the PI3K/Akt signaling cascade (see below).

Taken together, these results further support the opposing functions for CFIm59 and CFIm68 in APA regulation as a transcriptome-wide phenomenon, and that CFIm59 and CFIm68 targets only partially overlap. This further prompts a reconsideration of how CFIm PAS specificity is achieved (see Discussion).

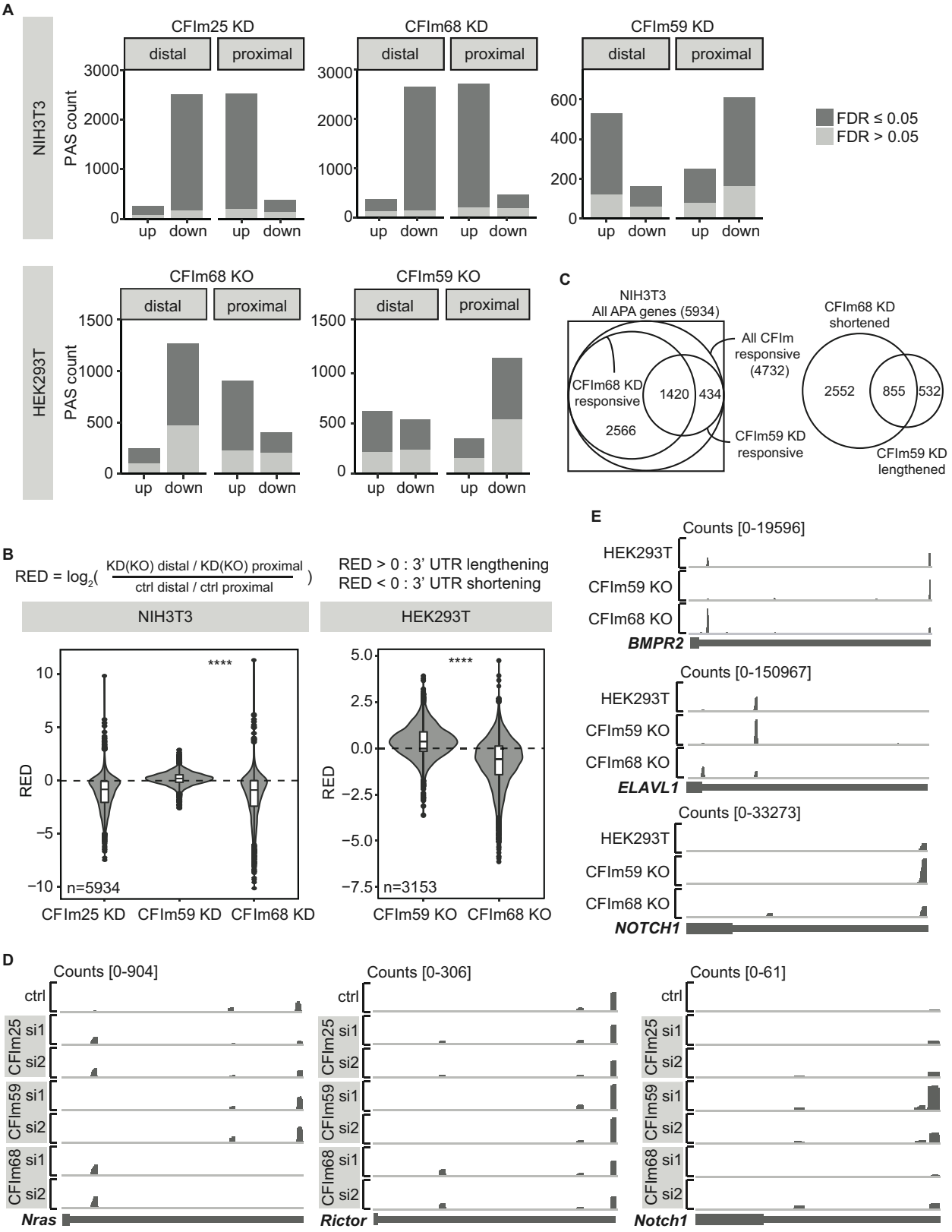

A
NIH3T3
HEK293T
CFIm25 KD
CFIm68 KD
CFIm59 KD
CFIm68 KO
CFIm59 KO
distal
proximal
PAS count
up
down
FDR ≤ 0.05
FDR > 0.05
B
RED = log2(KD(KO) distal / KD(KO) proximal / ctrl distal / ctrl proximal)
RED > 0 : 3' UTR lengthening
RED < 0 : 3' UTR shortening
NIH3T3
HEK293T
RED
****
n=5934
n=3153
CFIm25 KD
CFIm59 KD
CFIm68 KD
CFIm59 KO
CFIm68 KO
C
NIH3T3
All APA genes (5934)
CFIm68 KD
responsive
All CFIm
responsive
(4732)
CFIm59 KD
responsive
1420
434
2566
CFIm68 KD
shortened
2552
855
532
CFIm59 KD
lengthened
D
Counts [0-904]
Counts [0-306]
Counts [0-61]
ctrl
CFIm25
CFIm59
CFIm68
si1
si2
Nras
Rictor
Notch1
E
Counts [0-19596]
Counts [0-150967]
Counts [0-33273]
HEK293T
CFIm59 KO
CFIm68 KO
BMPR2
ELAVL1
NOTCH1

Figure 2.4: Distinct and opposing roles for CFIm59 and CFIm68 on global APA. (A) Up- and downregulated distal and proximal PAS counts upon CFIm KD or KO. Analyses of our NIH3T3 3' mRNA-seq and public datasets of HEK293T PAS-seq were performed to tabulate each PAS expression direction upon CFIm disruption. Statistical significance was calculated with DEXSeq or Fisher's exact test and corrected for multiple testing. Dark gray indicates PAS differential expression with false discovery rate (FDR) ≤ 0.05 and light gray for FDR > 0.05. (B) Relative expression difference (RED) score distribution of NIH3T3 and HEK293T upon CFIm disruption. RED of each gene was calculated as the log2 difference between KD (or KO) distal/proximal and control distal/proximal ratio. Paired two-tailed Student's *t*-test was used (****$P \leq 0.0001$). (C) Overlap of 3'UTRs shortened upon CFIm68 KD and lengthened upon CFIm59 KD in NIH3T3. (D and E) Genome tracks of example cancer-related genes that lengthen upon CFIm59 depletion and shorten upon CFIm25 and -68 depletion in NIH3T3 (D) and HEK293T (E) cells.

2.3.6 CFIm APA regulation in the PI3K/Akt pathway

The PTEN phosphatase hydrolyzes phosphatidylinositol 3,4,5-triphosphate (PIP_3) to phosphatidylinositol 4,5-bisphosphate (PIP_2), and this activity antagonizes PI3K/Akt (protein kinase B) signaling. Given the global role of CFIm in APA regulation, the physiological implications of *PTEN* APA regulation cannot be inferred without integrating its impact on the PI3K/Akt cascade. Among the genes detected in our datasets, 359 genes were assigned to the PI3K/Akt pathway by Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, 2019; Kanehisa *et al.*, 2020; Kanehisa and Goto, 2000), of which 136 (38%) expressed alternative 3'UTR isoforms (APA), and 88 (25%) were responsive to CFIm perturbations (Figure 2.5A). Assignment to the Akt signalling cascade by Gene Ontology (GO) (2021; Ashburner *et al.*, 2000) led to comparable proportions. Of 206 detected genes assigned to the cascade, 76 (37%) underwent APA and 44 (21%) were responsive to CFIm perturbations (Figure 2.5A). PI3K/Akt cascade

components that are responsive to CFIm include upstream activators such as receptor tyrosine kinases (RTK), the extracellular matrix (ECM), and G protein-coupled receptors (GPCR), as well as downstream targets of *Akt* such as *TSC1*, *Myc*, and *Bcl2* oncogenes (Figure 2.5B). Perhaps most strikingly, the catalytic p110α (*Pik3ca*) and regulatory p55γ (*Pik3r3*) subunits of PI3K, as well as *Akt3* responded to CFIm25, CFIm68 and CFIm59 KD similarly to *Pten*: CFIm25 and CFIm68 KD leading to 3'UTR shortening with negative RED scores, and CFIm59 KD leading to 3'UTR lengthening with positive RED scores (Figure 2.5C, D).

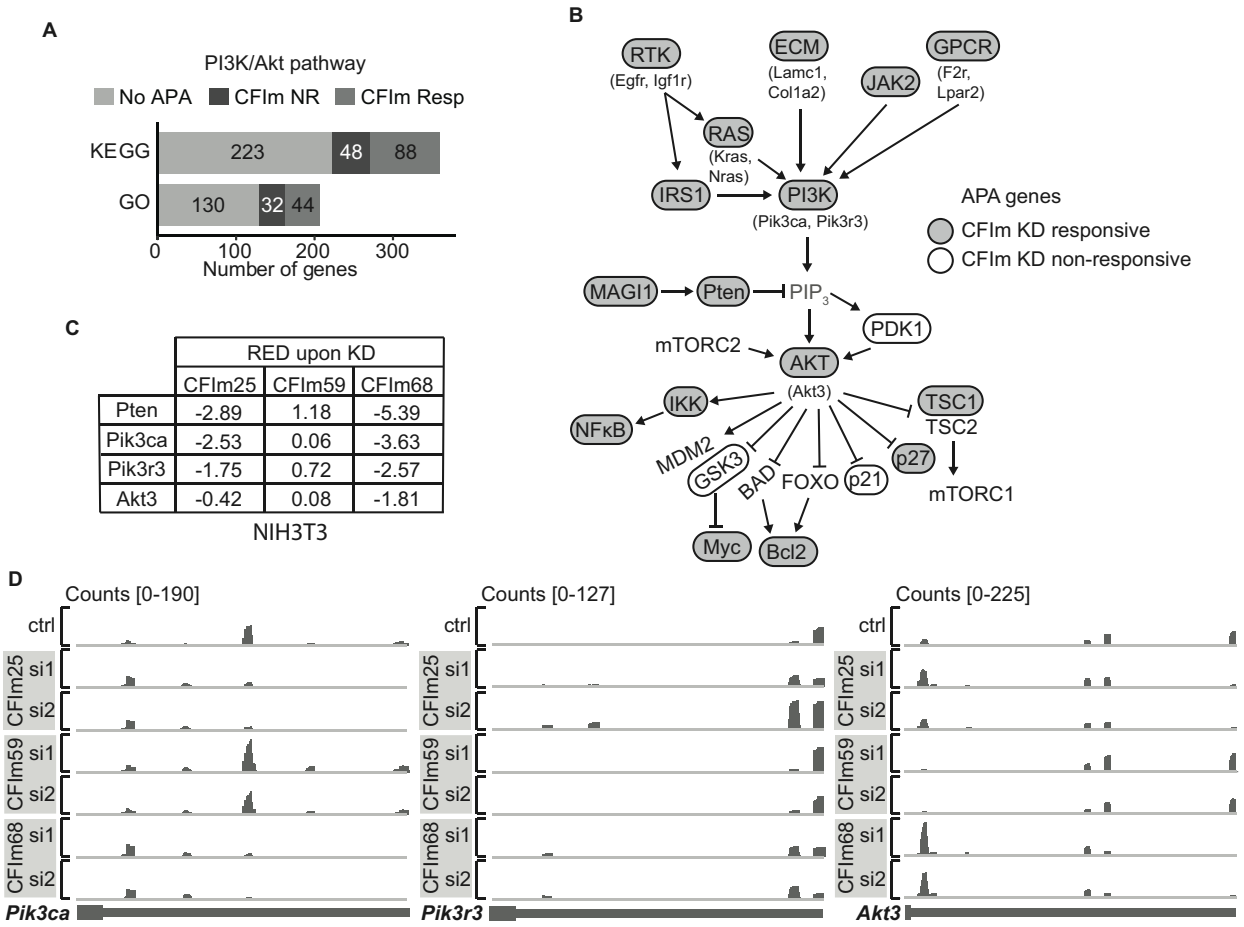


Figure 2.5: CFIm APA regulation in the PI3K/Akt pathway. (A) Tabulated APA status of genes designated to the PI3K/Akt pathway by KEGG, or the Akt signalling cascade by GO, from our NIH3T3 3'UTR-seq. No APA: genes with no detectable alterative 3'UTR isoform. CFIm NR: genes

not responsive to CFIm KD. CFIm Resp: genes with detectable switch in 3'UTR isoform usage upon CFIm KD. (B) Schematic of the PI3K/Akt cascades. Genes responsive to CFIm KD are circled and shaded in gray, with example genes in brackets. Genes non-responsive to CFIm KD but with more than one detectable 3'UTR isoform are circled without shading. Genes with no detectable APA are not circled nor shaded. (C and D) RED scores (C) and gene tracks (D) of core PI3K/Akt components Pten, Pik3ca, Pik3r3, and Akt3, upon CFIm KD in NIH3T3.

2.3.7 Distinct associations of *PTEN* APA, CFIm59 and CFIm68 across onco-transcriptomes

Pten and PI3K/Akt cascade misregulation is closely associated with several types of cancers. We next queried whether APA regulation of *PTEN* was altered in *PTEN*-driven cancers and compared it with other core PI3K/Akt cascade components. To enable enough depth across cancer subtypes, we relied on APA scoring by Percentage of Distal polyA site Usage Index (PDUI) from published mRNA-seq datasets. A higher PDUI means increased distal PAS usage (Xia *et al.*, 2014). We integrated the PDUI scores of prostate cancer and glioblastoma patients of known subtypes from The Cancer 3'UTR Atlas (TC3A) datasets (Xia *et al.*, 2014) with PDUI of normal tissues taken from APAatlas (Hong *et al.*, 2019). In prostate cancers, and regardless of the transcriptomic subtype, the 3'UTR of *PTEN* was significantly lengthened. Curiously and in contrast, the 3'UTR of *AKT3* was significantly shortened, whereas no clear trend could be discerned from those datasets for PIK3CA and PIK3R3. In glioblastoma subtypes the trend was opposite, with *PTEN* mRNAs being expressed with more proximal PAS usage, and AKT3 3'UTRs being lengthened. Among other CFIm-responsive genes, PIK3CA 3'UTRs were significantly shortened as well (Figure 2.6A and Figure A1.5). These results highlight the heterogeneity of APA changes across cancer subtypes and argue against onco-transcriptome 3'UTR misregulation through loss or impairment of any single CFIm subunit.

Lastly, we compared the expression of *PTEN* mRNA and *PTEN* PDUI with CFIm59 and CFIm68 expression across cancers. We performed this analysis across 34 types of cancers where mRNA-seq datasets were available from The Cancer Genome Atlas (TCGA), and gathered the corresponding PDUI from TC3A. Overall, *PTEN* mRNA was negatively correlated with CFIm59 or CFIm68 mRNA abundance (Figure 2.6B), in general agreement with our observation of *Pten* dosage up-regulation upon CFIm59 or CFIm68 KD (Figure 2.1C). *PTEN* mRNA expression correlated with *PTEN* 3'UTR lengthening, as indicated by *PTEN* mRNA PDUI (Figure 2.6C). However, comparisons of *PTEN* PDUI with mRNA abundance of CFIm59 or CFIm68 across cancer types revealed opposite trends. *PTEN* PDUI was negatively correlated with CFIm59 mRNA expression, as opposed to a strong positive correlation with CFIm68 across all cancer types (Figure 2.6C).

These findings further corroborate our overall conclusion that CFIm68 and CFIm59 exert distinct and opposite functions on APA dynamics in physiological and oncogenic transcriptomes.

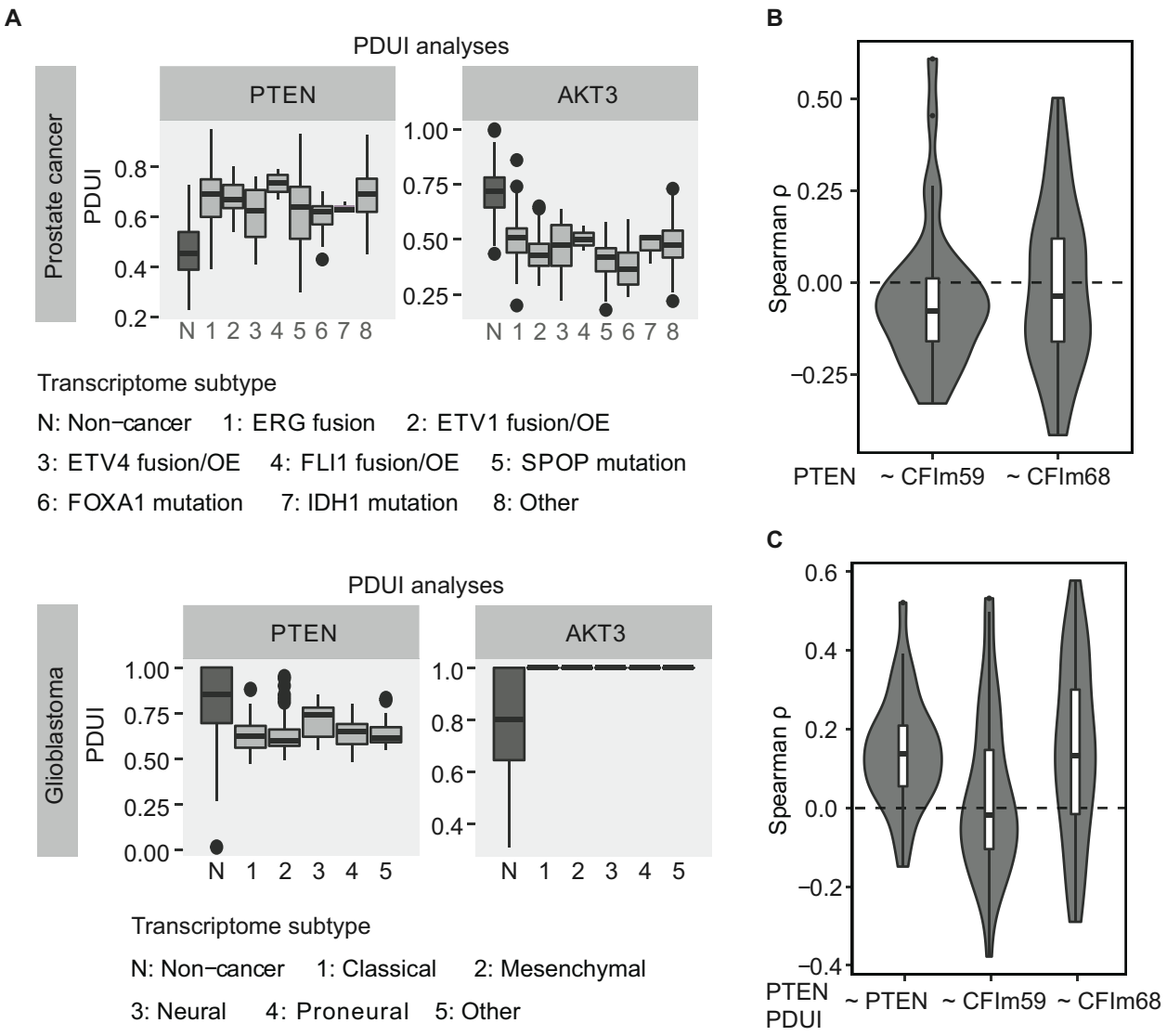


Figure 2.6: PTEN APA in cancers. (A) PDUI analyses of PTEN and AKT3 in prostate cancer and glioblastoma. Each cancer is divided into transcriptomic subtypes defined by The Cancer Genome Atlas (TCGA) project as shown in the bottom legend. PDUIs of cancer datasets were collated from The Cancer 3'UTR Atlas (TC3A). PDUIs of non-cancer prostate and brain tissues were mined from APAatlas. (B, C) Spearman correlations of CFIm59 and CFIm68 mRNA level with *PTEN* expression (B), and of *PTEN*, CFIm59 and CFIm68 mRNAs with *PTEN* PDUI (C) across 34 cancer types (detailed in Materials and Methods).

2.4 Discussion

Here we identify the CFIm complex as a direct regulator of *Pten* mRNA APA and show that this role has a significant impact on PTEN protein dosage. We further uncovered distinct and opposing roles for the CFIm59 and CFIm68 subunits in APA regulation on *Pten* mRNA and globally at the transcriptome level. CFIm59 depletion led to 3'UTR lengthening, while CFIm68 depletion led to 3'UTR shortening and favored distal PAS use. These results prompt a reconsideration of how CFIm specificity is established and reveal the breadth of APA regulation in the PI3K/Akt signaling cascade, as well as transcriptome-wide.

Even slight changes in PTEN dosage impact the output of the PI3K/Akt cascade, and its oncogenic down-regulation contributes to cancer initiation and progression (He, 2010). While CFIm perturbation results in an increase in both *PTEN* mRNA and protein expression, APA is merely one component of a much broader network of transcriptional, post-transcriptional, and post-translational regulation mechanisms controlling PTEN dosage (Kim *et al.*, 2010; Tay *et al.*, 2011; Tian *et al.*, 2013; Xiao *et al.*, 2008). For example, part of *Pten* mRNA upregulation upon CFIm KD could be indirectly affected by *Pten* transcriptional regulators such as EGFR (Virolle *et al.*, 2001) and NFκB (Vasudevan *et al.*, 2004) which are also CFIm sensitive. A recent study identified *PTEN* transcriptional enhancers as one of the contributors to its APA regulation (Kwon *et al.*, 2022). Furthermore, another recent study proposed a mechanism of PTEN protein destabilization through APA regulation of WWP2 E3 ubiquitin ligase by CFIm59 (Fang *et al.*, 2020). As such, the contribution of CFIm itself on PTEN dosage does not solely result from its direct involvement in *PTEN* mRNA APA.

How 3'UTR isoforms, individually or as co-expressed combinations, control or contribute to *PTEN* mRNA translation is also the result of input from a complex regulatory network. The 3.3k

PTEN 3'UTR, by far the best studied of its mRNA isoforms, is thought to be a nexus of several post-transcriptional regulation mechanisms (Kim *et al.*, 2010; Tay *et al.*, 2011; Tian *et al.*, 2013; Xiao *et al.*, 2008). Dozens of miRNAs are predicted to bind this 3'UTR, and its competition with a pseudogene served as conceptual framework for the ceRNA hypothesis. Even such indirect connections seem to exert influence on PTEN dosage; QTL analyses suggested that the *PTEN* 3'UTR pseudogene ceRNA drives a significant part of PTEN variation in cancers (Park *et al.*, 2018). It would stand to reason that longer 3'UTR isoforms, which encode more predicted regulatory elements, are better repressed by miRNAs, and that shorter isoforms are de-repressed. However, our previous work instead revealed that longer 3'UTR isoforms contribute to the bulk of PTEN dosage, are largely resistant to miRNA regulation (presumably through folding structures (Kertesz *et al.*, 2007)), and are in fact more stable than shortest isoforms (Thivierge *et al.*, 2018). These conclusions were yet again confirmed through our polysome profiling and mRNA stability assays (Figure A1.2B, D). In agreement with this, a drastic 6- to 10-fold increase in the short 300 nt *Pten* 3'UTR isoform upon CFIm68 KD only resulted in a 1.5-fold increase in PTEN protein, and a similar increase was observed upon CFIm59 KD, with merely 2- to 4-fold increase in the long isoforms (Figures 2.1, 2.2, 2.7A). In light of our results and given the multi-layered regulation converging on the 3'UTRs of *PTEN* mRNAs, a direct assessment of their translatability under oncogenic contexts appears as an important next research frontier.

2.4.1 Distinct selectivity for CFIm59 and CFIm68 in APA regulation

Our results are consistent with a model whereby CFIm59 and CFIm68 have partially overlapping – but distinct – preferences for specific subsets of UGUA elements and functionally compete in directing PAS usage on pre-mRNAs that bear several UGUA clusters such as *PTEN* (Figure 2.7B). Consistent with this model, CFIm59 and CFIm68 can partially compensate for each

other upon loss of one of the subunits (Figure 2.7C). This is illustrated by our analyses of UGUA mutations in wt and CFIm59- and CFIm68-KO cell lines (Figure 2.3). Indeed, CFIm68 appeared to drive most of UGUA4 and 5 input on the 3.3k isoform PAS, this activity was bolstered in CFIm59 KO, and UGUA4 and 5 still selectively promoted the nearby PAS in CFIm68 KO cells. Our model is consistent with prior reports of partial compensation between CFIm59 and CFIm68, namely in the recruitment of the CPSF subunit Fip1 to promote cleavage (Zhu *et al.*, 2018). Neither is it at odds with the reported enrichment of UGUA sites upstream of distal PAS, which had been regarded as a mechanism of distal PAS promotion by CFIm. However, that prior model did not account for the many cases such as *Pten* mRNA, where multiple UGUA sites are encoded near highly utilized proximal PAS, nor for the opposing functions of CFIm59 and CFIm68 subunits. CFIm-responsive transcripts that encode both proximal and distal UGUA elements identified in our analyses thus revealed a more complex dynamic between the CFIm subunits.

It may be hypothesized that CFIm59 acts as a negative regulator of CFIm68 through their direct interaction in a sub-population of CFIm complexes. This possibility could be considered as CFIm59 and CFIm68 can co-immunoprecipitate from HEK293 and glioblastoma cell lines (Chu *et al.*, 2019; Gruber *et al.*, 2012). However, overexpression of CFIm59 does not mimic the effect of CFIm68 depletion on APA, at least for some of the CFIm-sensitive mRNAs (Kim *et al.*, 2010). Considering our results, a more likely explanation is that distinct CFIm59 and CFIm68 CFIm complexes, associated with different UGUA elements, compete in their output on a subset of PAS. In line with this, only 32% of CFIm-sensitive PAS are responsive to both CFIm59 and CFIm68 KD. Mining of public PAR-CLIP datasets performed in HEK293 cells (Martin *et al.*, 2012) further supports this view. While both CFIm59 and CFIm68 signals peak at ~50 nt upstream of PAS cleavage sites across all APA genes (Figure A1.6A, B), CFIm68 peaks are selectively and

significantly enriched upstream of distal cleavage sites in mRNAs that were shortened upon CFIm68 depletion, as previously noticed (Zhu *et al.*, 2018). In contrast, CFIm59 peaks are instead enriched upstream of proximal cleavage sites, and across all APA mRNAs (Figure A1.6B). Lastly, the distinct functions of CFIm59 and -68 is yet further supported by the coupled expression of the CFIm25 and -68 subunits, or CFIm25 and -59, while CFIm59 and -68 KD do not affect each other (Figure 2.1D).

The CFIm25 subunit directly binds to the UGUA elements in pre-mRNAs, but it is likely that CFIm25/CFIm59 achieve a distinct preference from CFIm25/68 complexes by directly contacting nearby additional sequences or through RNA-binding co-factors. CFIm59 and CFIm68 share an overall protein structure reminiscent of classic splicing SR proteins, which consist of a central proline-rich region flanked by an N-terminal RRM and a C-terminal RS-like domain. On their own, subtle structural differences in the width of the RNA exit cleft could lead to distinct RNA preferences (Yang *et al.*, 2011). The most obvious difference between the CFIm68 and -59 subunits is the glycine-arginine rich (GAR) motif missing from CFIm59. Furthermore, the RS region of CFIm68 is thought to mediate interactions that are distinct from those of CFIm59 (Dettwiler *et al.*, 2004). Lastly, potential interactors of CFIm59 captured by co-fractionation and BioID RNA binding proteins including HuR and HNRNPA1, have previously been implicated in APA (Go *et al.*, 2021; Havugimana *et al.*, 2012; Jia *et al.*, 2019; Wei *et al.*, 2020; Youn *et al.*, 2018). Any of those differences may also contribute to the specificity of PAS enhancement. While it stands to reason that CFIm59 PAS specificity can be defined through a variety of interacting partners, some RNA elements may be more prevalent than others near responsive UGUA sites. Our simple motif enrichment analysis identified a motif significantly enriched near proximal PAS in mRNAs whose 3'UTR is lengthened upon CFIm59 depletion, against proximal PAS in mRNAs

with shortened 3'UTR when CFIm68 is depleted (Figure A1.6C). Future work will be required to determine the significance of such motifs and identify the determinants of their molecular recognition.

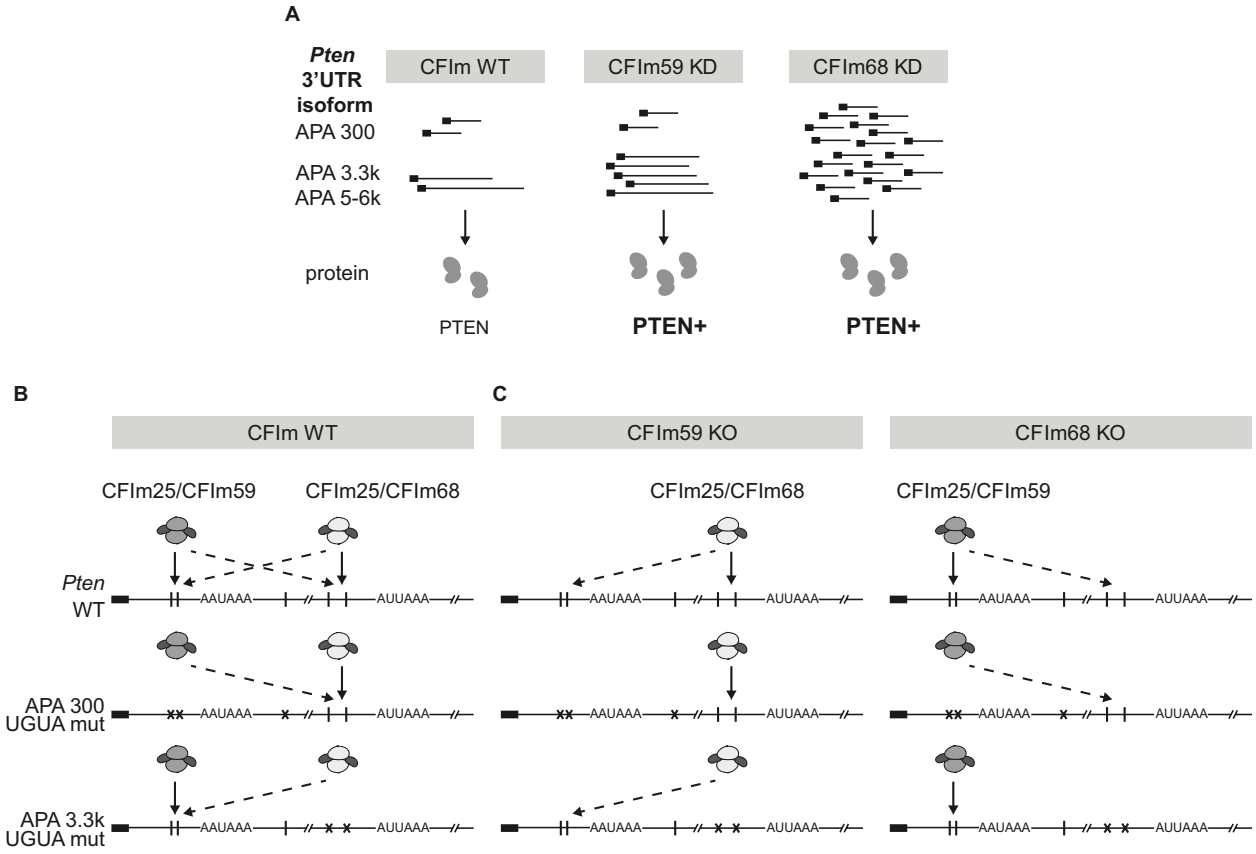


Figure 2.7: Model of *Pten* APA regulation by CFIm. (A) CFIm59 and -68 depletion both upregulate PTEN protein, but through different effects on APA. Because long isoforms (APA 3.3k and APA 5-6k) are better translated than the short, a ~3-fold upregulation of long isoforms upon CFIm59 KD leads to a similar (~1.5-fold) PTEN protein increase as a ~10-fold increase in the short (APA 300) isoform upon CFIm68 KD. (B, C) CFIm25/CFIm59 complex promotes usage of APA 300 nt more efficiently (solid arrow) than APA 3.3k (dashed arrow), whereas CFIm25/CFIm68 is a better promoter of APA 3.3k than APA 300 nt. In the presence of both CFIm59 and -68 (B), loss of UGUA motif around either major PAS leads to APA shift in favor of the other PAS. In the absence of either CFIm59 or -68 (C) the same trend of APA shift occurs with UGUA mutations

due to the partial compensation between the two complexes, but the magnitude depends on the targeting specificity of the complex present.

2.4.2 Breadth of APA regulation in the PI3K/AKT/PTEN axis and cancer

In our datasets, 80% of all genes identified with multiple functional PAS were responsive to depletion of at least one of the CFIm subunits. Several of these genes, including genes with opposing responses to CFIm59 and CFIm68 perturbations, are embedded in the PI3K/AKT/PTEN cascade. This consolidates CFIm as a master regulator of APA but also raises critical questions on the significance of this regulation in cancer. Considering the breadth of APA regulation on transcriptomes and their sensitivity to cellular and physiological changes, it should be expected that the positioning of regulatory elements in 3'UTRs has been harmonized through evolution for a fine-tuned output, whether enhancement or inhibition in the right cells and at the right time. In line with this concept, a previous 3'UTR-seq survey revealed the enrichment of miRNA-binding sites immediately upstream of APA sites in both pro-differentiation and anti-proliferative pre-mRNA sequences, which is expected to lead to a more potent output (Hoffman *et al.*, 2016). As PI3K/AKT/PTEN signalling controls cellular states including proliferation, one could view the outcome of its signaling and the APA regulation among the mRNA of its core components as a feedback network for this crucial cascade.

Landmark papers have highlighted the prevalence of APA dysregulation in cancer (Mayr and Bartel, 2009; Xia *et al.*, 2014; Xiang *et al.*, 2017). A persisting view that originates from those publications is that cancer cells tend to shorten 3'UTRs to 'evade' regulation by miRNAs and/or other trans-acting factors. Our APA analyses in tumor RNA-seq stratifications indicate that the outcome varies in different types of cancers, and loss of any individual CFIm subunit cannot account for the surveyed changes. This can be further inferred from several other publications. For

example, CFIm25 perturbations were reported to contribute to glioblastoma tumor growth *in vivo* (Masamha *et al.*, 2014). A significant correlation was established between worst outcomes (poor survival) in glioblastoma patient and reduced CFIm25 and CFIm59 expression, but not with CFIm68 (Chu *et al.*, 2019). In contrast, greater expression of CFIm59 and CFIm68 has been associated with increased tumor burden in mouse models of liver cancer (Fang *et al.*, 2020; Tan *et al.*, 2021). We reason that ultimately, the selective pressure for or against APA in any gene will very likely depend on the function of the gene itself, on the structure and sequence of its 3'UTRs, but also on the gene regulatory network that prevails in a specific cell fate and state. Consequently, the cellular impacts of *PTEN* APA should be considered in its cellular context and integrated with other APA regulation in the PI3K/Akt cascade. It may be premature to predict the oncogenic significance of any APA regulation or mis-regulation until comprehensive catalogs of dynamic PAS, of their regulatory elements and of the cellular cues they integrate are completed. Beyond the CFIm complex and its subunits, only a handful of APA regulators have been identified, and our study is among the first to detail how specific regulatory elements impact on individual 3'UTRs. Much of this important work on APA lies ahead.

2.5 Data availability

3'UTR sequencing datasets are available at GEO under the accession number GSE183704.

2.6 Acknowledgements

We would like to apologize to authors whose directly related work may have not been cited in this manuscript. We thank Caroline Thivierge for advising in designing and troubleshooting experiments. We thank Dr. Wei Li for his invaluable feedback on the manuscript, as well as Dr. Mathieu Flamand, and Dr. Hamed Najafabadi along with his student Gabrielle Perron for their initial guidance on the bioinformatics analyses.

2.7 Funding

This work was supported by the Canadian Institutes of Health Research (CIHR) grant (MOP-123352) to T.F.D.; personal funding from Ms. Georgette Duchaine, McGill University Rosalind and Morris Goodman Cancer Institute Canderel studentship award, Fonds de recherche du Québec Santé (FRQS) Master's training award, and CIHR Doctoral Research Award to H-W. T.

2.8 Materials and methods

2.8.1 Cell culture and transfection

NIH3T3 cells (ATCC) were cultured in Dulbecco's modified Eagle medium (DMEM) with 10% calf serum (Cytiva). HEK293T cells kindly shared by Dr. Alan Engelman and Dr. Yonsheng Shi labs were cultured in DMEM supplemented with 10% fetal bovine serum (FBS) (Wisent). All cell lines were grown in a humidified incubator at 37 °C with 5% CO_2. Knockdown in NIH3T3 was performed by reverse transfection of siRNAs using Lipofectamine RNAiMAX (ThermoFisher) with a final concentration of 10 μM for 72 hours. The following siRNAs were used: negative ctrl (Qiagen 1027281), CFIm25-si1 (Qiagen SI04414732), CFIm25-si2 (Qiagen SI04414739), CFIm59-si1 (Qiagen SI00858655), CFIm59-si2 (forward: 5'-GUCCUCAUCUCCUCUCUUATT-3', reverse: 5'-UAAGAGAGGAGAUGAGGACTT-3'), CFIm68-si1 (Qiagen SI00958741), and CFIm68-si2 (Qiagen SI00958748). Transient transfection rescue of CFIm59 and CFIm68 in its respective KO cell line was performed using Lipofectamine 2000 (ThermoFisher) following the manufacturer's instruction and collected 24 hrs post-transfection.

2.8.2 RNA isolation and Northern blots

Total RNA from mammalian cell lines was extracted with either the miRNEasy mini kit (Qiagen) or the Monarch total RNA miniprep kit (NEB). Depending on the level of expression, 1-4 µg of total RNA was loaded on 1% agarose-glyoxal gel prepared with NorthernMax-Gly gel prep running buffer (Ambion) and transferred to nylon membranes (BrightStar-Plus, Invitrogen). Probes were synthesized using DECAprime II (Invitrogen) with α-^{32}P dATP, or MEGAscript T3 kit (Invitrogen) with α-^{32}P UTP. Hybridization was performed overnight at 42 or 68 °C in ULTRAHyb hybridization buffer (Ambion). Membranes were then washed and exposed overnight onto imaging plates and imaged on Typhoon phosphorimager (GE). Primers used to amplify both human and mouse *Pten* probes were (TDO1774) 5'-ACCAGGACCAGAGGAAACCT-3' and (TDO1775) 5'- GAATGCTGATCTTCATCAAAAGG-3'.

2.8.3 *Pten* 3'UTR isoform-specific RT-qPCR

The protocol used was developed in (Thivierge *et al.*, 2018). Briefly, cDNA was generated from 500 µg of total RNA using an oligo-dT$_{12}$ anchor primer (TDO679, see Table A1.1) with SuperScript III reverse transcriptase (Invitrogen). To quantify isoforms 300 and 3.3k, first-round amplification (PCR1) was performed with 1:4 diluted reverse transcription (RT) reaction using isoform specific primers (Table A1.1) with a short extension time. The following round of real-time PCR (qPCR2) was performed with 1:1,000 or 1:100,000 diluted PCR1 product using SsoAdvanced Universal SYBR green supermix (Bio-Rad). To quantify the *Pten* 5-6k isoform, total *Pten* (targeting ORF), and Hprt1 internal control, qPCR was performed directly on the 1:4 diluted RT reaction without the PCR1 step. Relative expressions were quantified using the ΔΔCt method and normalized to Hprt1.

2.8.4 Reporter constructs

Mouse *Pten* sequences were cloned into the pcDNA4/TO plasmid, including the coding sequence and the 6.1k 3'UTR from the genomic sequence. Mutagenesis was performed on each targeted UGUA using all-around PCR with non-overlapping primers. Primers used for mutations are detailed in Table A1.1.

2.8.5 Western blotting

Total cell lysates were prepared using lysis buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1% IGEPAL CA-630, 1% sodium deoxycholate, 0.1% sodium dodecyl sulphate, and 1 mM ethylenediaminetetraacetic acid) supplemented with protease and phosphatase inhibitors (Sigma and Roche). Proteins were immobilized on PVDF membranes (Millipore) and probed with the following antibodies diluted in the Odyssey blocking buffer (Li-COR): anti-CFIm68 (Bethyl A301-358A-T, 1:1,000), anti-CFIm25 (Proteintech 10322-1-AP, 1:250), anti-CFIm59 (Sigma HPA041094, 1:500), anti-β-actin (CST 8H10D10, 1:5,000), anti-PTEN (CST 9552, 1:1,000). Secondary IR dye antibodies against mouse or rabbit (Li-COR) were used at 1:10,000. Blots were scanned using the Li-COR imaging system and quantified with the Image Studio Lite suite.

2.8.6 3'UTR-seq and data processing

Using 500 ng of total RNA prepared as described, 3'UTR libraries were generated using QuantSeq 3' mRNA library prep kit REV (Lexogen) following the manufacturer's instructions. Quality control and sequencing of the libraries were performed by the IRIC genomics platform (Université de Montréal) using NextSeq500 SR75 to obtain approximately 20 million reads per sample. Raw reads were preprocessed to remove any adapter sequences using Trimmomatic v0.36 and mapped to mm10 genome with STAR v2.7.2b. Subsequent identification and quantification of PAS peaks were performed according to (Lianoglou *et al.*, 2013). Briefly, reads from all samples were merged

for peak calling using CLIPanalyze (https://bitbucket.org/leslielab/clipanalyze). Internal priming events were then removed along with low-usage events to compile a curated list of valid APA sites. This list was then referenced to count APA events in each sample using featureCounts (Liao *et al.*, 2013). Differential expression of each APA event between conditions was performed through repurposing the DEXSeq package (Anders *et al.*, 2012) originally developed for the analysis of differential exon expression.

For RED score analyses, genes with only one identified PAS were first filtered out. The top two expressing PAS across all samples of each remaining gene were identified and defined as the proximal or distal site depending on their relative position. Responsive genes are defined as $|RED| \geq \log_2(1.5)$.

2.8.7 Polysome profiling

NIH3T3 cells (2 x 15 cm plates) were transfected with siRNA targeting CFlm59, CFlm68 or non-targeting control siRNA for 72 hrs and harvested. Confluency at harvesting was ~80%. Polysome profiles were generated as previously described (Gandin *et al.*, 2014). Cells were pre-treated with 100 µg/mL cycloheximide for 5 min and scraped in ice-cold, cycloheximide supplemented (100 µg/mL) PBS, using rubber scrapers. Cells were then pelleted in 15 mL conical tubes at 240 x g for 5 min at 4°C, and lysed on ice in 500 µL of hypotonic lysis buffer (5 mM Tris HCl, pH 7.5, 2.5 mM $MgCl_2$, 1.5 mM KCl, 100 µg/mL cycloheximide, 1 mM DTT, 0.5% Triton, 0.5% sodium deoxycholate) for 15 min. Subsequently, lysates were cleared through a 15 min, 20,817 x g centrifugation at 4°C. Cleared lysates were diluted with hypotonic lysis buffer to set their optical density (OD) at 260 nm to 10-20. Sucrose gradients (5-50%) were generated in polypropylene tubes (Beckman Coulter, 331372, 14 x 89 mm) using a gradient maker (Biocomp Gradient Master 108), according to the manufacturer's instructions. A volume of 500µl was removed from the top

of the sucrose gradient to allow for sample loading. Lysates were then carefully layered over the top of linear sucrose gradients and centrifuged at 4°C for 2 hrs at 222,228 x g (36,000 rpm) using an ultracentrifuge (Beckman Coulter, Optima XPN-80, rotor SW41Ti), while 10% of the sample was kept as input. Ultracentrifuged samples were fractionated into 2 mL tubes using a density gradient fractionation system (Brandel, BR-188-177), resulting in 12 fractions of 800 µL each. Input samples, as well as sucrose gradient fractions, were mixed with 1000 µL of TRIzol LS reagent (Ambion) and kept at -80°C for subsequent RNA isolation.

2.8.8 mRNA stability assay

NIH3T3 cells were seeded in 6-well plates and treated in a time course (0-360 min) with Actinomycin-D (5 µg/mL) at 70-80% confluency. Cells were then harvested in QIAZol (Qiagen) for total RNA extraction using the Monarch total RNA miniprep kit (NEB). Quantification of each *Pten* 3'UTR isoform was then performed using *Pten* 3'UTR isoform-specific RT-qPCR as described before.

2.8.9 Generation of KO cells

CRISPR knockout of CFIm59 and CFIm68 was performed in NIH3T3 cells using the two vector lentiviral system (Sanjana *et al.*, 2014). Cells were first infected with lentivirus generated using lentiCas9-Blast (Addgene plasmid #52962), selected in blasticidin, and subsequently isolated as single colonies. The selected clone with high Cas9 expression underwent a second round of infection by lentivirus generated using lentiGuide-puro (Addgene plasmid #52963) and single cell isolated following puromycin selection. Target guide sequences are: 5' agtgtacacaacgtttggtg 3' (CFIm68) and 5' acgctcatcattggcgactc 3' (CFIm59).

2.8.10 Analysis of PAR-CLIP and motif enrichment

Published PAR-CLIP dataset of CFIm59 and CFIm68 (Martin *et al.*, 2012) were subsetted into sequence regions corresponding to different set of genes responsive or non-responsive to CFIm59 and CFIm68 KO. Signal matrices of the selected 400 nt regions centered around PAS cleavage sites were then computed and plotted using deepTools (Ramírez *et al.*, 2016). Motif enrichment on 400 nt long sequences centered around PAS cleavage sites of different sets of mRNAs was performed with STREME (Bailey, 2021) using default settings. One significant hit was obtained as a relative motif enrichment of a query sequence set to a user-input control sequence set. No significant hit was obtained using the program generated scramble sequence set as the control.

2.8.11 Cancer APA analysis

Integrated analysis of mRNA expression and PDUI (Figure 2.6B, C) include the following cancer types (TCGA abbreviations): ACC, BLCA, BRCA, CESC, CHOL, COAD, DLBC, ESCA, GBM, HNSC, KICH, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC, UCS, UVM, and uncategorized cancers.

2.8.12 Statistical analyses

Plotted quantification of western blots, northern blots and RT-qPCR are presented as mean ± standard deviation unless otherwise indicated. Statistical tests were performed using one-way Analysis Of Variance (ANOVA) with post-hoc Dunnett's test between all treatment and control groups, unless otherwise indicated. All plots were generated using the gglot2 package (Wickham, 2016) in R.

Chapter 3:

Transcriptome-wide assessment for the impact of 3'UTR shortening on mRNA stability

Hsin-Wei Tseng[1,2], Thomas F. Duchaine[1,2,*]

[1] Rosalind and Morris Goodman Cancer Institute, McGill University, Montréal, H3G1Y6, Canada.

[2] Department of Biochemistry, McGill University, Montréal, H3G1Y6, Canada.

* Correspondence: thomas.duchaine@mcgill.ca

3.1 Abstract

The intricate balance between messenger RNA (mRNA) transcription and decay is central to the precise maintenance and regulation of all cellular processes. Determinants of mRNA stability, or *cis*-regulatory elements, are enriched in the sequence and structural features of the 3'-UnTranslated Region (3'UTR). Regulation for the length of the 3'UTR, a process known as alternative polyadenylation (APA), further dictates the inclusion or exclusion of these *cis*-regulatory elements. Disruptions of APA through regulators such as the CFIm complex have been linked with diseases including cancer, but the effect on mRNA stability has not been systematically studied. Here we characterize the transcriptome-wide effect of APA dysregulation on mRNA stability by leveraging a cell panel bearing a knockout of CFIm68, a subunit of CFIm, which exhibits an overall 3'UTR shortening. Upon CFIm68 knockout, we observed a surprisingly modest change in the mRNA stability, and this overall change is uncorrelated with APA shift. Nonetheless, APA shift and stability change are significantly correlated for a limited subset of transcripts which express 3'UTR isoforms with significantly different stability. For these transcripts, 3'UTR sequence analysis highlights transcript-specific impact for the enrichment of *cis*-elements, including miRNA binding sites and GC content, which may govern transcript stability. Finally, we identify a putative novel mechanism involving CFIm68 as a regulator of mRNA stability through modulation of the exon junction complex (EJC) dependent nonsense-mediated decay (NMD) pathway.

3.2 Introduction

The metabolism and functions of messenger RNAs (mRNAs) are tightly regulated at every step along their life cycles. *cis*-regulatory elements, preferentially situated in the 3' untranslated region (3'UTR), unbound from the sequence constraints bearing on the coding region, can facilitate post-transcriptional regulation by governing over mRNA translation, localization, and stability. The AU-rich elements (ARE), for instance, are targeted by RNA-binding proteins (RBPs) such as HuR (also known as ELAVL1), TTP, KHSRP, and others that can competitively or collaboratively regulate mRNA translation and stability under different circumstances (Cammas *et al.*, 2014; Chen *et al.*, 2001; Lykke-Andersen and Wagner, 2005). Perhaps the most characterized and abundant *cis*-elements in the 3'UTRs are microRNA (miRNA) binding sites, targeted by the miRNA-induced silencing complex (miRISC). Efficient miRISC binding directs transcript silencing through mRNA translation repression and decay (Filipowicz *et al.*, 2008). Other less defined *cis*-elements encoded by the 3'UTR sometimes act as localization signals for mRNAs, and in some cases for the encoded protein product, towards distinct spatial domains of polarized cell types such as neurons (An *et al.*, 2008; Taliaferro *et al.*, 2016). For some transcripts, separate expression of 3'UTR fragments which altogether cover the full-length 3'UTR sequence fails to recapitulate the properties seen with the full-length 3'UTR, suggesting that cooperative or synergistic action among *cis*-elements of the 3'UTR contribute to their regulatory roles (Besse *et al.*, 2009; Kristjánsdóttir *et al.*, 2015).

The length of 3'UTRs can additionally be regulated through alternative polyadenylation (APA). During APA, alternative polyadenylation signals (PAS), canonically the AAUAAA motif (Wickens and Stephenson, 1984), are used for transcript 3' end maturation that involves the two-step process of cleavage and polyadenylation (CPA). APA is estimated to occur for more than half

of all human and mouse genes to generate alternative mRNA 3'UTR isoforms, which give rise to identical proteins (Gruber *et al.*, 2016; Lianoglou *et al.*, 2013). Selective inclusion and exclusion of cis-elements in the 3'UTR can thus be fine-tuned through APA and contribute to context-specific mRNA and protein expression. Consistently, tissue and disease-specific APA patterns have been observed (Derti *et al.*, 2012; Xia *et al.*, 2014).

A growing number of APA regulators have been characterized over the years, with some affecting global APA while others impacting on specific genes. Among the best characterized is the mammalian cleavage factor I (CFIm) complex, consisting of three members: CFIm25, CFIm68, and CFIm59, forming heterotetrameric complexes composed of a homodimer of CFIm25 and an alternative homodimer of CFIm68 or CFIm59 (Yang *et al.*, 2011). Sequence specificity of CFIm is conferred by CFIm25 which recognizes and binds UGUA motifs in proximity to the PAS to facilitate cleavage and polyadenylation (Yang *et al.*, 2010; Zhu *et al.*, 2018). Depletion of either CFIm25 or CFIm68 results in a widespread shortening of 3'UTRs (Martin *et al.*, 2012; Zhu *et al.*, 2018), whereas CFIm59 depletion has a more selective but overall opposite effect (Tseng *et al.*, 2022). Pathological expression of CFIm25 have also been associated with diseases such as glioblastoma and neuropsychiatric disorders (Gennarino *et al.*, 2015; Masamha *et al.*, 2014). Notwithstanding this growing evidence for APA dysregulation in diseases, the functional relevance of observed coordinated global APA shifts remains unclear.

A frequent assumption is that longer 3'UTRs are less stable as they potentially harbor more destabilizing elements such as miRNA binding sites. However, this assumption is largely based on predicted, but unvalidated miRNA binding sites. Despite the prevalence of this assumption, anecdotal evidence supports that it is incorrect, as shorter mRNA 3'UTR isoforms can be either more or less stable than the longer isoforms (Mayr and Bartel, 2009; Thivierge *et al.*, 2018; Tushev

et al., 2018). The few studies that investigated the relationship between 3'UTR length and mRNA stability across transcriptomes have also reached inconsistent conclusions. For example, one study concluded with a negative correlation, while another reported no significant correlation between 3'UTR length and mRNA stability (Sharova *et al.*, 2009; Yang *et al.*, 2003). A subsequent study comparing the stability of 3'UTR isoforms expressed by each gene product concluded on a limited effect APA shift can exert on transcript stability, while another found ~8 times more genes expressing a more stable short 3'UTR isoform (Spies *et al.*, 2013; Wang *et al.*, 2018). An important limitation is that, while those studies aim to infer the effect of global APA disruption on mRNA stability changes, they were conducted at steady-state, without experimental APA perturbation. As such, the consequences and changes in mRNA stability due to a global APA shift remain to be directly investigated.

In this study, we employed transcriptome-wide perturbation of APA by knocking out CFIm68 to induce global shortening of the 3'UTRs. Systematic comparison of individual transcripts as well as all 3'UTR isoforms expressed per transcription unit revealed an *on average* modest effect CFIm68 depletion exerts on stability, and an overall highly variable relationship between APA shift and stability change. Within the set of transcription units expressing 3'UTR isoforms with significantly different stability, sequence feature analysis indicated a highly gene-specific enrichment of stability determinants. Lastly, we uncovered a novel role for CFIm68 as a potential direct regulator of factors in the exon junction complex (EJC) dependent nonsense-mediated decay (NMD) pathway. Together, our analyses suggest a complex and context-dependent relationship between APA regulators, APA, and the overall mRNA stability.

3.3 Results

3.3.1 3'UTR Decay-seq recapitulates transcriptome-wide mRNA stability

To interrogate transcript stability changes at the transcriptome-wide scale, we performed 3'UTR-seq on a time-course of Actinomycin-D treated mouse fibroblast NIH3T3 cells collected at 0-, 30-, 150-, and 360-minute time points. 3'UTR-seq captures the precise 3' cleavage site of every transcript to profile APA for each transcription unit (TU). By collapsing read counts for all 3'UTR isoforms mapped to one TU we can additionally quantify the total transcript expression for each TU. After Actinomycin-D treatment, the rate of transcript decay, reflected by the slope of transcript counts over time (herein referred to as the Stability Coefficient), was then calculated as a proxy of stability for every transcript (using individual transcript counts) and every TU (using collapsed transcript counts per TU. Herein referred to as the TU stability). The bigger the Stability Coefficient, the more stable the individual transcript or TU is. Using this method that we termed 3'UTR Decay-seq, by comparing CFIm68 knockout (KO) with wildtype (WT) cells, we were able to infer differential changes in total transcript (or TU) expression and transcript (or TU) stability as a response to APA shifts (Figure 3.1A).

As a control for this method, we could recapitulate the relative stability of TUs with known fast (Myc), medium (Pten), and slow (Hprt) transcript turnover (Figure 3.1B) (Thivierge *et al.*, 2018). Similarly, we could distinguish the relative stability of different 3'UTR transcript isoforms, faithfully recapitulating the greater stability of longer *Pten* isoforms (3.3k and 5.4k nt after the stop codon) compared to the short (300 nt) (Figure 3.1C), as characterized in previous studies (Thivierge *et al.*, 2018; Tseng *et al.*, 2022). Lastly, using collapsed transcript counts per TU, the overall TU stability could be quantified in each of WT and CFIm68 KO conditions (Figure 3.1D).

Together, these results validate the effectiveness of 3'UTR Decay-seq in mRNA stability quantification at both per-transcript and per-TU level.

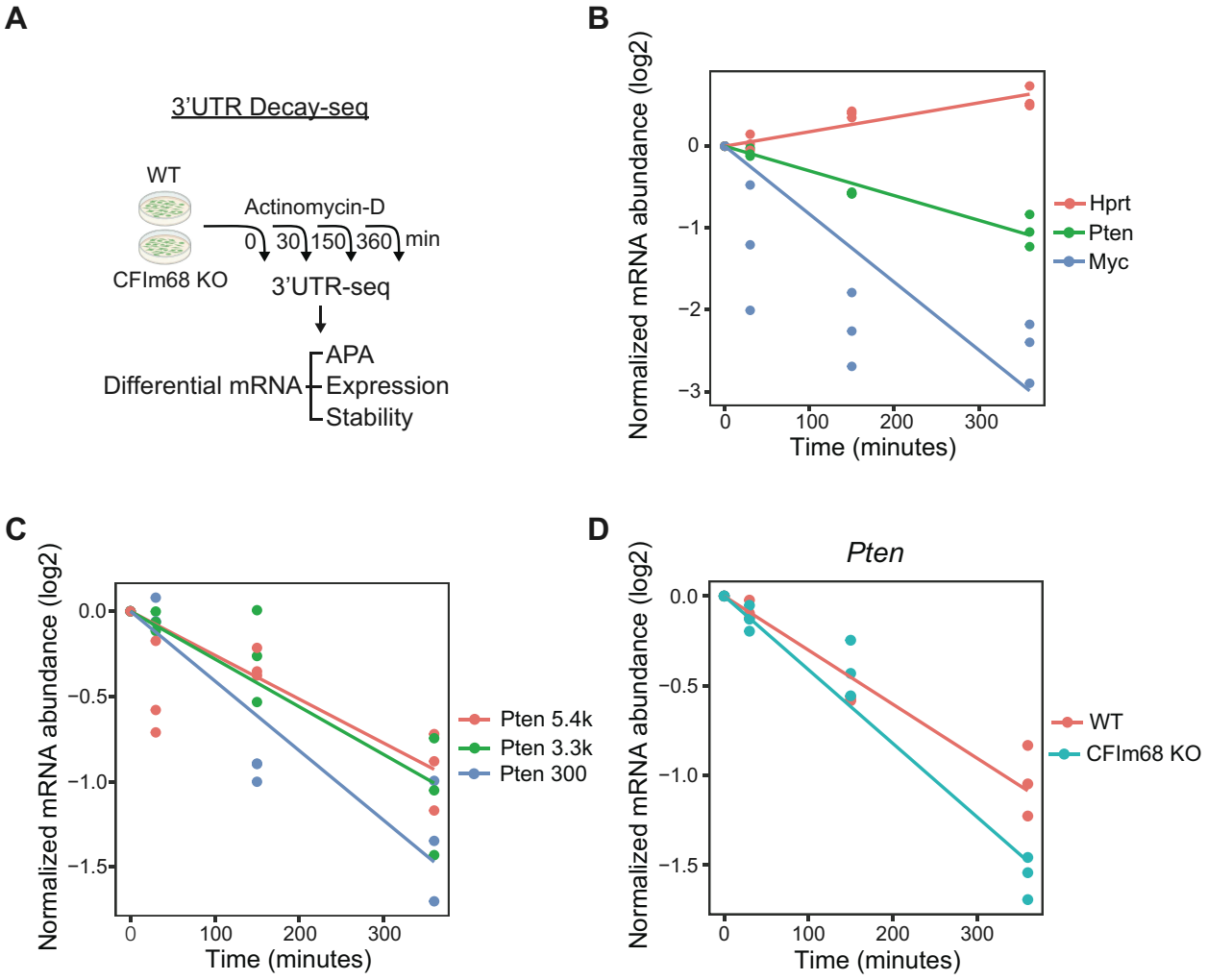
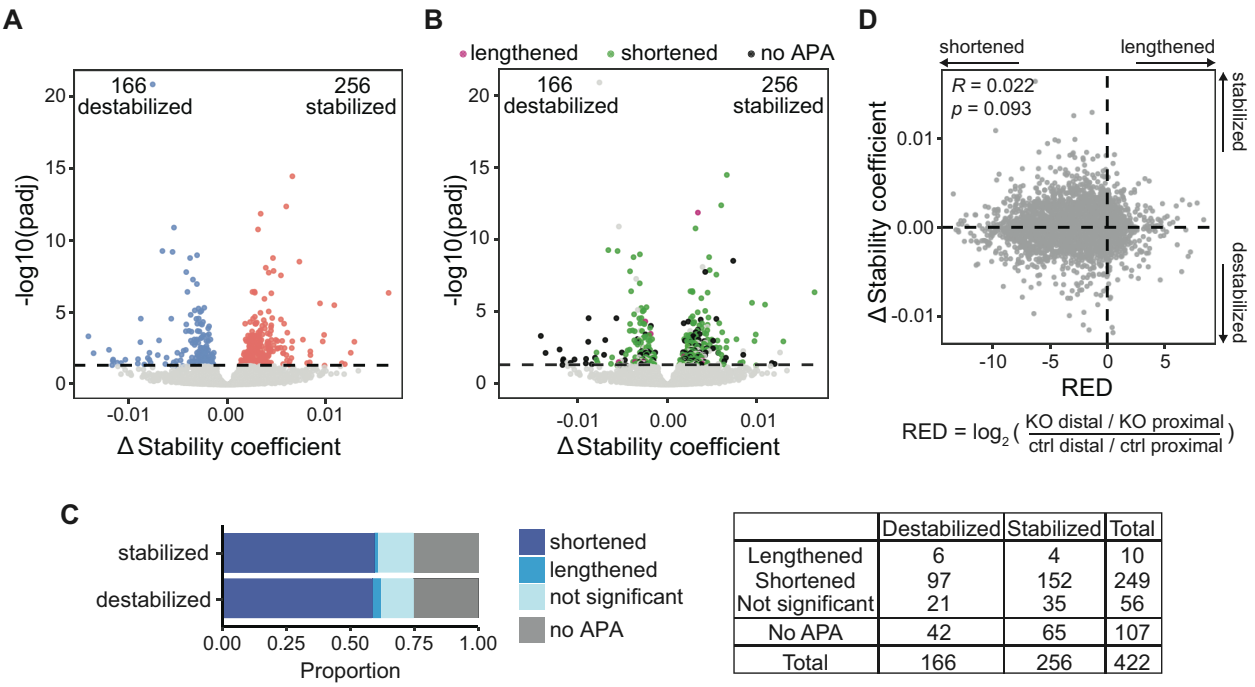


Figure 3.1: 3'UTR Decay-seq recapitulates transcriptome-wide mRNA stability. (A) Schematic of 3'UTR Decay-seq design. Wildtype (WT) and CFIm68 KO NIH3T3 cells were treated with Actinomycin D, collected at 0-, 30-, 150-, and 360-minute time points, and subjected to 3'UTR-seq. The sequencing data was then used to infer changes in mRNA APA, expression, and stability. (B) The relative stability of fast- (*Myc*), mid- (*Pten*), and slow-decaying (*Hprt*) transcription units (TUs) plotted as the logarithm of transcript counts over time. (C) *Pten* relative stability for the longer (3.3k and 5.4k after stop codon) and short (300) 3'UTR isoforms. (D) *Pten* TU stability change upon CFIm68 KO inferred by collapsing expression of all *Pten* 3'UTR isoforms.

3.3.2 CFIm68 depletion induces a global 3'UTR shortening that does not correlate with the average changes in TU stability

To assess changes in TU stability upon global 3'UTR shortening, we performed 3'UTR Decay-seq in CFIm68 KO and WT cells. APA is quantified by the Relative Expression Difference (RED) of predefined distal (long) and proximal (short) 3'UTR isoforms in KO over WT cells (see Materials and Methods) (Li *et al.*, 2015). From the 5849 detected TUs expressing multiple 3'UTR isoforms (referred to as APA genes), 4061 (69%) TUs expressed differential APA, and 3821 (65%) had significantly shorter 3'UTRs (RED < 0) while 240 (4%) had significantly longer 3'UTRs (RED > 0) in CFIm68 KO compared to WT cells (Figure A2.1A). CFIm68 KO also resulted in 166 destabilized (Δ Stability coefficient < 0) and 256 stabilized (Δ Stability coefficient > 0) TUs (Figure 3.2A), indicating that the majority of TUs affected by differential APA had no significant change in their stability. From the 422 TUs exhibiting differential stability, 315 (71%) were APA genes while the other 107 (29%) had no detectable alternative 3'UTR isoform in our dataset (no APA) (Figure 3.2B, C). From the 256 stabilized TUs, 4 (1.5%) were lengthened and 152 (59%) were shortened; among the 166 destabilized TUs, 6 (3.6%) were lengthened and 97 (58%) were shortened (Figure 3.2B, C). When plotted across all TUs, upon CFIm68 KO, changes in stability also had no correlation with changes in APA (Figure 3.2D). This lack of correlation persists even after stratifying the TUs according to their shift in APA direction upon CFIm68 KO (significantly lengthened, significantly shortened, or not significant) (Figure A2.2).



	Destabilized	Stabilized	Total
Lengthened	6	4	10
Shortened	97	152	249
Not significant	21	35	56
No APA	42	65	107
Total	166	256	422

Figure 3.2: CFIm68 depletion induces a global 3'UTR shortening that does not correlate with the average changes in TU stability. (A) Volcano plot of transcription units (TUs) significantly stabilized (red) and destabilized (blue) upon CFIm68 KO. (B) Same plot as (A) with points colored according to differential APA upon CFIm68 KO. Green marks TUs that are shortened, magenta for TUs that are lengthened, and black for TUs with no detectable differential APA. (C) Proportion and TU counts for stabilized and destabilized TUs. Left, stacked bar plot with different shades of blue for TUs expressing differential APA (shortened, lengthened, and not significant), and gray for TUs that do not express APA isoforms (no APA). Right, counts of TUs for each group of stability and APA change. (D) Correlation between APA and stability changes for each TU upon CFIm68 KO. Pearson's *R* and *p*-value are displayed.

Moreover, we observed that CFIm68 KO-induced 3'UTR shortening had no correlation with the averaged changes in mRNA expression level across shortened TUs (Figure 3.3A), reflecting the highly variable changes in the stability. However, when stratified according to their stability

change upon CFIm68 KO, an increase in 3'UTR shortening correlated with a decrease in mRNA expression for destabilized TUs, while the opposite correlation was also seen for stabilized TUs (Figure 3.3B).

Overall, these results demonstrate that CFIm68 KO impacts the stability only for a limited subset of TUs but overall and on average stability changes have no clear correlation with the direction nor the magnitude of APA shift.

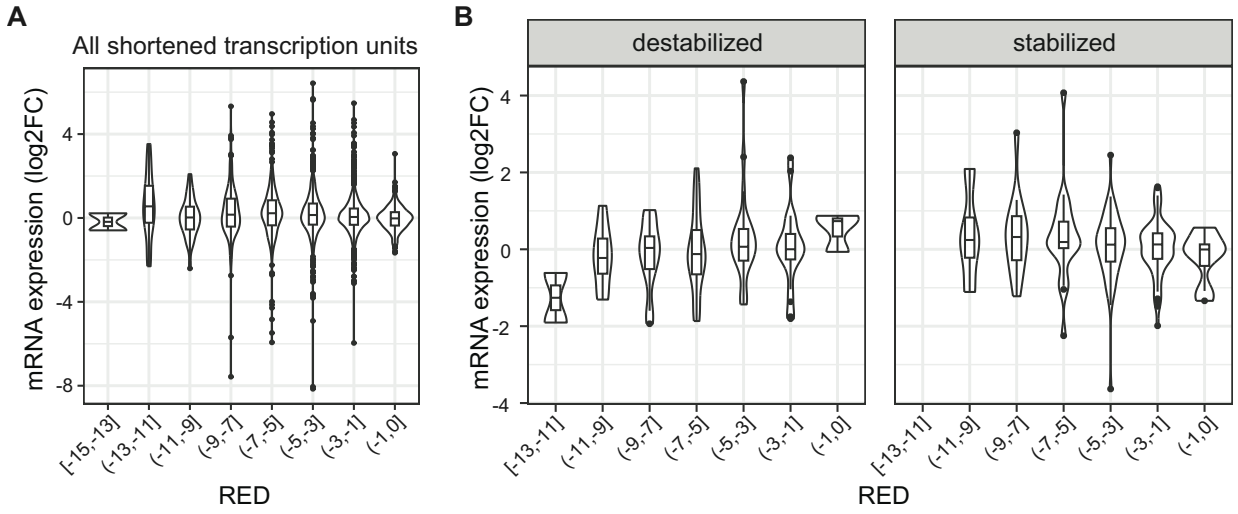


Figure 3.3: CFIm68 KO changes mRNA expression for a limited subset of TUs depending on its impact on TU stability. The changes in mRNA expression (in log2 fold change) and APA (in RED scores, binned) are plotted across all shortened transcription units (TUs) in (A), and across destabilized or stabilized TUs in (B).

3.3.3 Condition-dependent correlation between 3'UTR length and transcript stability

The restricted effect that global 3'UTR shortening has on mRNA stability raises the question of whether and to what extent the length of 3'UTR predisposes the stability of the transcript. To investigate this relationship across all expressed transcripts, we first eliminated outlier transcripts by restricting our analysis to transcripts with up to 5500 nt-long 3'UTRs, which represent ~95% of all captured transcripts (Figure A2.3). By plotting the Stability Coefficient of every transcript

against its 3'UTR length, we revealed a moderate but significant negative correlation in the WT cells (Pearson's R = -0.13, p < 2.2e-16), meaning that there is a moderate decrease in the transcript stability as 3'UTRs become longer (Figure 3.4A, top). However, this correlation was attenuated in CFIm68 KO cells (Pearson's R = -0.074, p < 2.2e-16) (Figure 3.4A, bottom). Interestingly, when comparing the stability of transcripts grouped by similar 3'UTR lengths (difference within 500 nt), despite the lack of statistical significance for most groups, there was a trend of transcript stabilization in CFIm68 KO cells compared to WT cells for transcripts with longer 3'UTRs (over 2000 nt) (Figure 3.4B).

Together, these results demonstrate that the correlation between 3'UTR length and transcript stability depends on the cell state and is sensitive to perturbations. As mentioned above, CFIm68 KO also impacts the stability for a subset of transcripts depending on the length of the 3'UTR.

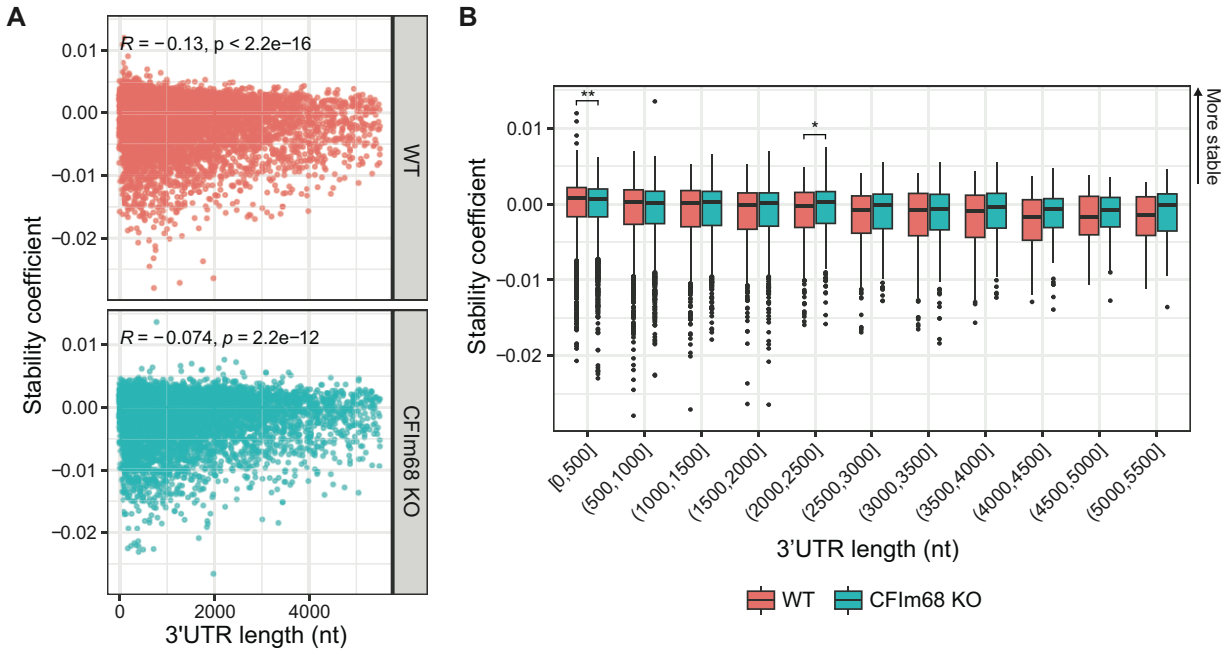


Figure 3.4: Condition-dependent correlation between 3'UTR length and transcript stability. The stability coefficient is plotted against the 3'UTR length for each transcript in wildtype (WT) and CFIm68 KO cells. Every transcript is represented by a point in (A). Pearson's R and p-values

are displayed. In (B) the conditions (WT and CFIm68 KO) are plotted side by side for comparison. The transcripts are grouped every 500 nt according to their 3'UTR length. Wilcox rank sum test was performed. (*$P \leq 0.05$, **$P \leq 0.01$)

3.3.4 The impact of CFIm68 knockout on TU stability depends on 3'UTR-specific features

To understand the impact of APA disruption on TU stability we investigated isoforms originating from the same TU. To this end, we compared the Stability Coefficient of paired distal and proximal isoforms for each TU under the WT condition. This led to a highly positive correlation (Pearson's $R = 0.72$, $p < 2.2e\text{-}16$) that was also symmetric (Figure 3.5A). This indicates that for the 5849 APA genes studied here, their distal and proximal 3'UTR isoforms had largely indistinguishable stability (5156/5849, 88%).

Nonetheless, two groups of TUs stood out; one group had a significantly more stable distal isoform than the proximal (255/5849 TUs, 4.3%, referred to as *dist-stable* TUs), while the other was characterized by a proximal isoform more stable than the distal (434/5849 TUs, 7.4%, referred to as *prox-stable* TUs) (Figure 3.5A, B). We defined proximal and distal isoforms originated from the same TU as sharing the coding sequence (CDS) and a constitutive 3'UTR (cUTR), and which differ solely in the alternative 3'UTR (aUTR) region (Figure 3.5C). As such, we investigated how the aUTR might differ in length and the differences in the encoded sequence(s) between the prox-stable and dist-stable TUs that could contribute to the differences in relative isoform stability.

Length of the aUTR alone was not a determinant of transcript stability, as there was no significant difference between the median aUTR length of dist-stable (1486 nt) and prox-stable TUs (1255 nt) (Figure 3.5D). The length of the aUTR, however, did correlate positively with the magnitude of APA shift upon CFIm68 KO, in both lengthened and shortened TUs (Figure A2.4), as previously reported (Li *et al.*, 2015).

Further analysis into the sequences of aUTR between the two groups of TUs revealed that aUTR of prox-stable TUs harboured on average significantly more predicted and conserved miRNA binding sites predicted using TargetScan (Friedman *et al.*, 2009) (prox-stable: median 5.8; dist-stable: median 1.6, sites per kb) (Figure 3.5E). We hypothesize that at least some of those predicted miRNA-binding sites may contribute to the destabilization of distal isoforms for the prox-stable TUs. Conversely, the aUTR of dist-stable TUs tended to be more GC-rich, suggesting the possible formation of more stable secondary structures (prox-stable: median 41%; dist-stable: median 45%) (Figure 3.5F). To further substantiate this possibility, we performed *in silico* RNA structure modelling of the aUTRs using RNAfold (Hofacker, 2003; Lorenz *et al.*, 2011). In line with the higher GC content in the aUTR of dist-stable TUs, the length-adjusted minimum free energy of folding was significantly lower in these regions, suggesting that aUTR of dist-stable TUs favoured the formation of RNA secondary structures (prox-stable: median -0.27; dist-stable: median -0.29, kcal/mol per nt) (Figure 3.5G). Lastly, as RNA turnover machineries are well conserved across eukaryotes, and coevolve with their corresponding RNA *cis*-regulatory elements (Cheng *et al.*, 2017; Heck and Wilusz, 2018), we further explored the relative sequence conservation of aUTRs using the phyloP (phylogenetic p-value) method (Hubisz *et al.*, 2010; Pollard *et al.*, 2010). Interestingly, per-base conservation across 60 vertebrates revealed that aUTRs of prox-stable TUs are better conserved than aUTRs of dist-stable TUs (prox-stable: median 0.35; dist-stable: median 0.19, phyloP score) (Figure 3.5H). This observation partially reflects on the concentration of conserved miRNA binding sites in the aUTRs of prox-stable TUs.

Lastly, upon global 3'UTR shortening by CFIm68 KO, we observed an overall positive correlation between stability changes and RED score among dist-stable TUs, and a negative correlation for prox-stable TUs (Figure 3.5I). For instance, for dist-stable TUs like *Thap1*, *Dr1*,

and *Grpel2*, the overall stability decreased as the 3'UTRs became shorter. Conversely, for prox-stable TUs like Wee1, Pum2, and Tle4, shortened 3'UTRs increased stability.

Taken together, our results indicate that 3'UTR length is not a reliable predictor of transcript stability. Regulation of RNA stability is highly dependent on the individual mRNA embedded in gene regulatory networks. Some alternative 3'UTRs of mRNAs form stable secondary structures that may prevent interactions with *trans*-acting factors, while others encode well-conserved *cis*-regulatory elements such as miRNA-binding sites.

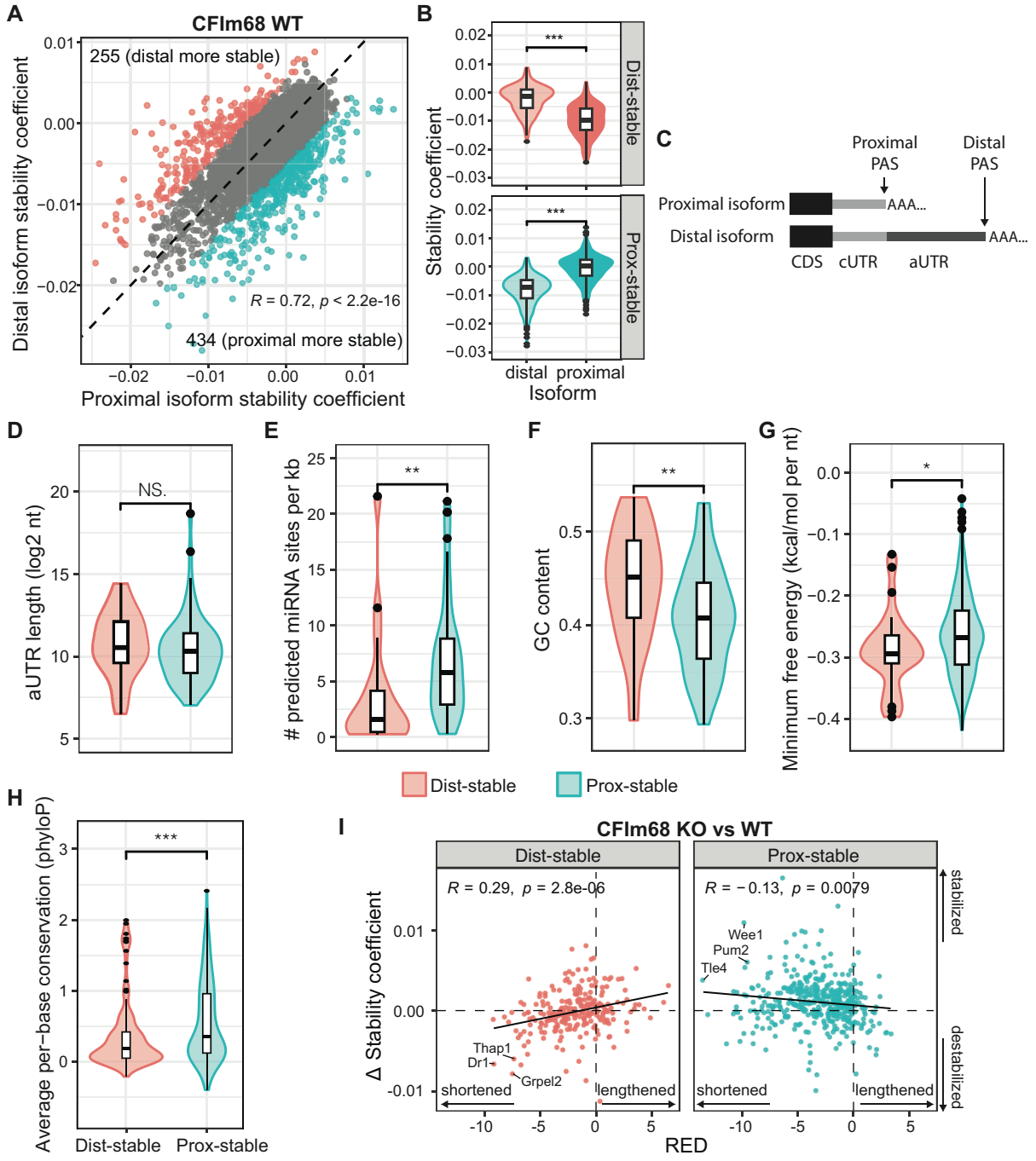


Figure 3.5: The impact of CFIm68 knockout on TU stability depends on 3'UTR-specific features. (A) Correlation of distal and proximal 3'UTR isoform stability coefficients for all transcription units (TUs) expressing APA isoforms. Throughout all panels, pink is used for TUs expressing a more stable distal isoform than the proximal (dist-stable), while green is used for TUs expressing a more stable proximal isoform than the distal (prox-stable). Pearson's *R* and *p*-

values are displayed. (B) The stability coefficients of the distal and proximal 3'UTR isoforms are compared for dist-stable and prox-stable TUs. (C) mRNA proximal and distal 3'UTR isoforms produced by the usage of alternative polyadenylation signals (PAS) share the coding sequence (CDS) and constitutive 3'UTR (cUTR), while differing in the alternative 3'UTR (aUTR) sequence. (D-H) Different attributes of the aUTR are compared between dist-stable and prox-stable TUs: (D) aUTR length distribution, (E) number of predicted miRNA binding sites, (F) GC content, (G) minimum free energy in folding calculated by *in silico* modelling of RNA secondary structures, and (H) per-base sequence conservation. (I) Changes in stability coefficient against changes in APA upon CFIm68 KO for dist-stable and prox-stable TUs. Example TUs for each group are indicated. Pearson's *R* and *p*-values are displayed. Unless otherwise indicated, the *P*-values were determined using Wilcoxon Signed Rank Test (*** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$, NS. not significant).

3.3.5 A possible role for CFIm68 in the regulation of mRNA stability through EJC/NMD

Depletion of the master APA regulator CFIm68 alters the stability of a subset of TUs undergoing APA. However, approximately one-quarter of all TUs with altered stability upon CFIm68 KO did not express any alternative 3'UTR isoform (Figure 3.2B, C), which suggests that CFIm68 can control mRNA stability without a direct impact on mRNA APA. This led us to hypothesize that component(s) of mRNA turnover may lie among APA genes under direct regulation by CFIm68. We performed a Gene Set Enrichment Analysis (GSEA) on differentially expressed genes upon CFIm68 depletion, using all pathways in the Canonical Pathways dataset retrieved from MSigDB (Liberzon *et al.*, 2011; Subramanian *et al.*, 2005). Interestingly, among the pathways that were most significantly enriched or depleted were mRNAs associated with ribosomes, translation, and nonsense-mediated mRNA decay (NMD) (Figure 3.6A).

NMD is a major mRNA turnover pathway that eliminates truncated transcripts, but the pathway is also known to regulate a subset of normal (non-truncated) transcripts (Hug *et al.*, 2016; Yi *et al.*, 2021). Among the routes of NMD activation, the most potent and best-studied mechanism is coupled to the exon junction complex (EJC) (Kurosaki *et al.*, 2019). Using datasets from both CFIm68-depleted mouse NIH3T3 and human HEK293T cells (Gruber *et al.*, 2012), we compared the differential expression of mRNAs annotated as NMD pathway components. The analysis revealed that among the most robustly and consistently downregulated mRNAs were core EJC members, including *Rbm8a*, *Magoh*, *Eif4a3*, *Rnps1*, and *Casc3* (Figure 3.6B). Crucial positive regulators and effectors of NMD such as *Smg1*, *Smg5*, and *Smg7*, were downregulated upon CFIm68 depletion. Notably, *Smg8*, a negative regulator of Smg1 kinase activity (Usuki *et al.*, 2013), was consistently upregulated. These trends were also observed upon CFIm25 knockdown, which also elicits global 3'UTR shortening akin to CFIm68 depletion. This was not universal across all components of NMD; aside from *Upf3a*, we did not detect the same consistent trend for other functional subunits including *Upf1*, *Upf2*, and *Upf3b* (Figure 3.6B).

We next turned to search for target mRNAs among the NMD/EJC pathway that may be under direct APA regulation by CFIm68. Two mRNAs, *Rbm8a* and *Smg1*, expressed alternative 3'UTR isoforms, and depletion of CFIm68 in HEK293T cells shifted their expression towards the proximal 3'UTR isoform (Figure 3.6C, 3'UTR-seq tracks). Interestingly, PAR-CLIP analyses on CFIm25 and CFIm68 proteins (Martin *et al.*, 2012) further revealed cross-linking clusters near UGUA sites surrounding the distal 3'UTR end of both *SMG1* and *RBM8A*, as well as at the proximal site of *SMG1* (Figure 3.6C, PAR-CLIP tracks). These results support a model wherein CFIm25 and CFIm68 control *Smg1* and *Rbm8a* mRNAs APA through direct binding to alternative PAS regulatory sequences.

Lastly, we searched our CFIm68 depletion datasets for significantly upregulated or stabilized mRNAs previously identified as putative NMD targets (Colombo *et al.*, 2017; Gangras *et al.*, 2020; McIlwain *et al.*, 2010) but found no enrichment. We reason that the newly identified pathway may be dictated by distinct determinants of specificity.

Taking these results together, here we propose a novel route of mRNA stability regulation by CFIm68 via direct APA regulation of factors in the EJC and NMD pathways (Figure 3.7).

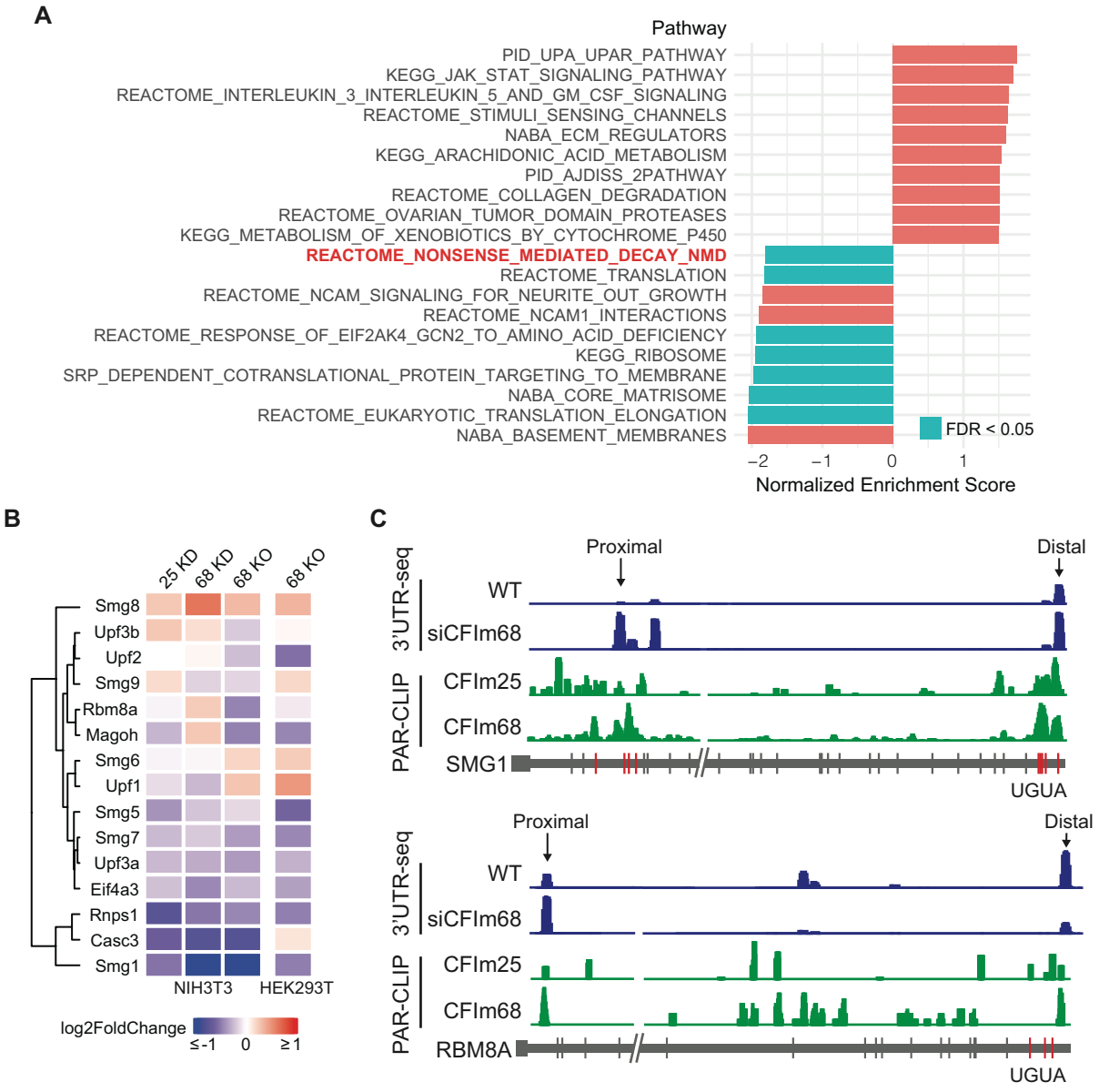


Figure 3.6: A possible role for CFIm68 in the regulation of mRNA stability through EJC/NMD. (A) Gene Set Enrichment Analysis (GSEA) of differentially expressed mRNAs upon CFIm68 KO using annotated Canonical Pathways from MSigDB. Green marks for significantly enriched or depleted pathways with a false discovery rate (FDR) cut-off of 0.05, and red for non-significant pathways. (B) Heatmap for core factors in the NMD and EJC machineries upon depletion of CFIm25 or CFIm68 in mouse NIH3T3 or human HEK293T cell lines. (C) Genome tracks of 3'UTR-seq (dark blue) and PAR-CLIP (green) for SMG1 (top) and RBM8A (bottom) in HEK293T. UGUA sites are indicated and the ones directly overlapping PAR-CLIP signals surrounding distal and proximal PASs are colored in red.

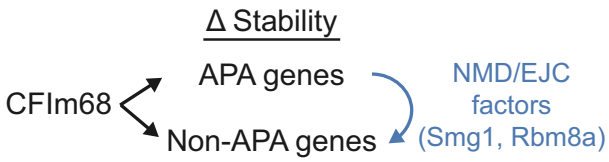


Figure 3.7: Model for the regulation of mRNAs by CFIm68. Depletion of CFIm68 can alter the stability of mRNAs through the direct regulation of APA. CFIm68 depletion also affects the stability of mRNAs without APA, through direct APA regulation of factors in the NMD and EJC machineries such as Smg1 and Rbm8a.

3.4 Discussion

This study prompts a careful reinterpretation of the role of APA and APA regulators on mRNA turnover and expression. A common assumption is that transcripts with longer 3'UTRs are selectively prone to greater destabilization signals by miRNA targeting and RBPs. In previous studies, the effect of APA on mRNA stability and protein output was often inferred from a limited number of mRNAs selected for their biological functions in specific contexts, such as in rapidly dividing or transformed cells where a global 3'UTR shortening is observed (Mayr and Bartel, 2009;

Sandberg *et al.*, 2008). Clearly, these results were not representative of all APA isoforms. In the current study, we employed transcriptome-wide perturbation of APA by depleting CFIm68 and described a much more complex portrait of the effects of APA shifts on mRNA stability, at least in NIH3T3 cells. Our findings are consistent with a previous transcriptome-wide study conducted under steady state (without perturbations), which found a striking correlation between the half-lives of the proximal and distal isoforms for the same gene (Spies *et al.*, 2013) This study also revealed that mRNA stability is minimally affected by 3'UTR isoform choice. Here, we demonstrate that the stability of transcripts is regulated by certain determinants that are mRNA-specific and do not often correlate with 3'UTR length. Such determinants are easily overlooked in studies that average across transcriptomes. Interestingly, stability changes among the non-APA mRNAs upon CFIm68 KO led us to identify a novel link between CFIm68, NMD and EJC. Within this novel cascade, the stability of a select group of transcripts is modulated by CFIm68 via direct APA regulation, but also indirectly via NMD/EJC pathways, whose components are under direct CFIm68 regulation (Figure 3.7, Model).

3.4.1 Unique sequence features of 3'UTRs govern transcript stability

The process of APA fine-tunes the inclusion and exclusion of *cis*-regulatory elements in the 3'UTR that can have complex interactions with one another. Among them are miRNA binding sites that can elicit potent silencing. However, silencing capacity can be modulated by a myriad of parameters that are unique to each 3'UTR, and as such the mere presence of a predicted miRNA binding site is not a proof of functional output. Equally important is the expression of miRNAs matching the predicted sites in the same cell. Expression of miRNAs varies broadly with biological contexts, which integrates cell fate, tissue type, and the presence of stress signals, for instance. Accessibility of the miRNA binding site may be limited as 3'UTR sequences may favor the

formation of complex secondary structures (Wu and Bartel, 2017). miRNA binding sites situated in closed stem-loop structures, for example, are less potent than ones positioned in open structures (Kertesz *et al.*, 2007). In the opposite scenario, miRNA binding sites located near either end of a 3'UTR tend to be more efficiently utilized than ones situated in the middle of a long 3'UTR (Grimson *et al.*, 2007). Consequently, a lengthening of the 3'UTR via APA could lead to addition of inaccessible miRNA binding sites, buried by secondary structures. Conversely, shortening of the 3'UTR could potentiate miRNA binding sites as they become closer to the 3' end, for example.

miRNA binding sites that are evolutionarily conserved appear to be more efficacious than non-conserved sites, even when found in a similar 3'UTR sequence context (Bartel, 2009; McGeary *et al.*, 2019). Here, we showed that fewer conserved miRNA target sites were found in the aUTR of dist-stable TUs compared to that of prox-stable TUs (Figure 3.5E). This is in agreement with a reduced sequence conservation of the dist-stable aUTRs (Figure 3.5H), as miRNA binding sites account for an important fraction of conserved motifs in 3'UTRs. This has previously been estimated at up to ~45% (Xie *et al.*, 2005).

Although less characterized, 3'UTRs can also encode stabilizing elements, which could also be selectively included or excluded through APA. Aside from burying miRNA target sites, 3'UTR secondary structures can stabilize mRNAs by facilitating 3' end processing and blocking the degradation by the exosome complex (Anderson and Parker, 1998; Aw *et al.*, 2016; Wu and Bartel, 2017). In yeast, poly(U) elements near the 3' end of mRNAs stabilize a subset of transcripts through the formation of secondary structures with the poly(A) tail (Geisberg *et al.*, 2014). On the other hand, the ubiquitously expressed *trans*-acting factor HuR can compete with other factors for the binding to AU-rich elements in the 3'UTRs, leading to an improved transcript stability (Mukherjee *et al.*, 2011). Lastly, increasing evidence supports the emerging role of RNA

modifications in both transcript stabilization and destabilization, depending on their interactions with RBPs and the accumulation of those marks can also be modulated through APA (Boo and Kim, 2020).

The pervasive nature of mRNA stability regulation via 3'UTR *cis*-regulatory elements is unequivocal and their importance for cellular functions and fitness is not questioned here. It is possible that gene network feedback loops converging on key mRNAs have mitigated the effects of APA on the stability of transcripts in our experimental setup. Broad strokes of global dynamics in APA are observed across processes such as development, cellular proliferation, and differentiation. It thus stands to reason that these physiological processes exert selective pressure on the distribution of regulatory elements in 3'UTRs, and on gene regulation networks. Ultimately, the biological significance of shifts in mRNA APA should be studied in specific physiological contexts and ideally by detailing specific 3'UTRs.

3.4.2 Role of CFIm68 in mRNA turnover

We discovered that global 3'UTR shortening induced by CFIm68 depletion was accompanied by a coordinated downregulation of components in the NMD and EJC pathways. Our search for known NMD targets did not reveal an enrichment in CFIm68 depletion datasets. This could have been expected, given the pleiotropic effects CFIm68 depletion should elicit, which should include spurious wide differential expression of mRNAs (Figure A2.5). Additionally, endogenous targets for NMD have been notoriously difficult to identify and genome-wide studies largely disagree on the candidates (Colombo *et al.*, 2017).

Our PAR-CLIP data-mining analysis supports the direct regulation of Smg1 (the kinase for the rate-limiting phosphorylation activation step in NMD (Yamashita *et al.*, 2001)) and Rbm8a (a core component of EJC) by both CFIm25 and CFIm68. In line with an impaired NMD function,

depletion of CFIm68 also resulted in the depletion of mRNAs within translation-associated pathways (Figure 3.6A). It is known that NMD depends on translation, and is strongly linked with inefficient or aberrant translation termination (Karousis and Mühlemann, 2019). Downregulation of translation along with NMD and EJC components could thus jointly contribute to NMD dysregulation in response to CFIm68 depletion. The extent to which CFIm68 can regulate mRNA turnover through NMD/EJC, and whether this novel link contributes to molecular and physiological phenotypes observed in CFIm68-associated diseases should be investigated.

3.5 Acknowledgements

We would like to apologize to authors whose directly related work may have not been cited in this manuscript. We thank Dr. Hamed Najafabadi and Dr. Gabrielle Perron for their input and guidance on the bioinformatics analyses. We thank all the lab members for their valuable feedback on the project.

3.6 Funding

This work was supported by the Canadian Institutes of Health Research (CIHR) grant (MOP-123352) to T.F.D.; Fonds de recherche du Québec Santé (FRQS) Doctoral training award, and CIHR Doctoral Research Award to H-W. T.

3.7 Materials and methods

3.7.1 3'UTR-seq sample preparation and sequencing

NIH3T3 cells (ATCC) were cultured in Dulbecco's modified Eagle medium (DMEM) with 10% calf serum (Cytiva) in a humidified incubator at 37 °C with 5% CO_2. CFIm68-KO NIH3T3 cells were generated previously (Tseng *et al.*, 2022). For this study, cells were seeded in 6-well plates and treated with Actinomycin-D (5 μg/mL) at 70-80% confluency. Treated cells were then

harvested at 0-, 30-, 150-, and 360-minute time points in QIAZol (Qiagen) for total RNA extraction using the Monarch total RNA miniprep kit (NEB). Using 500 ng of total RNA, 3'UTR libraries were generated using QuantSeq 3' mRNA library prep kit REV (Lexogen) following the manufacturer's instructions. Quality control and quantification of the prepared sequencing libraries were performed on a Bioanalyzer (Agilent) with the High Sensitivity DNA Kit (Agilent). Samples were then sequenced on the Illumina NextSeq 500 platform for 75 cycles with single end reads by the IRIC genomics platform (Université de Montréal).

3.7.2 3'UTR-seq data processing

3.7.2.1 Pre-processing, read alignment, and counting features

Raw reads were processed first by trimming adapters (Trimmomatic v0.36) and examined for quality control with FastQC v0.12.1. Sequences were then mapped to the mm10 genome with STAR v2.7.2b. On average we obtained around 10 million mapped reads per sample. Quantification of APA usage was performed as previously described (Tseng *et al.*, 2022). In brief, peaks were first identified using mapped reads from all samples using CLIPanalyze (https://bitbucket.org/leslielab/clipanalyze). Internal priming events were then removed along with low-usage events to compile a curated list of valid APA sites. This list was then referenced to count APA events in each sample using featureCounts (Liao *et al.*, 2013).

3.7.2.2 Differential APA

Only samples at the 0-minute time point were used for differential APA calculation. Size factors calculated by DESeq2 (Love *et al.*, 2014) were first extracted using all counts for each sample. Subsequently, for each sample, read counts were split into proximal and distal isoform reads by taking the top two expressing isoforms per gene. DESeq2 was supplied with previously extracted size factors for each originating sample, and the design formula ~ *location* + *condition* + *location :*

condition, where *location* consists of proximal and distal, and *condition* consists of WT and CFIm68-KO. The effect size from the interaction term was then extracted as the RED (Relative Expression Difference) score, which is equivalent to the log ratio of the proximal to the distal site in CFIm68-KO over WT cells.

3.7.2.3 Differential transcript stability

To estimate the stability coefficient for every transcript, which is equivalent to the log of transcript counts over time, we supplied DESeq2 with the design formula ~ *condition + time : condition*, where *condition* consists of WT and CFIm68-KO, and *time* is the series of time points in minutes as a continuous variable. The interaction term for each condition can then be extracted with the effect size being the stability coefficient used in this study. Differential TU stability was estimated with DESeq2 supplied with the same design formula, using TU counts (gene counts) outputted by STAR, but the difference between the interaction terms for each condition was extracted instead.

3.7.2.4 Differential gene expression

Only samples from the 0-minute time point were used for the calculation of differential gene expression. DESeq2 was used with the design formula ~ *condition*, where *condition* consists of WT and CFIm68-KO, and the corresponding result was extracted. Differential gene expression of CFIm25 and CFIm68 KD NIH3T3 (Tseng *et al.*, 2022), and CFIm68 KO HEK293T (Zhu *et al.*, 2018) was estimated similarly using published sequencing data. Gene Set Enrichment Analysis was performed on differentially expressed genes using the R package fgsea (Korotkevich *et al.*, 2021) against annotated Canonical Pathways taken from MSigDB (Liberzon *et al.*, 2011; Mootha *et al.*, 2003; Subramanian *et al.*, 2005).

3.7.3 Characterization of aUTR

The proximal and distal isoforms were first identified as the top two expressing isoforms for each TU. The location and sequence of the aUTR can then be deduced for each TU. Annotation of predicted conserved miRNA binding sites for conserved miRNA families was downloaded from TargetScan 8.0 (Friedman *et al.*, 2009). RNAfold (Hofacker, 2003; Lorenz *et al.*, 2011) was used for *in silico* modelling of RNA secondary structures and estimation of the minimum free energy for folding. The phyloP (phylogenetic P-value) scores quantifying the per-base conservation across 60 vertebrates were taken from the UCSC Genome Browser (Lee *et al.*, 2022), originally calculated with the PHAST package (Pollard *et al.*, 2010).

3.7.4 Statistical analysis and data visualization

All statistical significance cut-off was set at $P < 0.05$. Data visualization was performed in R using ggplot2 (Wickham, 2016) and ComplexHeatmap (Gu *et al.*, 2016).

Chapter 4:

Tracking APA through metastasis at single-cell resolution

Hsin-Wei Tseng[1,2], Phuong U. Le[4], Rima Ezzeddine[1,2], Charlotte Girondel[1,3], Marco Biondini[1,3], Matthew Dankner[1,3], Kevin Petrecca[4], Peter M. Siegel[1,3], Thomas F. Duchaine[1,2,*]

[1] Rosalind and Morris Goodman Cancer Institute, McGill University, Montréal, QC, Canada.

[2] Department of Biochemistry, McGill University, Montréal, QC, Canada.

[3] Department of Medicine, McGill University, Montréal, QC, Canada

[4] Department of Neurosciences, Montreal Neurological Institute-Hospital, McGill University, Montréal, QC, Canada

* Correspondence: thomas.duchaine@mcgill.ca

4.1 Abstract

Global 3'UTR shortening in cancer through alternative polyadenylation (APA) is a well-documented phenomenon that is frequently associated with the post-transcriptional activation of oncogenes. While aberrant expression of key APA regulators is thought to contribute to cancer progression, how APA adapts and shapes the process of metastasis is not well understood. Moreover, the extent of intra-tumoral 3'UTR heterogeneity and how it may contribute to distinct cell identities within the mosaicism of tumor tissues also remain elusive. Here we conduct the first longitudinal study of tumor APA profiling at single-cell resolution and demonstrate distinctive APA profiles between matching, paired primary and metastatic tumors. APA profiles resolved cell origins better than gene expression profiles. Detailed profiling of 3'UTR shortening from the primary to metastatic tumors revealed a correlation with a decrease in proliferation signatures. Lastly, we identified a quiescent stem-like cell population with a unique APA expression profile. Overall, our results shed light into the process of APA adaptation and its contribution to shaping metastasis.

4.2 Introduction

The 3' untranslated region (3'UTR) of a messenger RNA (mRNA) can be dynamically regulated through alternative polyadenylation (APA) at different polyadenylation signals (PAS) to generate mRNA isoforms differing only in the 3'UTRs (Tian and Manley, 2017). The majority of mammalian genes are under APA control, which dictates the inclusion and exclusion of *cis*-regulatory elements in 3'UTRs, such as target sites for microRNAs (miRNAs) and RNA binding proteins (RBPs). These regulatory sites, in turn, confer post-transcriptional control over mRNA stability, subcellular localization, and translational output (Tian and Manley, 2017). APA regulators, composed of components of the mRNA cleavage and polyadenylation machinery and their interactors, are often expressed in a tissue-specific manner, which in turn leads to tissue-specific APA patterns (Lianoglou *et al.*, 2013; Zhang *et al.*, 2005). Interestingly, alterations in APA can often occur independently of changes in gene expression, such that APA patterns can improve cell fate and state signatures derived from gene expression analyses (Lianoglou *et al.*, 2013). In line with this, recent advances in single-cell APA profiling have enabled the discovery of cell type-specific APA patterns in known neuronal subtypes, as well as the identification of subpopulations within those subtypes that were previously indistinguishable based on gene expression patterns (Yang *et al.*, 2021). Additionally, distinct APA patterns have uncovered novel cell subpopulations during human embryonic development (Gao *et al.*, 2021). While changes in APA can be specific to certain mRNAs, coordinated and transcriptome-wide changes in APA occur during major biological transitions. For instance, overall 3'UTR shortening is correlated with increasing proliferation, such as in T-cell activation and cell transformation (Mayr and Bartel, 2009; Sandberg *et al.*, 2008). Conversely, 3'UTRs are globally lengthened during embryonic development and differentiation processes (Ji *et al.*, 2009).

While the biological consequences of these orchestrated APA reorganizations remain largely unknown, APA dysregulation has been associated with various diseases. Notably, global 3'UTR shortening is seen across diverse cancer types when compared to corresponding normal tissues (Masamha *et al.*, 2014; Mayr and Bartel, 2009; Xia *et al.*, 2014). These widespread shortening of the 3'UTRs feature cancer-type-specific distinctions, although the 3'UTRs of certain mRNAs are also recurrently shortened across cancers (Xia *et al.*, 2014). APA changes enriched in cancer can affect protein and energy metabolism mRNAs, supporting that APA dysregulation indeed plays a functional role in cancer (Burri and Zavolan, 2021). Moreover, a subset of 3'UTR alterations in breast cancer identifies patient subgroups with greater chances of relapse and metastasis, providing prognostic power beyond predictions based on traditional clinicopathologic factors (Wang *et al.*, 2016; Wang *et al.*, 2018; Xia *et al.*, 2014).

Aberrant expression of some APA regulators has been directly implicated in cancer progression. For instance, CFIm25, a member of the mammalian cleavage factor I (CFIm) complex, is downregulated in different types of cancers (Masamha and Wagner, 2017; Masamha *et al.*, 2014; Sun *et al.*, 2017). In glioblastoma, global 3'UTR shortening potentiates the *CCND1* and *Pak1* oncogenes, which in turn contribute to enhancing cell proliferation and tumor growth (Chu *et al.*, 2019; Masamha *et al.*, 2014). Similarly, PCF11, a subunit of the CFIIm complex, is an important APA regulator of the WNT signalling cascade and controls the cell cycle, proliferation, apoptosis, and neurodifferentiation. Knockdown of PCF11 promotes neurodifferentiation and low PCF11 expression is associated with favorable neuroblastoma prognosis as well as spontaneous regression (Ogorodnikov *et al.*, 2018).

Efforts to elucidate the relationship between APA, APA regulators, and metastasis are ongoing, given that metastasis is the leading cause of death across all cancer types (Steeg, 2006).

Thus far, genome-wide studies have largely focused on APA alterations in paired comparisons of tumors and their adjacent normal tissues, and few have examined changes in APA along the metastatic process. Furthermore, these studies are based overwhelmingly on bulk RNA-seq experiments. For these reasons, the extent of intra-tumoral APA heterogeneity and how it may contribute to adaptations favoring metastasis remain unexplored.

Here, we conduct the first longitudinal study of APA using paired primary and metastatic tumors at single-cell resolution. Leveraging a uniquely organotrophic breast cancer patient-derived xenograft (PDX) model, we demonstrate that distinct tumor APA profiles are better at resolving cell origins than gene expression profiles. We identify a trend towards 3'UTR shortening as tumors progress from primary to metastasis, and this surprisingly correlates with a decrease in proliferative signatures. Finally, we identify a quiescent stem-like cell population with a novel APA profile. Together, our results unveil the intra-tumoral APA diversity and support the importance of APA adaptation in the progression of cancer to metastasis.

4.3 Results

4.3.1 APA better resolves tumor origin than gene expression

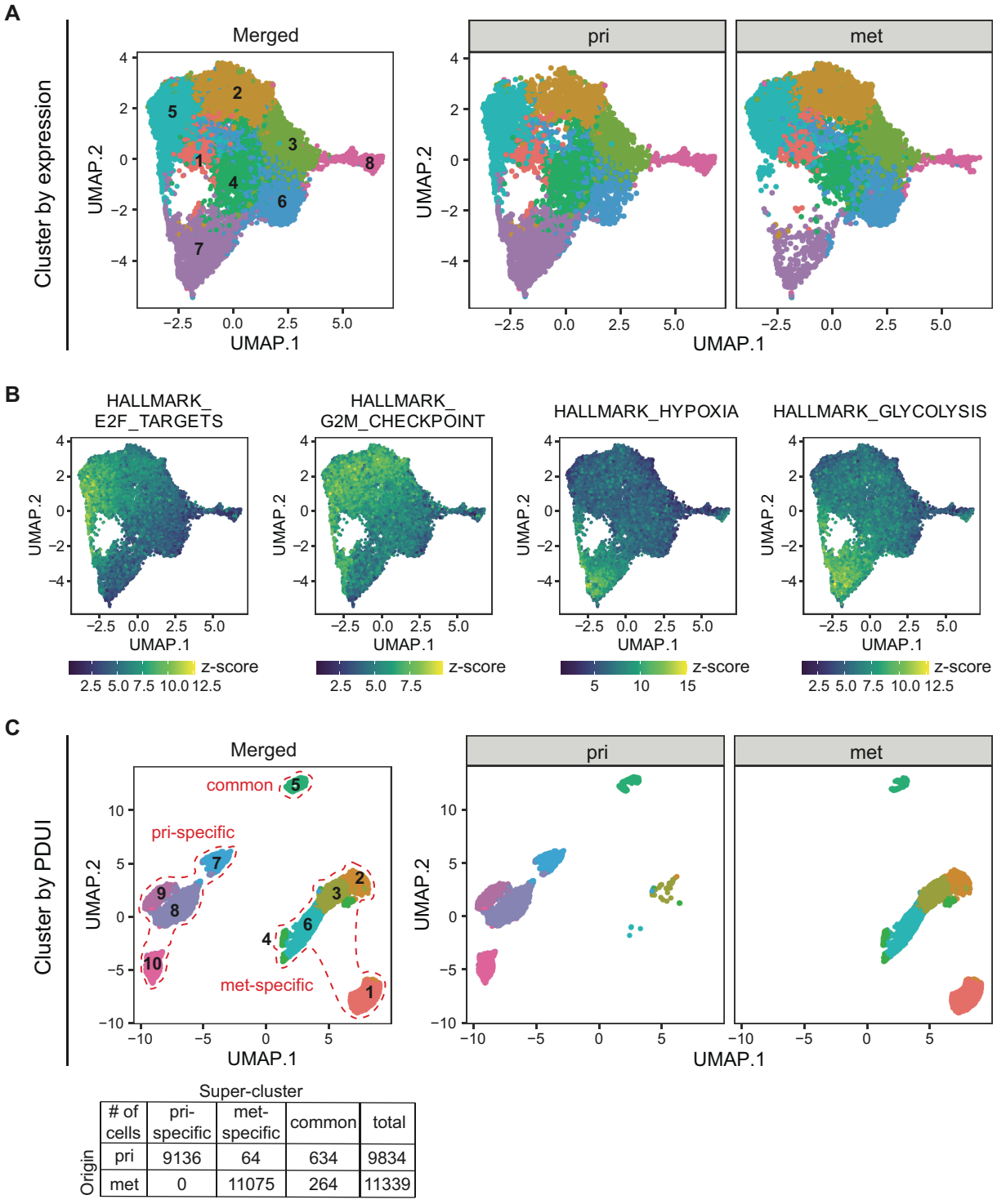
To study the diversity of changes in APA through metastasis, we performed single-cell RNA-seq of primary and metastatic tumors in a well-established PDX model of ER+ breast cancer, GCRC1971 (Savage *et al.*, 2020). This PDX model is driven in part by the amplification of FGFR1, and is uniquely organotrophic towards skull-base metastasis, ensuring reproducibility and reducing background variability. Between 91 to 96% of our sequencing reads mapped unambiguously to the reference human genome, and over 98% of identified cells were annotated as epithelial cells, confirming the purity and quality of our libraries. After filtering low-quality cells, we retained 9834 cells from the primary tumors, and 11339 cells from the metastatic tumors for downstream analyses (Figure A4.1).

To assess the intra-tumoral diversity of cancer cells, we first performed a clustering analysis with gene expression by merging datasets to include all primary and metastatic tumor cells. Cells clustered into eight populations (Figure 4.1A, Merged). When we categorized the cells based on their corresponding origins, primary (pri) or metastasis (met), we observed a slight redistribution of cell populations between different clusters. Clusters 4, 7, and 8 were relatively depleted, whereas the other clusters were enriched from primary to metastatic tumors (Figure 4.1A, pri and met). Notwithstanding those differences, all identified clusters in the merged dataset were present in detectable proportions in both primary and metastatic tumors. Following the identification of gene expression clusters, we next performed pathway analysis to investigate the relevant functional annotation of each cluster. Clusters 1, 2, and 5 scored highly in proliferation-related signatures, and featured gene sets such as E2F targets, and G2M checkpoint (Figure 4.1B). On the other hand, cluster 7 featured highly upregulated hypoxia and glycolysis signatures compared to the other

clusters (Figure 4.1B). All clusters associated with more than one cancer hallmark pathways as annotated in MSigDB (Liberzon *et al.*, 2011; Subramanian *et al.*, 2005). Intriguingly, cluster 8 featured a depletion in almost all significant gene expression signatures (Figure A3.2).

To investigate the 3'UTR diversity among tumor cells, the relative expression of 3'UTR isoforms was quantified using the scDaPars pipeline (Gao *et al.*, 2021) for each expressed gene in each cell as the Percentage of Distal Poly(A) usage Index (PDUI). The general interpretation of PDUI is that a larger PDUI corresponds to a longer averaged 3'UTR length of all mRNAs expressed from one gene. Cells were clustered into 10 distinct populations using PDUI calculated for the 799 genes that passed the expression threshold across all cells (Figure 4.1C, Merged). Strikingly, three super-clusters of cells emerged when we assigned each cell according to their origin, pri or met (Figure 4.1C, pri, met, and table). Clusters 7, 8, 9, and 10 arose exclusively from cells of primary tumors origin, which we termed the "pri-specific" super-cluster. Conversely, clusters 1, 2, 3, 4, and 6 were almost exclusively derived from metastatic tumors, and altogether were termed the "met-specific" super-cluster. Lastly, cluster 5 was comprised of a mixture of cells from both primary and metastatic tumors and was termed the "common" cluster (Figure 4.1C).

Together, these results demonstrate a high intra-tumoral heterogeneity in both gene expression and APA patterns. Strikingly, APA changes better distinguished cell origins than gene expression changes.



Origin	# of cells	pri-specific	met-specific	common	total
pri	9136	64	634	9834	
met	0	11075	264	11339	

Figure 4.1: APA better resolves tumor origin than gene expression. (A) UMAP plots for cells clustered by gene expression. Merged, combined datasets from all tumor cells; pri, cells originated from the primary tumors; met, cells originated from the metastatic tumors. The cluster

numbers assigned from 1 to 8 are indicated and color-coded. (B) UMAP plots for all cells clustered by gene expression as in (A), and colored according to the level of expression for the genes in each of the indicated cancer hallmark gene sets. (C) UMAP plots for cells clustered by PDUI in the merged dataset, primary tumors, or metastatic tumors. Cluster numbers are indicated from 1 to 10 and color-coded. Each cluster is also assigned to one of the three super-clusters, pri-specific, met-specific, and common. The number of cells originated from the primary or metastatic tumors assigned to each super-cluster is tabulated in the table below.

4.3.2 3'UTRs are shortened from primary to metastatic tumors

To map changes in APA that occur through metastasis, we performed a three-way differential APA analysis on all mRNAs passing the expression threshold among the three identified super-clusters: pri-specific, met-specific, and common. Strikingly, all 240 genes with significant PDUI changes between pri-specific and met-specific super-clusters presented a reduced PDUI in the met-specific super-cluster. Similarly, all 171 genes with significant PDUI changes between the met-specific and common super-clusters featured a reduction in PDUI in the common super-cluster. Further in line with this, all 415 genes with significantly different PDUI had a lower PDUI in the common super-cluster compared to the pri-specific super-cluster. This trend of PDUI reduction, namely the reducing 3'UTR length from the pri-specific super-cluster towards the met-specific super-cluster, and finally to the common super-cluster was clearly visible despite some heterogeneity within each super-cluster (Figure 4.2A). Furthermore, this reduction of 3'UTR length through metastasis was preserved even when transcripts of all 799 detectable genes were included to calculate the average PDUI of each cell (Figure 4.2B).

To infer the possible functional significance of 3'UTR shortening through metastasis, we performed pathway analysis on genes expressing mRNAs with significant APA differences between pri-specific and met-specific super-clusters. We performed this analysis using annotations

from three different databases: Gene Ontology (GO) (Ashburner *et al.*, 2000; Consortium *et al.*, 2023), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), and Reactome (Gillespie *et al.*, 2021). Interestingly, the top enriched terms (with p-values on the order of 10^{-9}) were all associated with energy metabolic pathways: electron transfer activity from GO, oxidative phosphorylation from KEGG, and the TCA cycle from Reactome (Figure A3.3).

Collectively, these analyses demonstrated a significant shortening of the 3'UTR through metastasis, particularly in mRNAs encoding the energy metabolic pathways.

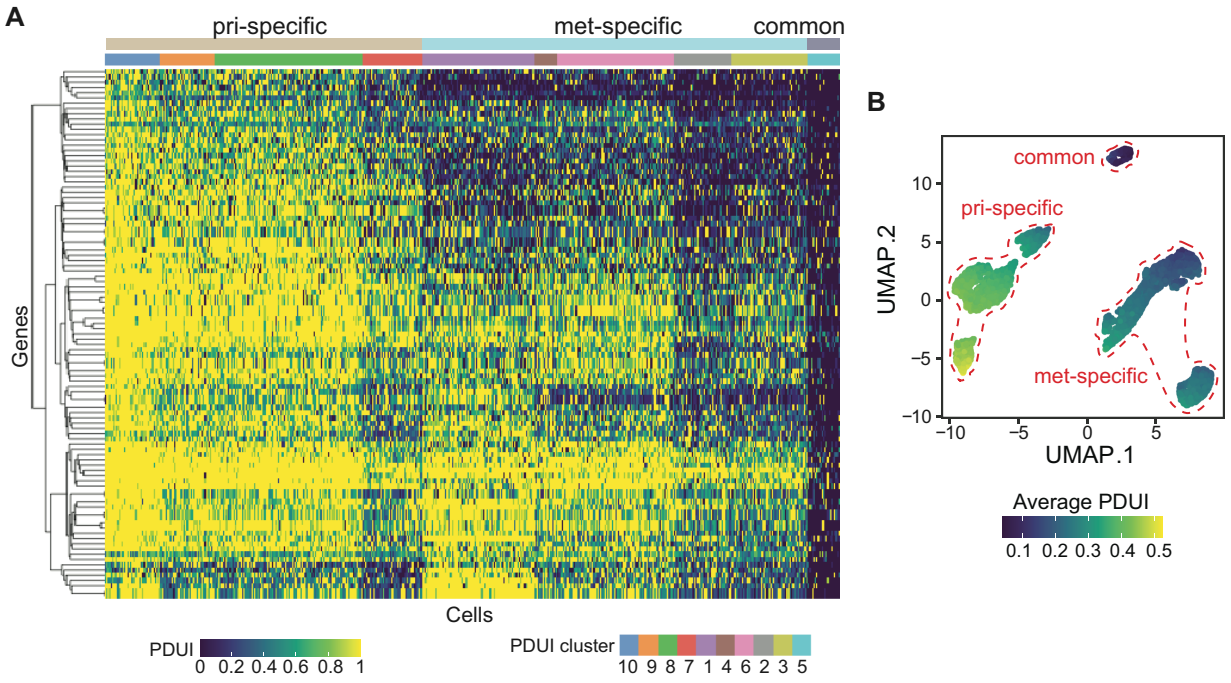


Figure 4.2: 3'UTRs are shortened from primary to metastatic tumors. (A) Heatmap of PDUI for all cells arranged according to its super-cluster assignment (pri-specific, met-specific, or common), and PDUI cluster assignment (1-10). (B) UMAP of all cells clustered by PDUI, and colored by the average PDUI of all 799 genes for each cell. Super-clusters are indicated.

4.3.3 3'UTR length positively correlates with proliferation

Global 3'UTR reprogramming is generally associated with transitions between cellular states and identities. For instance, shortening of the 3'UTRs was seen in highly proliferative cell types such as activated T cells and transformed cells (Mayr and Bartel, 2009; Sandberg *et al.*, 2008). Lengthening of 3'UTRs is also observed during embryonic cell development and differentiation (Ji *et al.*, 2009; Shepard *et al.*, 2011). We thus next investigated whether the shortening of 3'UTRs through metastasis in our single-cell dataset was correlated with changes in proliferative capability. To quantify proliferation, we utilized two previously compiled gene sets and independently calculated proliferation scores for each cell (see methods). The following two gene expression signatures were used: 1) the proliferation-correlated signature, which is based on time-course CFSE staining and cell sorting (Jiang *et al.*, 2021), and 2) the meta-PCNA signature, which is a collection of the top genes correlated with the expression of PCNA (Proliferating Cell Nuclear Antigen) across diverse tissue samples (Venet *et al.*, 2011). The scores obtained from these two distinct gene sets were highly and positively correlated (Figure A3.4). When plotted on cells clustered by PDUI, the proliferation signature scores formed clear gradients, suggesting a correlation with the PDUI scores (Figure 4.3A, B). Surprisingly, a strong positive correlation was observed for both the proliferation-correlated score (Pearson's $R = 0.55$, $p < 2.2e\text{-}16$) and meta-PCNA score (Pearson's $R = 0.27$, $p < 2.2e\text{-}16$) with the average PDUI scores across all cells (Figure 4.3C, D). In other words, cells with longer average 3'UTR lengths were associated with more proliferative gene signatures in our datasets, in stark contrast to the expected relationship between proliferation and 3'UTR length, based on previous work (Mayr and Bartel, 2009; Sandberg *et al.*, 2008).

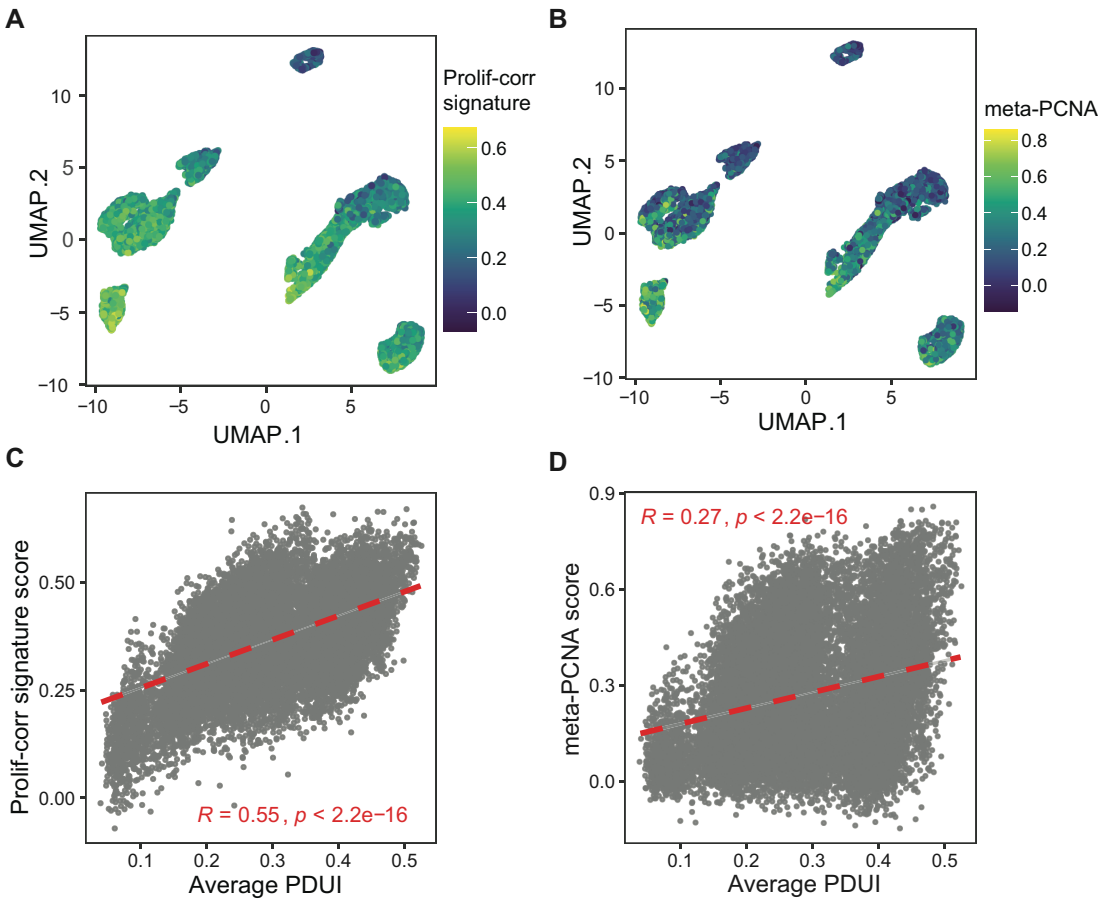


Figure 4.3: 3'UTR length positively correlates with proliferation. (A, B) UMAP plots of all cells clustered by PDUI and colored according to the proliferation-correlated signature score (A), or the meta-PCNA signature score (B). (C, D) Scatter plots of the proliferation-correlated signature score (C) or the meta-PCNA score (D) against the average PDUI for each cell. Pearson's *R* and *p*-values are indicated.

4.3.4 A common super-cluster characterizes a putative quiescent cancer stem cell population

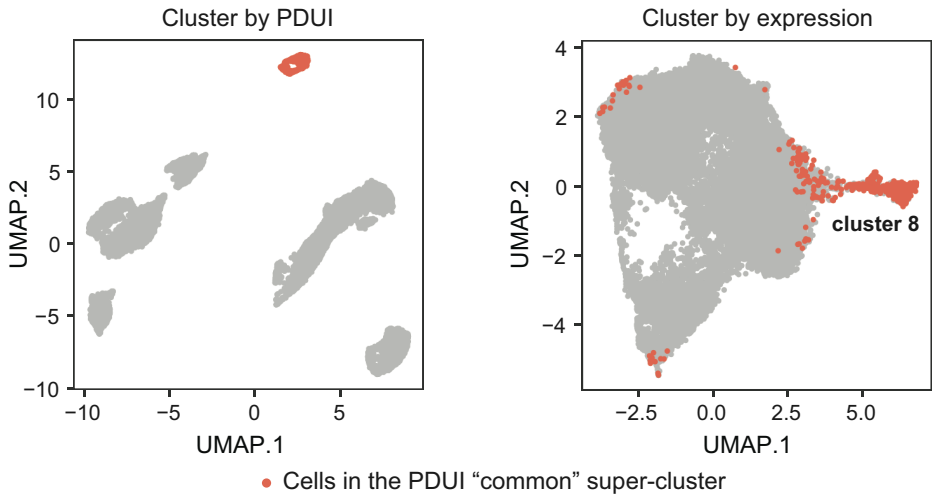
Throughout combined analyses of APA, gene expression, and proliferation signatures, the "common" PDUI super-cluster emerged as having the shortest average 3'UTR while being the least proliferative. Strikingly, the "common" super-cluster of cells also corresponded nearly exclusively to cluster 8 when cells were clustered by gene expression (Figure 4.4A). None of the

other PDUI clusters or super-clusters exhibited such a distinct APA and gene expression profile (Figure A3.5).

To better infer the functional identity of cluster 8 cells, we identified the driving expression changes for this cluster and performed gene set enrichment analysis (GSEA). All except one significant (padj < 0.05) pathway were depleted from this cluster of cells, and pathways associated with cell growth and proliferation (MYC targets and MTORC1 signalling), metabolism (fatty acid metabolism and oxidative phosphorylation), and reactive oxygen species, were among the most depleted (Figure 4.4B). These observations support the possibility that cluster 8 identifies a quiescent cell state in primary and metastatic tumor cells. Furthermore, the top upregulated markers of cluster 8 included some well-known long non-coding RNAs (lncRNAs): MALAT1, XIST1, and NEAT1, which are frequently associated with cancer stem cells, metastasis, and malignancy across a variety of cancer types (Chen *et al.*, 2020; Hussen *et al.*, 2022; Ma *et al.*, 2023). Other cancer aggressivity related genes including FGFR1, VEGFA, and WSB1 also ranked among the top of cluster 8 upregulated markers (Figure 4.4C). Interestingly, the short average 3'UTR length in cluster 8 was accompanied by an overall upregulation of the cleavage and polyadenylation core machinery (Figure A3.6). We note that this was observed during the reprogramming of iPS cells (Ji and Tian, 2009). Further in line with this, some markers previously associated with cancer cell stemness, including CD44, some of the SOX family genes (SOX2 and SOX9), ALDH family genes (ALDH3A2 and ALDH2), and KLF4 (Pouremamali *et al.*, 2022; Zanoni *et al.*, 2022; Zhao *et al.*, 2017; Zhou *et al.*, 2019), were also marginally upregulated in cluster 8 compared with other clusters, although these changes characterized a smaller subset of cells (Figure A3.7).

In summary, we identified a distinct population of cells expressing markedly short 3'UTRs, which is characterized by a gene expression profile consistent with quiescent cancer stem cells.

A



B

Pathway	Gene ranks	NES	pval	padj
HALLMARK_MITOTIC_SPINDLE		1.32	$5.4 \cdot 10^{-3}$	$1.4 \cdot 10^{-2}$
HALLMARK_HEDGEHOG_SIGNALING		1.47	$2.9 \cdot 10^{-2}$	$6.2 \cdot 10^{-2}$
HALLMARK_HYPOXIA		1.21	$3.5 \cdot 10^{-2}$	$7.1 \cdot 10^{-2}$
HALLMARK_UV_RESPONSE_DN		1.21	$3.9 \cdot 10^{-2}$	$7.6 \cdot 10^{-2}$
HALLMARK_ANGIOGENESIS		1.29	$1.4 \cdot 10^{-1}$	$2.3 \cdot 10^{-1}$
HALLMARK_KRAS_SIGNALING_DN		0.82	$9.7 \cdot 10^{-1}$	$9.7 \cdot 10^{-1}$
HALLMARK_MYC_TARGETS_V2		−1.64	$2.4 \cdot 10^{-4}$	$1.2 \cdot 10^{-3}$
HALLMARK_ALLOGRAFT_REJECTION		−1.56	$5.5 \cdot 10^{-5}$	$3.1 \cdot 10^{-4}$
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY		−1.82	$7.5 \cdot 10^{-7}$	$4.7 \cdot 10^{-6}$
HALLMARK_MTORC1_SIGNALING		−1.60	$7.4 \cdot 10^{-7}$	$4.7 \cdot 10^{-6}$
HALLMARK_FATTY_ACID_METABOLISM		−1.67	$5.4 \cdot 10^{-7}$	$4.5 \cdot 10^{-6}$
HALLMARK_XENOBIOTIC_METABOLISM		−1.67	$4.9 \cdot 10^{-8}$	$4.9 \cdot 10^{-7}$
HALLMARK_DNA_REPAIR		−1.77	$6.3 \cdot 10^{-10}$	$7.9 \cdot 10^{-9}$
HALLMARK_ADIPOGENESIS		−1.72	$3.3 \cdot 10^{-10}$	$5.5 \cdot 10^{-9}$
HALLMARK_MYC_TARGETS_V1		−2.03	$3.3 \cdot 10^{-27}$	$8.2 \cdot 10^{-26}$
HALLMARK_OXIDATIVE_PHOSPHORYLATION		−2.06	$2.1 \cdot 10^{-28}$	$1.1 \cdot 10^{-26}$

C

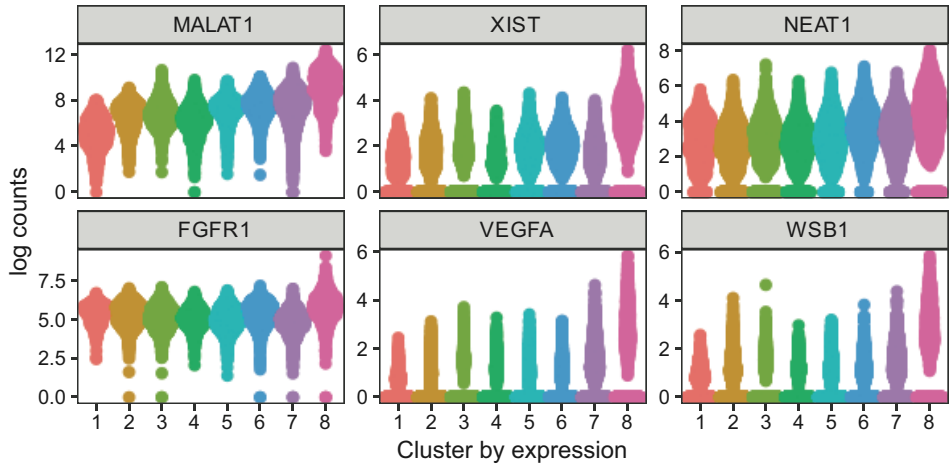


Figure 4.4: A common super-cluster characterizes a putative quiescent cancer stem cell population. (A) UMAP plots of all cells clustered by PDUI (left) or gene expression (right). Red points indicate cells originated from the PDUI “common” super-cluster. The location of cluster 8 from gene expression clustering is indicated on the right UMAP plot. (B) Gene set enrichment analysis for the marker mRNAs from cluster 8. Both the p-values prior to multiple-testing adjustment (pval) and after (padj) are indicated. NES, normalized enrichment score. (C) Top marker mRNAs that are upregulated in cluster 8 cells.

4.4 Discussion

General shortening of the 3'UTR in tumor, as well as the specificity of APA patterns in specific subtypes of cancers have now been well documented. Albeit it being cancer-type dependent, 3'UTR shortening in well-known oncogenes, notably those related to proliferation, correlates with oncogenic activation and metastasis. In some cases, these changes significantly added prognostic power beyond common clinical and molecular covariates (Singh *et al.*, 2009; Xia *et al.*, 2014; Yuan *et al.*, 2021). Nonetheless, whether and how transcriptome-wide APA reorganizes through the process of metastasis, the primary cause of death for over 90% of cancer patients (Steeg, 2006), has not been directly assessed. Furthermore, the extent of APA heterogeneity and how it may contribute to distinct cellular identities and functions within a mosaic tumor population also remain poorly understood.

Here we performed a single-cell longitudinal study of APA using paired tumors to capture cellular heterogeneity and changes through metastasis. We demonstrated the superior resolution of APA over gene expression profiles in identifying cellular identities and observed a wide-scale shortening of the 3'UTRs from primary to metastatic tumors. Surprisingly, this shortening correlated with a reduction in proliferative signatures. Lastly, we identified a novel population of cells exhibiting atypically short 3'UTRs matched with quiescence and stem-cell gene signatures.

4.4.1 Single-cell APA analyses provide unique functional insights

Metastasis is a complex, multi-step process through which tumor cells undergo extensive selection and adaptation to colonize distal sites. In a model of clonal seeding from the primary tumor, the relative abundance of subclones can vary between paired tumors (Merino *et al.*, 2019). This behaviour is consistent with the heterogeneous gene expression profiles that nonetheless conserved common features while transitioning from primary to metastatic tumors. However,

changes in our APA data were much more pronounced than gene expression changes alone, and clearly distinguished between primary and metastatic tumors. The apparent inconsistency in the magnitude and direction of changes between gene expression profiles and APA is coherent with the previously identified lack of correlation between tissue-specific mRNA expression changes and APA (Lianoglou *et al.*, 2013), and argues in favor of the functional importance of APA regulation in metastasis.

Perhaps above all other functional implications, our results and other's support the importance of APA regulation in metabolic adaptation. For instance, genes with significant APA changes between the primary and metastatic tumors were enriched in the oxidative phosphorylation pathway. These changes may reflect a metabolic reprogramming that is crucial to meet anabolic energetic demands in progressing cancers (Ward and Thompson, 2012). This result may also be specific for our model with its unique organotropism, as distinctive metabolic signatures can arise depending on the site of metastasis (Dupuy *et al.*, 2015). Indeed, APA regulation of metabolic adaptation is an emerging theme in both normal and pathological cellular processes. For instance, APA dysregulation in distinct subsets of metabolic mRNAs was recently implicated in the impairment of stem cell activation (Sommerkamp *et al.*, 2020), in cancers (Tan *et al.*, 2021), in endocrine diseases (Zhao *et al.*, 2005), and in non-alcoholic fatty liver disease (Jobbins *et al.*, 2022). In addition, master APA regulators are known to exert broad control over metabolic networks, such as for mTOR and its associated pathways (Tseng *et al.*, 2022; Zhang and Tian, 2023). Conceivably, changes in the energy and other metabolic demands through metastasis could be the main driver for the APA reconfiguration we observed in our system of study.

Another salient observation stemming from our analysis is the correlation between 3'UTR shortening and the reduction in proliferative signatures from primary to metastatic tumors. This is

unlike the global APA reprogramming often described for cell proliferation-associated processes, such as 3'UTR shortening in T cell activation (Sandberg *et al.*, 2008) for example, or during oncogenic transformation (Mayr and Bartel, 2009). While these studies pointed to increased proliferation as a strong driver for global 3'UTR shortening across select cell types and processes, our results indicate that this association is not universal. Other exceptions have been reported. For instance, in the differentiation process of certain secretory cell types, such as syncytiotrophoblast and plasma cells, 3'UTRs are globally shortened and either coupled with a reduced, or a complete loss of association with proliferation signatures, depending on the cell type (Cheng *et al.*, 2020). Interestingly, 3'UTR shortening in these cells correlates strongly with gene signatures portraying the transport of secreted proteins, supporting the possibility that this process is a functionally relevant driver of APA reprogramming in these cells. In another example, cells in the tumor environment of lung adenocarcinoma exhibit APA changes in gene sets enriched in various pathways in a cell type-dependent manner, even though globally all cell types trend towards 3'UTR shortening (Burri and Zavolan, 2021). Together, these observations substantiate the concept that APA reprogramming can be driven by a variety of cell type-specific processes, including but not limited to proliferation. An important remaining question in line with our study is the extent to which APA changes can contribute to metabolic remodelling or to unique proliferation processes in metastasis. It will also be interesting to directly inquire whether and how APA changes in specific metabolic pathways synergize with one another in the metastatic process, and thus may contribute to gene expression coordination during cell adaptation and systems reprogramming.

4.4.2 A unique APA profile in a putative quiescent cancer stem cell population

Using single-cell APA analyses, we identified a unique population of cells present in both primary and metastatic tumors that may represent quiescent cancer stem cells (CSC). Beyond their

presence in both primary and metastatic tumors, three other findings support this hypothesis: 1- previously identified CSC markers were upregulated in at least a sub-population of these cells, 2- top markers of these cells are commonly associated with increased cancer malignancy, and 3- aside from presenting a distinctive APA pattern, these cells also exhibit a unique gene expression profile (cluster 8) that is consistent with a dormant or quiescent cell state.

CSCs have been identified in multiple cancer types including in solid breast cancer tumors, which were first identified as a lineage of $CD44^{+}CD24^{-/low}$ cells of low abundance, expressed at as few as 100 cells per tumor, and is capable of initiating new tumors (Al-Hajj *et al.*, 2003). Additional CSC surface markers have since been identified in breast cancer across various subtypes, but the extent of detection appears to vary across models (Zhang *et al.*, 2020; Zheng *et al.*, 2021). For this reason, while the identification of the fraction of cells in cluster 8 expressing previously identified CSC markers supports their hypothesized CSC identity, ultimately a robust demonstration will require their isolation and similar functional tumor initiation assays.

Other than cell surface markers that distinguish CSCs from other tumor cells, numerous signaling programs and their regulators were identified as contributors of CSC maintenance and functions. Among them is a group of long non-coding RNAs (lncRNAs), including MALAT1, XIST, and NEAT1, which were also among the top upregulated genes of cluster 8 cells in our data. These lncRNAs have been broadly implicated in maintaining and promoting CSC stemness, immune evasion, and chemoresistance, and are in turn correlated with cancer recurrence as well as poor prognosis for some patients (Schwerdtfeger *et al.*, 2021; Shen *et al.*, 2021). The underlying mechanisms are not fully understood, but some of their functions have been attributed to their ability to act as scaffolds to bring together proteins and RNP complexes, as well as sponging away tumor-suppressive miRNAs, as in the competitive endogenous RNA (ceRNA) hypothesis (Pa *et*

al., 2017; Schwerdtfeger *et al.*, 2021; Wu *et al.*, 2016). Of note, global 3'UTR shortening caused by the depletion of CFIm25 has been implicated in altering the putative ceRNA network and in the repression of tumor suppressors *in trans* (Park *et al.*, 2018). This raises the question of whether the markedly upregulated lncRNAs in cluster 8 can functionally synergize with the overall 3'UTR shortening reprogramming in driving CSC phenotypes.

Prior to this study, APA profile in CSC had not been characterized directly. As such, their cellular identity can only be inferred from accepted traits of CSC, or from signatures detected in normal stem cell counterparts. For example, the ability to remain in a prolonged quiescence state is considered a fundamental attribute of adult stem cells (Cheung and Rando, 2013). Activation of quiescent stem cells and the subsequent re-entry into cell cycle is typically associated with global 3'UTR shortening (Sommerkamp *et al.*, 2021). Similar to adult stem cells, CSCs are typically characterized as quiescent, or slow-cycling, which may be a crucial mechanism for their resistance to anti-proliferative chemotherapy and the subsequent relapse (Clevers, 2011). Interestingly, our data is consistent with the expression of massively shortened 3'UTRs in the putative CSC population, and their overall reduced metabolism and proliferation is consistent with quiescence. While our current understanding of CSC may not provide a definitive explanation or the full meaning of our observation, clues can be gained when considering key differences between CSC and normal adult stem cells. One notable difference is their degree of dependence on the stem cell niche, which is a specialized microenvironment stem cells reside. The maintenance of CSC self-renewal and its other properties is highly dependent on a host of chemokines and cytokines produced by the stroma and by other cells recruited to the niche in a positive feedback loop, mimicking a chronic inflammatory profile (Li and Neaves, 2006). In such a microenvironment, CSCs are also endowed with metabolic plasticity unparalleled by normal stem cells (Peiris-Pagès

et al., 2016). Global APA reprogramming has been previously linked with processes like inflammation as well as response to metabolic stresses and adaptation outside of the CSC context (Burri and Zavolan, 2021; Sommerkamp *et al.*, 2020). It stands to reason that these changes may well be important in shaping CSC attributes.

Understanding the critical differences in the biological processes between CSC and normal adult stem cells has deep clinical implications and may provide promising foundations for targeted cancer therapies. Our single-cell APA study provides the first glimpse into a fundamental distinction between normal stem cells and CSCs, which has thus far eluded conventional gene expression analyses. In studying APA at the single-cell level in cancer, we seem to have opened an exciting new dimension of CSC biology.

4.5 Acknowledgements

We would like to apologize to authors whose directly related work may have not been cited in this manuscript. We thank Rached Alkallas and Ariel Madrigal Aguirre for their input and guidance on the bioinformatics analyses. We thank Maxime Bellefeuille for the tumor dissociation protocol troubleshooting.

4.6 Funding

This work was supported by the Canadian Institutes of Health Research (CIHR) grant (MOP-123352) to T.F.D.; Fonds de recherche du Québec Santé (FRQS) Doctoral training award, and CIHR Doctoral Research Award to H-W. T.

4.7 Materials and methods

4.7.1 Patient-derived xenograft model of ER+ breast cancer

The GCRC1971 PDX model was originally established in Savage *et al*. (Savage *et al.*, 2020). For this study, ~1 mm^3 fresh tumor fragments were orthotopically transplanted in the mammary fat pad of NSG mice (The Jackson Laboratories, Strain # 005557). Primary tumor growth was monitored by weekly caliper measurements. Tumor volumes were calculated using the formula $\pi LW^2/6$ where L is the length and W is the width of the tumor. Mammary tumors were resected at a volume of approximately 500-600 mm^3, while skull-base metastatic tumors were collected at clinical endpoints approximately 7 to 8 weeks after resection. All fresh tissue samples were placed in cold phosphate buffered saline (PBS) and immediately processed for single-cell RNA-seq library prep.

4.7.2 Single cell dissociation and sequencing

To isolate single cells, fresh tumor tissues were washed three times with cold sterile PBS containing 1% penicillin and streptomycin (PS), and minced with scalpels into ~1 mm^3 or less in size. Tumor fragments were then incubated in ~ 8 ml collagenase/dispase solution per tumor (1 mg/ml; Roche, Cat # 11097113001) containing DNAse I (50 U/ml; Biorad, Cat # 7326826) and $MgCl_2$ (1 mM) for 1 hour at 37ºC with occasional mixing by pipetting. Subsequently, the digested tissues were vigorously pipetted and passed through a 70 μm strainer to remove any remaining clumps. Cells were then centrifuged for 10 minutes at 350 g to remove the collagenase/dispase solution. Cell pellets were resuspended in 21 ml PBS+PS and 9 ml of Percoll (Sigma, Cat # 17091-01) and centrifuged at 31,000 g for 30 minutes at 4ºC to remove blood cells and debris. The layer with tumor cells was collected and counted using trypan blue. After counting, the cells were centrifuged again for 10 min at 350 g to remove the solution. Lastly, the cells were diluted to 1000

cells/μL in PBS + 0.04% bovine serum albumin (BSA), and 100 μL of the sample was used for single-cell capture.

One library was generated for each tumor sample using the Chromium Next Single Cell 3' GEM, Library & Gel Bead Kit v3.1 and Chip G Kit (10x Genomics) with a Chromium Controller following the manufacturer's instructions. The libraries were further converted using the MGIEasy Universal DNA Library Prep Set (MGI), quality controlled, and sequenced on the MGI DNBSEQ-G400 sequencer with PE100.

4.7.3 Single-cell RNA-seq data processing and gene expression analyses

Cell barcodes and UMI (universal molecular identifier) were demultiplexed and reads were mapped to the combined mm10 and hg19 reference genomes using the Cell Ranger v3.0.1 pipeline (10x Genomics). The following analyses were performed with R packages based on the Bioconductor OSCA toolbox (Amezquita *et al.*, 2020). Reads mapping to the mouse genome were first removed. Empty droplets and doublets were also filtered out with *DropletUtils* (Griffiths et al., 2018; Lun et al., 2019) and *scDblFinder* (Germain et al., 2022), respectively. Subsequently, low quality cells with UMI < 1000, total gene counts < 300, or having a proportion of mitochondrial reads > 20% were removed. Different samples were then integrated hierarchically using the top 5000 variable genes with *batchlor* (Haghverdi et al., 2018), excluding genes highly correlated with cell cycle phases assigned by *cyclone* as implemented in *scran* (Lun et al., 2016; Scialdone et al., 2015). Cells were then assigned to clusters using the Louvain method with k = 15 nearest neighbors and visualized using *ggplot2* (Wickham, 2016) after UMAP dimension reduction. Pathway analyses were performed using the package *fgsea* (Korotkevich *et al.*, 2021) with annotated cancer hallmark pathways retrieved from MSigDB v2023.1 (Liberzon *et al.*, 2015), either by the function *fgsea* for pre-ranked markers of a single cluster, or *geseca* across all cells.

The proliferation scores were calculated using ssGSEA as implemented in *GSVA* (Hänzelmann *et al.*, 2013) by supplying logcounts and proliferation gene sets (Jiang *et al.*, 2021; Venet *et al.*, 2011).

4.7.4 APA analyses in single-cell RNA-seq

To quantify APA as PDUI, we performed the scDaPars pipeline as previously described (Gao *et al.*, 2021). Specific steps are as follows. First, reads mapped to the mouse reference genome were filtered out along with reads with low-quality UMI base calls (≤ 10). Duplicated UMIs were then collapsed using UMI-tools (Smith *et al.*, 2017). Bam files from each sample were subsequently split into one bam file per cell barcode and converted to wiggle files as input to DaPars2 (Feng *et al.*, 2018; Li *et al.*, 2021) to calculate the raw PDUI scores. Lastly, the raw PDUI scores were used as input to the *scDaPars* R package to calculate the imputed PDUI as the final scores used for subsequent analyses.

To reveal APA patterns, cells were clustered based on PDUI scores using the Louvain method with the first 10 PCs and k = 20 nearest neighbors. UMAP visualization was generated with *ggplot2* while heatmaps were generated with *ComplexHeatmap* (Gu *et al.*, 2016).

Differential APA between the super-clusters was performed using pairwise Wilcoxon rank sum test implemented in *scran*. Subsequent pathway analysis of top differential APA genes was performed using g:Profiler (Kolberg *et al.*, 2023; Raudvere *et al.*, 2019).

Chapter 5:

General discussion

The handful of transcripts known to undergo APA nearly three decades ago has now greatly expanded to include most mammalian mRNAs (Edwalds-Gilbert et al., 1997; Hoque et al., 2013). The development of new experimental and bioinformatics tools also enabled the identification of pervasive and coordinated changes in APA patterns across different physiological and pathological contexts. Some of these changes in 3'UTRs were demonstrated early on to regulate aspects of mRNA stability, translation, and protein abundance (Edwalds-Gilbert et al., 1997). However, we are still far from cracking the "APA code". This is in part because the number of APA events identified in humans and other species thus far has yet to reach the level of saturation (Ye et al., 2022), and also because many key questions regarding the regulation and consequences of APA under different physiological contexts, both for specific transcripts and globally, remain open.

In this discussion, I will highlight the contribution of our findings to the field, delineate further questions, and propose important future experiments to address them.

5.1 How are context-specific global APA patterns achieved?

Many systems studies have revealed that APA profiles are specific to the cell type (Lianoglou *et al.*, 2013; Xia *et al.*, 2014; Zhang *et al.*, 2005). Certain biological processes such as proliferation and differentiation also accompany specific global changes in APA (Ji *et al.*, 2009). While we now know that the relative expression of some components in the CPA machinery can greatly affect the general trend of APA in a cell, the mechanism responsible for this effect remains unclear. For some of the CPA factors that normally exist in a complex, their expression may affect the availability of the overall functional complex, much like for CSTF64 during B cell activation (Takagaki *et al.*, 1996). Alternatively, they may affect the relative composition of the heterogeneous CPA machinery, or perhaps tip the balance of competition with other RNA processing/regulating machineries. This is exemplified in U1 snRNP telescripting, through which the splicing and CPA machineries likely

both compete and cooperate in the regulation of APA (So *et al.*, 2019). These competitive effects may also be observed between members of the same complex, such as for CFIm. In Chapter 2 we demonstrated that changes in the relative expression of CFIm68 and CFIm59 can exert opposing effects on the overall APA pattern. However, in the absence of one or the other, their functions appear to partially compensate for each other. Thus, with all the possible APA regulators in mind, the extent to which each of these mechanisms contributes to the global APA patterns, and whether the limiting CPA factors vary across cell types, are still not known. A careful examination of the changes in CPA machinery composition under different physiological contexts may provide additional insights for the coordinated context-specific APA regulations.

Part of the APA control also lies in the regulation of the APA regulators themselves. While little is known about the transcriptional and post-transcriptional control of APA regulators, in some cases they do undergo auto-regulation (Dai *et al.*, 2012). In the case of CFIm, we and others have also observed the coupled protein expression among the subunits (Chu *et al.*, 2019). CFIm auto-regulates CFIm68 APA *in vitro*, and CFIm25 expresses 3'UTR isoforms with CFIm binding motifs near PASs (Brown and Gilmartin, 2003; Sartini *et al.*, 2008). However, it is unclear whether CFIm59 protein expression is also sensitive to APA and how the auto-regulation impacts CFIm's ability to in turn regulate APA. These are some of the important remaining questions in the study of CFIm, especially considering that CFIm59 and CFIm68 can have opposing effects on APA regulation.

In Chapter 4, we noted an enrichment for energy metabolic pathways among genes expressing shortened 3'UTRs in metastatic tumor cells compared to primary tumor cells. In our data, 3'UTR shortening is surprisingly correlated with a decreasing proliferation signature. These observations raise questions about the drivers of global APA changes, and whether uncoupling

from proliferation can be specific to cell types or physiological contexts. Notably, the concept of general 3'UTR shortening being associated with proliferation was also recently challenged in the differentiation of plasma cells and syncytiotrophoblasts (Cheng *et al.*, 2020). In these cells, increasing demands in the secretory protein metabolism was the main driver of global 3'UTR shortening, and there was little association with cellular proliferative ability.

While the 3'UTR shortening associated with proliferation in human fibroblasts and epithelial cells has been attributed to the regulation of several CPA factors by the cell cycle-related E2F transcription factors (Elkon *et al.*, 2012), it is unclear how changes in the cellular metabolic functions and demands may rewire the expression of components in the CPA machinery. In Chapter 4, we identified a putative quiescent cell population which expresses the shortest overall 3'UTRs, including for mRNAs in metabolic pathways. Isolation of these cells may provide an opportunity to study a novel driver of 3'UTR shortening that would be uniquely active in a non-proliferative context, as well as its connection with metabolic reprogramming.

5.2 Impact of APA on expression beyond regulation of mRNA stability

Regulation of mRNA stability and translational efficiency are two important aspects of post-transcriptional gene regulation that ultimately dictate protein abundance. Changes in APA impacting protein abundance through altered mRNA stability is well documented for various transcripts, particularly in the context of cancer. In Chapter 3 we also demonstrated that global APA perturbation significantly affects mRNA stability but only for a select subset of transcripts. On the other hand, while we know the importance of 3'UTR sequences in the translational control of some transcripts, how global changes in APA affect the overall translational output is not well understood. Polysome profiling experiments performed under steady-state conditions suggest that the relative amount of polysomes associated with the long and short 3'UTR isoforms is dependent

on the cell type and cell cycle status (Floor and Doudna, 2016; Fu *et al.*, 2018; Spies *et al.*, 2013). For instance, longer 3'UTR isoforms associated with slightly more polysomes than shorter isoforms in NIH3T3 cells grown to a near-confluent condition, when contact inhibition is pronounced and cells exit the cell cycle, but this trend was reversed in sub-confluent cells (Fu *et al.*, 2018). Changes in the cell state may further trigger a switch in the translation status, even for the same 3'UTR isoform. For instance, this was seen with translational activation of the long *BDNF* 3'UTR isoform upon neuronal activation (Lau *et al.*, 2010). Nonetheless, likely only a subset of APA shifts that occur during the reprogramming of genetic networks significantly impact translation efficiency. To understand and identify such APA events that yield biologically meaningful changes, we reason that the incorporation of global APA perturbation is critical. We achieved this in Chapter 3 through CFIm68 depletion, and we could next complement this thesis work with polysome profiling and 3'UTR-seq of polysome gradient fractions. An integrated study joining changes in mRNA stability with translation efficiency upon global APA perturbation could reveal their relative contribution towards gene expression.

Another critical role of APA is the regulation of transcript subcellular localization. Asymmetric distribution of 3'UTR isoforms is a feature described in several polarized cell types (Goering *et al.*, 2021; Taliaferro *et al.*, 2016). However, it is not clear to what extent global changes in APA, such as those seen in pathological conditions including cancer, disrupt transcript localization and indeed contribute to diseases. To study the localization of 3'UTR isoforms at the transcriptome-wide scale, techniques like biochemical fractionation or separation of cell projections from the cell body could be coupled with 3'UTR-seq. Towards this, it may also be suitable to use a different cell model than the cell lines used in this thesis.

5.3 Genomic approaches in APA studies

Approaches to study APA at genome-wide scale have progressed vastly over the years. The advent of next-generation sequencing technologies brought about a flood of bulk RNA-seq, and more recently, growing single-cell RNA-seq (scRNA-seq) datasets that are progressively mined for APA studies. However, each method comes with its limitations. Here I will provide a critical review of the genomic approaches applied in this thesis, possible improvements to circumvent some of the limitations, and highlight new frontiers of genomic approaches for APA studies.

Throughout Chapters 2 and 3, we have implemented 3'UTR-seq, which is a variation of the 3'-enriched bulk RNA-seq protocols, to quantify APA under several conditions. 3'-enriched RNA-seq is highly sensitive for the detection of mRNA 3'-ends and often considered the "ground truth" data used for the development of computational tools that estimate APA from standard RNA-seq output. However, it still suffers from some technical biases that may compromise the accuracy of APA detection. For instance, oligo(dT) primers are often used for the capture of poly(A)-containing transcripts. This was the case in our own libraries. The possibility of mis-priming at A-rich regions within the transcripts is a known drawback of using oligo(dT) primers. It may lead to false identification of polyadenylation sites. While most of these internal priming events are effectively eliminated computationally, a small percentage of polyadenylation sites without upstream known PAS hexamers can be systematically lost. Nonetheless, for ours and like many other studies, the strong confidence in the APA events that were called was efficient in tracing changes in APA trends. To further strengthen the confidence of polyadenylation site identification, alternative 3'-enriched RNA-seq protocols that bypass the use of oligo(dT) primers could be used.

All bulk RNA-seq methods also share the obvious limitation of averaging sequencing reads from many cells that may be in different states, of heterogeneous types, or cells that were treated

or perturbed to different extents during an experiment. This averaging effect hampers the ability to detect gene expression changes of lesser magnitudes, or shifts that only occur in a subset of cells. These problems led to the development of scRNA-seq technologies. Many of the scRNA-seq protocols specifically capture mRNA 3'-ends and thus are perfectly suited for profiling APA events. This is the strategy that was adopted and implemented in Chapter 4. However, as scRNA-seq technologies currently stand, high levels of noise and the sparsity of reads at low-usage PASs are some of the major challenges. These problems limit the confident detection of polyadenylation sites to the more abundant mRNA isoforms, which yield sufficient sequencing coverage. Nonetheless, in a preliminary benchmarking of select polyadenylation site prediction tools, all tested scRNA-seq tools outperformed bulk RNA-seq tools (applied on standard RNA-seq data) in accuracy and consistency (Ye *et al.*, 2022). While this may be due to the advantage of enrichment in 3'-end reads produced by scRNA-seq compared to standard RNA-seq, it demonstrates the confidence and promise of these approaches.

Different methods of APA quantification have also been developed to capture global APA changes. Currently there are three main methods that are used. The first method defines changes in the ratio of proximal to distal isoforms between two conditions (Li *et al.*, 2015). This type of method includes the RED score (relative expression difference) implemented in Chapters 2 and 3, for example. It is straightforward to apply in the comparison of two conditions, but not suited for more than two conditions. Additionally, it requires a prior definition of the proximal and distal transcript isoforms for each gene, which may be too simplified as many genes express more than two mRNA isoforms. The second method defines an APA score for each gene as the fraction of the longest (or shortest) isoform in all expressed transcripts of the gene (Xia *et al.*, 2014). This method is well suited for multiple conditions, such as for scRNA-seq data in Chapter 4, where we

considered each cell separately. Similarly to method 1, method 2 provides a simple metric to trace and interpret APA changes across conditions for each transcription unit as 3'UTR lengthening or shortening. However, both methods define or group isoforms as two or one group, thus losing resolution when more than two isoforms are produced by a single transcription unit. To gain more resolution, a third method may be used, where each expressed polyadenylation site is identified and individually quantified as a fraction of all expressed isoforms of the same gene. This effectively quantifies the rate of utilization for each polyadenylation site, which can then be compared to the corresponding site in a different condition (Kowalski *et al.*, 2023). While this method faithfully allows for the detection of more subtle changes beyond simple lengthening and shortening events, its results are also more complicated to interpret.

In general, and since different genomic approaches present different limitations, integration of two or more methods would be ideal to further improve confidence in the detection of APA events. One acceptable method may be to take the consensus of two different computational tools with complementary strengths and different weaknesses. Taking the intersection of predicted polyadenylation sites and annotated sites in current databases is another possibility. Integration of data from different omics approaches, bulk RNA-seq and scRNA-seq for instance, can also be done as a verification, when available. Nonetheless, these methods of integration may result in a small number of polyadenylation sites identified (limited sensitivity) since the current tools tend to yield only limited overlap in their results (Ye *et al.*, 2022), and the databases are far from being complete.

Lastly, emerging long-read sequencing technologies such as Oxford Nanopore and PacBio platforms present a new opportunity of genomic approaches to study APA (Krause *et al.*, 2019; Liu *et al.*, 2019; Polenkowski *et al.*, 2023). While there is still room for improvement in the

coverage and affordability of these technologies, they are powerful tools for interrogating the linkage between APA and other co-transcriptional events, the hallmarks of which are recognizable in full-length transcripts. For example, RNA modifications such as N6-methyladenosine (m^6A) can be captured through direct RNA sequencing by Oxford Nanopore, to investigate their potential connections with APA, which is a new and quickly expanding field of study (Chen *et al.*, 2022).

5.4 Cell adaptation in experimental and cancer contexts

In the face of changing environments, the remarkable ability of cells to adapt and prevail is the basis of evolution and one of the causes of cancer. Under selective pressure, adaptation may be reflected by the temporary reorganization in the regulatory and genetic networks, or more permanently through genetic mutations. The effects of cell adaptation are also encountered in experimental settings where immortalized cell lines are cultured on plastic dishes for extended periods of time. The genetic background of cell lines that have been passaged and maintained for decades in culture can be substantially different from the original source cells, which may be reflected in their divergent responses to chemical stimuli, for instance (Gutbier *et al.*, 2018). Nonetheless, cell adaptation can also be manifested in a shorter time frame. During our CFIm depletion experiments in Chapters 2 and 3, the observed effect on global APA change was consistently stronger in transient siRNA knockdown compared to CRISPR knockout cells. CFIm knockdown cells were harvested after 48-72 hours of siRNA treatment, while the knockout cells underwent several selective processes including lentiviral transduction, single-cell selection, and clonal expansion, which overall spanned well over a month. The differences in phenotypes and molecular impact displayed in knockdown and knockout cells have previously been described in other cell lines and animal models (El-Brolosy *et al.*, 2019; Ma *et al.*, 2019). These effects are often attributed to genetic compensation events in the knockout cells where, for instance, the

expression of homologous genes may be transcriptionally upregulated to compensate for the introduced mutation. While not all genes display an obvious phenotypic discrepancy between their knockout and knockdown models, it may be an important factor to consider in an experimental design and for appropriate interpretations of the data. Contrasting our results with even more acute depletion of APA regulators such as CFIm68 using protein degron systems, for example, would likely yield insight into the potential role of APA in systems adaptation.

Successful cell adaptation is a crucial part of cancer onset and metastasis, as well as a root cause of chemoresistance. Metastasis constitutes the primary cause of death among cancer patients (Steeg, 2006). However, the immense selective pressure faced by the tumor cells during metastasis makes it a highly inefficient process, with only less than 0.1% of tumor cells released in circulation ultimately establishing metastasis in a distant location (Luzzi *et al.*, 1998). Some of the adaptive challenges include physical barriers (need for motility and withstanding blood pressure), immune surveillance, and nutrient deprivation. Much of the cell adaptation elicited by these changes in the tumor environment is manifested in the rewiring of cancer genetic networks. Expression changes for hundreds of genes are reported to comprise a "metastasis signature" in primary tumor cells that determine their metastatic potential, namely the combination of cancer cell traits that enables metastatic dissemination (Hunter *et al.*, 2003). In contrast, genomic studies comparing metastatic to primary tumors also identified many specific somatic mutations in metastatic tumors that promote invasion and metastasis, including for *PTEN*, *TP53*, and *PIK3CA* (Birkbak and McGranahan, 2020; Robinson *et al.*, 2017).

Similarly, APA signatures constituted by changes in select mRNA panels identified in primary tumors have been associated with increased invasion and metastatic potential (Huang *et al.*, 2023; Miles *et al.*, 2016; Wang *et al.*, 2016). However, little is known about APA changes from

matching primary to metastatic tumors, or their contribution to cancer cell adaptation through the metastasis process. Our work in Chapter 4 starts to address this and points to the potential role of APA in the metabolic reprogramming of metastatic cancer cells in a model of breast cancer. Our work provides a promising framework; similar analysis can now be applied to other models, of different cancer types, or with different treatment outcomes and clinical profiles.

An additional consideration that may further our understanding for the role of APA in metastasis lies in the study of the tumor microenvironment, which is comprised of a complex ecosystem of heterogeneous cell types including stromal and immune cells. Crosstalk between cells in the microenvironment and tumor cells improves the survival, development, and overall adaptation of the cancer cells (de Visser and Joyce, 2023). Recent studies leveraging single-cell sequencing technologies revealed distinct changes in APA patterns, primarily in general 3'UTR shortening, in different cell types within the tumor microenvironment compared to the corresponding normal tissues (Burri and Zavolan, 2021; Huang *et al.*, 2023; Kim *et al.*, 2019). This finding supports the importance of APA regulation in shaping the tumor microenvironment. It also raises the question of how APA may contribute to shaping the metastatic tumor microenvironment or niche, and hence aid in metastatic cancer cell adaptation. This question may be addressed by comparing APA profiles of the different cell types that constitute both the metastatic and primary tumor microenvironments. Of note, mouse microenvironments were systematically eliminated from the individual profiles of human tumor cells in Chapter 4. Instead of a limitation of our approach, this may be an opportunity for a future series of analyses. When established in an immunocompetent PDX model, computational resolution of the mouse stromal components from within primary and metastatic tumor profiles would allow a sharp focus on the microenvironment APA dynamics.

5.5 Conclusion

The work presented in this thesis deepens our understanding of the regulation and impact of APA on gene expression and mRNA stability, as well as on the broader genetic networks relevant to cancer and metastasis. Through biochemical and genomic approaches, our results also reveal the perils of generalizing or averaging APA dynamics and patterns, as every 3'UTR is unique and physiological contexts matter. Overall, my thesis provides an important new perspective on the interpretation of APA functions and regulations. Just as I was able to leverage the growing number of computational tools and emerging sequencing technologies available, a deeper, more extensive survey of APA across conditions should soon decipher the APA code.

References

The Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res*, **49**, D325-d334.

Agarwal, V., Bell, G.W., Nam, J.-W. and Bartel, D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, **4**, e05005.

Agarwal, V., Lopez-Darwin, S., Kelley, D.R. and Shendure, J. (2021) The landscape of alternative polyadenylation in single cells of the developing mouse embryo. *Nature Communications*, **12**, 5101.

Akamatsu, W., Okano, H.J., Osumi, N., Inoue, T., Nakamura, S., Sakakibara, S.-I., Miura, M., Matsuo, N., Darnell, R.B. and Okano, H. (1999) Mammalian ELAV-like neuronal RNA-binding proteins HuB and HuC promote neuronal development in both the central and the peripheral nervous systems. *Proceedings of the National Academy of Sciences*, **96**, 9885-9890.

Al-Hajj, M., Wicha, M.S., Benito-Hernandez, A., Morrison, S.J. and Clarke, M.F. (2003) Prospective identification of tumorigenic breast cancer cells. *Proceedings of the National Academy of Sciences*, **100**, 3983-3988.

Alimonti, A., Carracedo, A., Clohessy, J.G., Trotman, L.C., Nardella, C., Egia, A., Salmena, L., Sampieri, K., Haveman, W.J., Brogi, E. *et al.* (2010) Subtle variations in Pten dose determine cancer susceptibility. *Nature genetics*, **42**, 454-458.

Alkan, S.A., Martincic, K. and Milcarek, C. (2006) The hnRNPs F and H2 bind to similar sequences to influence gene expression. *Biochemical Journal*, **393**, 361-371.

Alt, F.W., Bothwell, A.L., Knapp, M., Siden, E., Mather, E., Koshland, M. and Baltimore, D. (1980) Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. *Cell*, **20**, 293-301.

Amara, S.G., Jonas, V., Rosenfeld, M.G., Ong, E.S. and Evans, R.M. (1982) Alternative RNA processing in calcitonin gene expression generates mRNAs encoding different polypeptide products. *Nature*, **298**, 240-244.

Amezquita, R.A., Lun, A.T.L., Becht, E., Carey, V.J., Carpp, L.N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C. *et al.* (2020) Orchestrating single-cell analysis with Bioconductor. *Nature Methods*, **17**, 137-145.

An, J.J., Gharami, K., Liao, G.-Y., Woo, N.H., Lau, A.G., Vanevski, F., Torre, E.R., Jones, K.R., Feng, Y., Lu, B. *et al.* (2008) Distinct role of long 3'UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal n
eurons. *Cell*, **134**, 175-187.

Anders, S., Reyes, A. and Huber, W. (2012) Detecting differential usage of exons from RNA-seq data. *Genome Research*, **22**, 2008-2017.

Anderson, J.S.J. and Parker, R. (1998) The 3' to 5' degradation of yeast mRNAs is a general mechanism for mRNA turnover that requires the SKI2 DEVH box protein and 3' to 5' exonucleases of the exosome complex. *The EMBO Journal*, **17**, 1497-1506.

Andres, S.F., Williams, K.N., Plesset, J.B., Headd, J.J., Mizuno, R., Chatterji, P., Lento, A.A., Klein-Szanto, A.J., Mick, R., Hamilton, K.E. *et al.* (2018) IMP1 3 ´ UTR shortening enhances metastatic burden in colorectal cancer. *Carcinogenesis*, **40**, 569-579.

Arribere, J.A. and Fire, A.Z. (2018) Nonsense mRNA suppression via nonstop decay. *eLife*, **7**, e33292.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, **25**, 25-29.

Aw, Jong Ghut A., Shen, Y., Wilm, A., Sun, M., Lim, Xin N., Boon, K.-L., Tapsin, S., Chan, Y.-S., Tan, C.-P., Sim, Adelene Y.L. *et al.* (2016) In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Molecular cell*, **62**, 603-617.

Bagga, P.S., Arhin, G.K. and Wilusz, J. (1998) DSEF-1 is a member of the hnRNP H family of RNA-binding proteins and stimulates pre-mRNA cleavage and polyadenylation in vitro. *Nucleic acids research*, **26**, 5343-5350.

Bailey, T.L. (2021) STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, **37**, 2834-2840.

Bakheet, T., Williams, B.R. and Khabar, K.S. (2006) ARED 3.0: the large and diverse AU-rich transcriptome. *Nucleic Acids Res*, **34**, D111-114.

Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215-233.

Baserga, S.J. and Steitz, J.A. (1993) The diverse world of small ribonucleoproteins. *Cold Spring Harbor Monograph Series*, **24**, 359-359.

Bashirullah, A., Cooperstock, R.L. and Lipshitz, H.D. (2001) Spatial and temporal control of RNA stability. *Proceedings of the National Academy of Sciences*, **98**, 7025-7028.

Bashirullah, A., Halsell, S.R., Cooperstock, R.L., Kloc, M., Karaiskakis, A., Fisher, W.W., Fu, W., Hamilton, J.K., Etkin, L.D. and Lipshitz, H.D. (1999) Joint action of two RNA degradation pathways controls the timing of maternal transcript elimination at the midblastula transition in Drosophila melanogaster. *The EMBO Journal*, **18**, 2610-2620.

Beaudoin, J.D. and Perreault, J.P. (2013) Exploring mRNA 3'-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening. *Nucleic Acids Res*, **41**, 5898-5911.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M. and Gautheret, D. (2000) Patterns of variant polyadenylation signal usage in human genes. *Genome Res*, **10**, 1001-1010.

Becalska, A.N., Kim, Y.R., Belletier, N.G., Lerit, D.A., Sinsimer, K.S. and Gavis, E.R. (2011) Aubergine is a component of a nanos mRNA localization complex. *Dev Biol*, **349**, 46-52.

Berg, M.G., Singh, L.N., Younis, I., Liu, Q., Pinto, A.M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L. *et al.* (2012) U1 snRNP determines mRNA length and regulates isoform expression. *Cell*, **150**, 53-64.

Bergalet, J. and Lécuyer, E. (2014) The functions and regulatory principles of mRNA intracellular trafficking. *Advances in experimental medicine and biology*, **825**, 57-96.

Berkovits, B.D. and Mayr, C. (2015) Alternative 3'UTRs act as scaffolds to regulate membrane protein localization. *Nature*, **522**, 363-367.

Bernstein, D., Hook, B., Hajarnavis, A., Opperman, L. and Wickens, M. (2005) Binding specificity and mRNA targets of a C. elegans PUF protein, FBF-1. *RNA (New York, N.Y.)*, **11**, 447-458.

Besse, F., López de Quinto, S., Marchand, V., Trucco, A. and Ephrussi, A. (2009) Drosophila PTB promotes formation of high-order RNP particles and represses oskar translation. *Genes & Development*, **23**, 195-207.

Beug, S.T., Cheung, H.H., LaCasse, E.C. and Korneluk, R.G. (2012) Modulation of immune signalling by inhibitors of apoptosis. *Trends Immunol*, **33**, 535-545.

Binder, R., Horowitz, J., Basilion, J., Koeller, D., Klausner, R. and Harford, J. (1994) Evidence that the pathway of transferrin receptor mRNA degradation involves an endonucleolytic cleavage within the 3 ´ UTR and does not involve poly (A) tail shortening. *The EMBO Journal*, **13**, 1969-1980.

Birkbak, N.J. and McGranahan, N. (2020) Cancer Genome Evolutionary Trajectories in Metastasis. *Cancer cell*, **37**, 8-19.

Blewett, N.H. and Goldstrohm, A.C. (2012) A eukaryotic translation initiation factor 4E-binding protein promotes mRNA decapping and is required for PUF repression. *Mol Cell Biol*, **32**, 4181-4194.

Boo, S.H. and Kim, Y.K. (2020) The emerging role of RNA modifications in the regulation of mRNA stability. *Experimental & Molecular Medicine*, **52**, 400-408.

Boreikaite, V., Elliott, T.S., Chin, J.W. and Passmore, L.A. (2022) RBBP6 activates the pre-mRNA 3' end processing machinery in humans. *Genes Dev*, **36**, 210-224.

Boreikaitė, V. and Passmore, L.A. (2023) 3'-End Processing of Eukaryotic mRNA: Machinery, Regulation, and Impact on Gene Expression. *Annual Review of Biochemistry*, **92**, 199-225.

Briata, P., Forcales, S.V., Ponassi, M., Corte, G., Chen, C.-Y., Karin, M., Puri, P.L. and Gherzi, R. (2005) p38-dependent phosphorylation of the mRNA decay-promoting factor KSRP controls the stability of select myogenic transcripts. *Molecular cell*, **20**, 891-903.

Broderick, J.A., Salomon, W.E., Ryder, S.P., Aronin, N. and Zamore, P.D. (2011) Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing. *RNA (New York, N.Y.)*, **17**, 1858-1869.

Brown, K.M. and Gilmartin, G.M. (2003) A Mechanism for the Regulation of Pre-mRNA 3' Processing by Human Cleavage Factor Im. *Molecular cell*, **12**, 1467-1476.

Brumbaugh, J., Di Stefano, B., Wang, X., Borkent, M., Forouzmand, E., Clowers, K.J., Ji, F., Schwarz, B.A., Kalocsay, M., Elledge, S.J. *et al.* (2018) Nudt21 Controls Cell Fate by Connecting Alternative Polyadenylation to Chromatin Signaling. *Cell*, **172**, 106-120.e121.

Bullock, S.L. and Ish-Horowicz, D. (2001) Conserved signals and machinery for RNA transport in Drosophila oogenesis and embryogenesis. *Nature*, **414**, 611-616.

Burri, D. and Zavolan, M. (2021) Shortening of 3p UTRs in most cell types composing tumor tissues implicates alternative polyadenylation in protein metabolism. *RNA (New York, N.Y.)*.

Cammas, A., Sanchez, B.J., Lian, X.J., Dormoy-Raclet, V., van der Giessen, K., de Silanes, I.L., Ma, J., Wilusz, C., Richardson, J., Gorospe, M. *et al.* (2014) Destabilization of nucleophosmin mRNA by the HuR/KSRP complex is required for muscle fibre formation. *Nature Communications*, **5**, 4190.

Campbell, Z.T., Bhimsaria, D., Valley, C.T., Rodriguez-Martinez, J.A., Menichelli, E., Williamson, J.R., Ansari, A.Z. and Wickens, M. (2012) Cooperativity in RNA-protein interactions: global analysis of RNA binding specificity. *Cell reports*, **1**, 570-581.

Carracedo, A., Alimonti, A. and Pandolfi, P.P. (2011) PTEN Level in Tumor Suppression: How Much Is Too Little? *Cancer Research*, **71**, 629-633.

Chan, S.L., Huppertz, I., Yao, C., Weng, L., Moresco, J.J., Yates, J.R., 3rd, Ule, J., Manley, J.L. and Shi, Y. (2014) CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3' processing. *Genes Dev*, **28**, 2370-2380.

Chang, J.-W., Zhang, W., Yeh, H.-S., de Jong, E.P., Jun, S., Kim, K.-H., Bae, S.S., Beckman, K., Hwang, T.H., Kim, K.-S. *et al.* (2015) mRNA 3'-UTR shortening is a molecular signature of mTORC1 activation. *Nature Communications*, **6**, 7218.

Chartron, J.W., Hunt, K.C.L. and Frydman, J. (2016) Cotranslational signal-independent SRP preloading during membrane targeting. *Nature*, **536**, 224-228.

Chen, C.Y., Chen, S.T., Juan, H.F. and Huang, H.C. (2012) Lengthening of 3'UTR increases with morphological complexity in animal evolution. *Bioinformatics*, **28**, 3178-3181.

Chen, C.Y., Gherzi, R., Ong, S.E., Chan, E.L., Raijmakers, R., Pruijn, G.J., Stoecklin, G., Moroni, C., Mann, M. and Karin, M. (2001) AU binding proteins recruit the exosome to degrade ARE-containing mRNAs. *Cell*, **107**, 451-464.

Chen, F., MacDonald, C.C. and Wilusz, J. (1995) Cleavage site determinants in the mammalian polyadenylation signal. *Nucleic Acids Res*, **23**, 2614-2620.

Chen, F. and Wilusz, J. (1998) Auxiliary downstream elements are required for efficient polyadenylation of mammalian pre-mRNAs. *Nucleic Acids Res*, **26**, 2891-2898.

Chen, L., Fu, Y., Hu, Z., Deng, K., Song, Z., Liu, S., Li, M., Ou, X., Wu, R., Liu, M. *et al.* (2022) Nuclear m6A reader YTHDC1 suppresses proximal alternative polyadenylation sites by interfering with the 3 ´ processing machinery. *EMBO reports*, **23**, e54686.

Chen, Q., Zhu, C. and Jin, Y. (2020) The Oncogenic and Tumor Suppressive Functions of the Long Noncoding RNA MALAT1: An Emerging Controversy. *Frontiers in Genetics*, **11**.

Chen, S., Wang, R., Zheng, D., Zhang, H., Chang, X., Wang, K., Li, W., Fan, J., Tian, B. and Cheng, H. (2019) The mRNA Export Receptor NXF1 Coordinates Transcriptional Dynamics, Alternative Polyadenylation, and mRNA Export. *Molecular cell*, **74**, 118-131.e117.

Chen, W., Jia, Q., Song, Y., Fu, H., Wei, G. and Ni, T. (2017) Alternative Polyadenylation: Methods, Findings, and Impacts. *Genomics, Proteomics & Bioinformatics*, **15**, 287-300.

Cheng, J., Maier, K.C., Avsec, Ž., Rus, P. and Gagneur, J. (2017) Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast. *RNA (New York, N.Y.)*, **23**, 1648-1659.

Cheng, L.C., Zheng, D., Baljinnyam, E., Sun, F., Ogami, K., Yeung, P.L., Hoque, M., Lu, C.-W., Manley, J.L. and Tian, B. (2020) Widespread transcript shortening through alternative polyadenylation in secretory cell differentiation. *Nature Communications*, **11**, 3182.

Cheung, T.H. and Rando, T.A. (2013) Molecular regulation of stem cell quiescence. *Nature Reviews Molecular Cell Biology*, **14**, 329-340.

Chu, Y., Elrod, N., Wang, C., Li, L., Chen, T., Routh, A., Xia, Z., Li, W., Wagner, E.J. and Ji, P. (2019) Nudt21 regulates the alternative polyadenylation of Pak1 and is predictive in the prognosis of glioblastoma patients. *Oncogene*, **38**, 4154-4168.

Chuvpilo, S., Zimmer, M., Kerstan, A., Glöckner, J., Avots, A., Escher, C., Fischer, C., Inashkina, I., Jankevics, E., Berberich-Siebelt, F. *et al.* (1999) Alternative Polyadenylation Events Contribute to the Induction of NF-ATc in Effector T Cells. *Immunity*, **10**, 261-269.

Clevers, H. (2011) The cancer stem cell: premises, promises and challenges. *Nature medicine*, **17**, 313-319.

Colombo, M., Karousis, E.D., Bourquin, J., Bruggmann, R. and Mühlemann, O. (2017) Transcriptome-wide identification of NMD-targeted human mRNAs reveals extensive redundancy between SMG6- and SMG7-mediated degradation pathways. *RNA (New York, N.Y.)*, **23**, 189-201.

Consortium, T.G.O., Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L. *et al.* (2023) The Gene Ontology knowledgebase in 2023. *Genetics*, **224**.

Cortazar, M.A., Sheridan, R.M., Erickson, B., Fong, N., Glover-Cutter, K., Brannan, K. and Bentley, D.L. (2019) Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a "Sitting Duck Torpedo" Mechanism. *Molecular cell*, **76**, 896-908.e894.

Crick, F. (1970) Central Dogma of Molecular Biology. *Nature*, **227**, 561-563.

Cui, Y. and Denis, C.L. (2003) In Vivo Evidence that Defects in the Transcriptional Elongation Factors RPB2, TFIIS, and SPT5 Enhance Upstream Poly(A) Site Utilization. *Molecular and Cellular Biology*, **23**, 7887-7901.

Dai, W., Zhang, G. and Makeyev, E.V. (2012) RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage. *Nucleic Acids Research*, **40**, 787-800.

Danckwardt, S., Gehring, N.H., Neu-Yilik, G., Hundsdoerfer, P., Pforsich, M., Frede, U., Hentze, M.W. and Kulozik, A.E. (2004) The prothrombin 3´ end formation signal reveals a unique architecture that is sensitive to thrombophilic gain-of-function mutations. *Blood*, **104**, 428-435.

de Klerk, E., Venema, A., Anvar, S.Y., Goeman, J.J., Hu, O., Trollet, C., Dickson, G., den Dunnen, J.T., van der Maarel, S.M., Raz, V. *et al.* (2012) Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation. *Nucleic Acids Research*, **40**, 9089-9101.

de Visser, K.E. and Joyce, J.A. (2023) The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth. *Cancer cell*, **41**, 374-403.

de Vries, H., Rüegsegger, U., Hübner, W., Friedlein, A., Langen, H. and Keller, W. (2000) Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. *Embo j*, **19**, 5895-5904.

DeJesus-Hernandez, M., Mackenzie, I.R., Boeve, B.F., Boxer, A.L., Baker, M., Rutherford, N.J., Nicholson, A.M., Finch, N.A., Flynn, H. and Adamson, J. (2011) Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron*, **72**, 245-256.

Derti, A., Garrett-Engele, P., Macisaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M. and Babak, T. (2012) A quantitative atlas of polyadenylation in five mammals. *Genome Res*, **22**, 1173-1183.

Dettwiler, S., Aringhieri, C., Cardinale, S., Keller, W. and Barabino, S.M. (2004) Distinct sequence motifs within the 68-kDa subunit of cleavage factor Im mediate RNA binding, protein-protein interactions, and subcellular localization. *The Journal of biological chemistry*, **279**, 35788-35797.

Di Giammartino, D.C., Li, W., Ogami, K., Yashinskie, J.J., Hoque, M., Tian, B. and Manley, J.L. (2014) RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3' UTRs. *Genes Dev*, **28**, 2248-2260.

Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N.J., Van Nostrand, E.L., Pratt, G.A. *et al.* (2018) Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular cell*, **70**, 854-867.e859.

Dubbury, S.J., Boutz, P.L. and Sharp, P.A. (2018) CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. *Nature*, **564**, 141-145.

Dupuy, F., Tabariès, S., Andrzejewski, S., Dong, Z., Blagih, J., Annis, Matthew G., Omeroglu, A., Gao, D., Leung, S., Amir, E. *et al.* (2015) PDK1-Dependent Metabolic Reprogramming Dictates Metastatic Potential in Breast Cancer. *Cell Metabolism*, **22**, 577-589.

Edwalds-Gilbert, G., Veraldi, K.L. and Milcarek, C. (1997) Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res*, **25**, 2547-2561.

El-Brolosy, M.A., Kontarakis, Z., Rossi, A., Kuenne, C., Günther, S., Fukuda, N., Kikhi, K., Boezio, G.L.M., Takacs, C.M., Lai, S.-L. *et al.* (2019) Genetic compensation triggered by mutant mRNA degradation. *Nature*, **568**, 193-197.

Elkon, R., Drost, J., van Haaften, G., Jenal, M., Schrier, M., Vrielink, J. and Agami, R. (2012) E2F mediates enhanced alternative polyadenylation in proliferation. *Genome Biol*, **13**, R59.

Engel, K.L., Arora, A., Goering, R., Lo, H.-Y.G. and Taliaferro, J.M. (2020) Mechanisms and consequences of subcellular RNA localization across diverse cell types. *Traffic*, **21**, 404-418.

Eom, T., Antar, L.N., Singer, R.H. and Bassell, G.J. (2003) Localization of a beta-actin messenger ribonucleoprotein complex with zipcode-binding protein modulates the density of dendritic filopodia and filopodial synapses. *J Neurosci*, **23**, 10433-10444.

Esquela-Kerscher, A. and Slack, F.J. (2006) Oncomirs — microRNAs with a role in cancer. *Nature Reviews Cancer*, **6**, 259-269.

Fang, S., Zhang, D., Weng, W., Lv, X., Zheng, L., Chen, M., Fan, X., Mao, J., Mao, C., Ye, Y. *et al.* (2020) CPSF7 regulates liver cancer growth and metastasis by facilitating WWP2-FL and targeting the WWP2/PTEN/AKT signaling pathway. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, **1867**, 118624.

Fansler, M.M., Mitschka, S. and Mayr, C. (2023) Comprehensive annotation of 3 ´ UTRs from primary cells and their quantification from scRNA-seq data. *bioRxiv*, 2021.2011.2022.469635.

Feng, X., Li, L., Wagner, E.J. and Li, W. (2018) TC3A: The Cancer 3' UTR Atlas. *Nucleic Acids Res*, **46**, D1027-d1030.

Fernandes, N. and Buchan, J.R. (2020) RPS28B mRNA acts as a scaffold promoting cis-translational interaction of proteins driving P-body assembly. *Nucleic Acids Research*, **48**, 6265-6279.

Ferrandon, D., Elphick, L., Nüsslein-Volhard, C. and St Johnston, D. (1994) Staufen protein associates with the 3'UTR of bicoid mRNA to form particles that move in a microtubule-dependent manner. *Cell*, **79**, 1221-1232.

Ferrandon, D., Koch, I., Westhof, E. and Nüsslein-Volhard, C. (1997) RNA-RNA interaction is required for the formation of specific bicoid mRNA 3' UTR-STAUFEN ribonucleoprotein particles. *Embo j*, **16**, 1751-1758.

Filipowicz, W., Bhattacharyya, S.N. and Sonenberg, N. (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature Reviews Genetics*, **9**, 102.

Fischer, J.W., Busa, V.F., Shao, Y. and Leung, A.K.L. (2020) Structure-Mediated RNA Decay by UPF1 and G3BP1. *Molecular cell*, **78**, 70-84.e76.

Flamand, M.N., Gan, H.H., Mayya, V.K., Gunsalus, K.C. and Duchaine, T.F. (2017) A non-canonical site reveals the cooperative mechanisms of microRNA-mediated silencing. *Nucleic Acids Research*, **45**, 7212-7225.

Flavell, S.W., Kim, T.-K., Gray, J.M., Harmin, D.A., Hemberg, M., Hong, E.J., Markenscoff-Papadimitriou, E., Bear, D.M. and Greenberg, M.E. (2008) Genome-Wide Analysis of MEF2 Transcriptional Program Reveals Synaptic Target Genes and Neuronal Activity-Dependent Polyadenylation Site Selection. *Neuron*, **60**, 1022-1038.

Floor, S.N. and Doudna, J.A. (2016) Tunable protein synthesis by transcript isoforms in human cells. *eLife*, e10921.

Forrest, K.M. and Gavis, E.R. (2003) Live Imaging of Endogenous RNA Reveals a Diffusion and Entrapment Mechanism for nanos mRNA Localization in Drosophila. *Current Biology*, **13**, 1159-1168.

Friedman, R.C., Farh, K.K.-H., Burge, C.B. and Bartel, D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, **19**, 92-105.

Fu, Y., Chen, L., Chen, C., Ge, Y., Kang, M., Song, Z., Li, J., Feng, Y., Huo, Z., He, G. *et al.* (2018) Crosstalk between alternative polyadenylation and miRNAs in the regulation of protein translational efficiency. *Genome Research*, **28**, 1656-1663.

Fu, Y., Sun, Y., Li, Y., Li, J., Rao, X., Chen, C. and Xu, A. (2011) Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Research*, **21**, 741-747.

Gandin, V., Sikström, K., Alain, T., Morita, M., McLaughlan, S., Larsson, O. and Topisirovic, I. (2014) Polysome fractionation and analysis of mammalian translatomes on a genome-wide scale. *JoVE (Journal of Visualized Experiments)*, e51455-e51455.

Gangras, P., Gallagher, T.L., Parthun, M.A., Yi, Z., Patton, R.D., Tietz, K.T., Deans, N.C., Bundschuh, R., Amacher, S.L. and Singh, G. (2020) Zebrafish rbm8a and magoh mutants reveal EJC developmental functions and new 3'UTR intron-containing NMD targets. *PLOS Genetics*, **16**, e1008830.

Gao, Y., Li, L., Amos, C.I. and Li, W. (2021) Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression. *Genome Research*.

Gebert, L.F. and MacRae, I.J. (2019) Regulation of microRNA function in animals. *Nature reviews Molecular cell biology*, **20**, 21-37.

Gehring, N.H., Frede, U., Neu-Yilik, G., Hundsdoerfer, P., Vetter, B., Hentze, M.W. and Kulozik, A.E. (2001) Increased efficiency of mRNA 3' end formation: a new genetic mechanism contributing to hereditary thrombophilia. *Nature genetics*, **28**, 389-392.

Geisberg, J.V., Moqtaderi, Z., Fan, X., Ozsolak, F. and Struhl, K. (2014) Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell*, **156**, 812-824.

Gennarino, V.A., Alcott, C.E., Chen, C.A., Chaudhury, A., Gillentine, M.A., Rosenfeld, J.A., Parikh, S., Wheless, J.W., Roeder, E.R., Horovitz, D.D. *et al.* (2015) NUDT21-spanning CNVs lead to neuropsychiatric disease and altered MeCP2 abundance via alternative polyadenylation. *Elife*, **4**.

Gennarino, V.A., Alcott, C.E., Chen, C.A., Chaudhury, A., Gillentine, M.A., Rosenfeld, J.A., Parikh, S., Wheless, J.W., Roeder, E.R., Horovitz, D.D. *et al.* (2015) NUDT21-spanning CNVs

lead to neuropsychiatric disease and altered MeCP2 abundance via alternative polyadenylation. *eLife*, **4**, e10782.

Germain, P., Lun, A., Garcia Meixide, C., Macnair, W. and Robinson, M. (2022) Doublet identification in single-cell sequencing data using scDblFinder [version 2; peer review: 2 approved]. *F1000Res*, **10**.

Geuens, T., Bouhy, D. and Timmerman, V. (2016) The hnRNP family: insights into their role in health and disease. *Hum Genet*, **135**, 851-867.

Ghosh, S., Ataman, M., Bak, M., Börsch, A., Schmidt, A., Buczak, K., Martin, G., Dimitriades, B., Herrmann, Christina J., Kanitz, A. *et al.* (2022) CFIm-mediated alternative polyadenylation remodels cellular signaling and miRNA biogenesis. *Nucleic Acids Research*.

Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C. *et al.* (2021) The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, **50**, D687-D692.

Go, C.D., Knight, J.D.R., Rajasekharan, A., Rathod, B., Hesketh, G.G., Abe, K.T., Youn, J.Y., Samavarchi-Tehrani, P., Zhang, H., Zhu, L.Y. *et al.* (2021) A proximity-dependent biotinylation map of a human cell. *Nature*, **595**, 120-124.

Goering, R., Engel, K.L., Gillen, A.E., Fong, N., Bentley, D.L. and Taliaferro, J.M. (2021) LABRAT reveals association of alternative polyadenylation with transcript localization, RNA binding protein expression, transcription speed, and cancer survival. *BMC Genomics*, **22**, 476.

Goldstrohm, A.C., Hall, T.M.T. and McKenney, K.M. (2018) Post-transcriptional Regulatory Functions of Mammalian Pumilio Proteins. *Trends Genet*, **34**, 972-990.

Goldstrohm, A.C., Hook, B.A., Seay, D.J. and Wickens, M. (2006) PUF proteins bind Pop2p to regulate messenger RNAs. *Nature Structural & Molecular Biology*, **13**, 533-539.

Graham, R.R., Kyogoku, C., Sigurdsson, S., Vlasova, I.A., Davies, L.R.L., Baechler, E.C., Plenge, R.M., Koeuth, T., Ortmann, W.A., Hom, G. *et al.* (2007) Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proceedings of the National Academy of Sciences*, **104**, 6758-6763.

Gregersen, L.H., Mitter, R., Ugalde, A.P., Nojima, T., Proudfoot, N.J., Agami, R., Stewart, A. and Svejstrup, J.Q. (2019) SCAF4 and SCAF8, mRNA Anti-Terminator Proteins. *Cell*, **177**, 1797-1813.e1718.

Griffiths, J.A., Richard, A.C., Bach, K., Lun, A.T.L. and Marioni, J.C. (2018) Detection and removal of barcode swapping in single-cell RNA-seq data. *Nature Communications*, **9**, 2667.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P. and Bartel, D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell*, **27**, 91-105.

Gromak, N., West, S. and Proudfoot, N.J. (2006) Pause sites promote transcriptional termination of mammalian RNA polymerase II. *Mol Cell Biol*, **26**, 3986-3996.

Grosso, A.R., de Almeida, S.F., Braga, J. and Carmo-Fonseca, M. (2012) Dynamic transitions in RNA polymerase II density profiles during transcription termination. *Genome Research*, **22**, 1447-1456.

Gruber, A.J., Gypas, F., Riba, A., Schmidt, R. and Zavolan, M. (2018) Terminal exon characterization with TECtool reveals an abundance of cell-specific isoforms. *Nature Methods*, **15**, 832-836.

Gruber, A.J., Schmidt, R., Ghosh, S., Martin, G., Gruber, A.R., van Nimwegen, E. and Zavolan, M. (2018) Discovery of physiological and cancer-related regulators of 3 ´ UTR processing with KAPAC. *Genome Biology*, **19**, 44.

Gruber, A.J., Schmidt, R., Gruber, A.R., Martin, G., Ghosh, S., Belmadani, M., Keller, W. and Zavolan, M. (2016) A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res*, **26**, 1145-1159.

Gruber, A.J. and Zavolan, M. (2019) Alternative cleavage and polyadenylation in health and disease. *Nature Reviews Genetics*, **20**, 599-614.

Gruber, A.R., Martin, G., Keller, W. and Zavolan, M. (2012) Cleavage factor Im is a key regulator of 3' UTR length. *RNA biology*, **9**, 1405-1412.

Gu, Z., Eils, R. and Schlesner, M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847-2849.

Guan, W.-L., Jiang, L.-L., Yin, X.-F. and Hu, H.-Y. (2023) PABPN1 aggregation is driven by Ala expansion and poly(A)-RNA binding, leading to CFIm25 sequestration that impairs alternative polyadenylation. *Journal of Biological Chemistry*, **299**.

Gunderson, S.I., Beyer, K., Martin, G., Keller, W., Boelens, W.C. and Mattaj, L.W. (1994) The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase. *Cell*, **76**, 531-541.

Gunderson, S.I., Polycarpou-Schwarz, M. and Mattaj, I.W. (1998) U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70K and poly(A) polymerase. *Molecular cell*, **1**, 255-264.

Gutbier, S., May, P., Berthelot, S., Krishna, A., Trefzer, T., Behbehani, M., Efremova, L., Delp, J., Gstraunthaler, G., Waldmann, T. *et al.* (2018) Major changes of cell function and toxicant sensitivity in cultured cells undergoing mild, quasi-natural genetic drift. *Archives of Toxicology*, **92**, 3487-3503.

Ha, K.C.H., Blencowe, B.J. and Morris, Q. (2018) QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biology*, **19**, 45.

Haghverdi, L., Lun, A.T.L., Morgan, M.D. and Marioni, J.C. (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, **36**, 421-427.

Hänzelmann, S., Castelo, R. and Guinney, J. (2013) GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, **14**, 7.

Harrison, B.J., Flight, R.M., Gomes, C., Venkat, G., Ellis, S.R., Sankar, U., Twiss, J.L., Rouchka, E.C. and Petruska, J.C. (2014) IB4-binding sensory neurons in the adult rat express a novel 3' UTR-extended isoform of CaMK4 that is associated with its localization to axons. *J Comp Neurol*, **522**, 308-336.

Hashimshony, T., Wagner, F., Sher, N. and Yanai, I. (2012) CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell reports*, **2**, 666-673.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S. *et al.* (2012) A census of human soluble protein complexes. *Cell*, **150**, 1068-1081.

He, L. (2010) Posttranscriptional regulation of PTEN dosage by noncoding RNAs. *Science signaling*, **3**, pe39.

Heck, A.M. and Wilusz, J. (2018) The Interplay between the RNA Decay and Translation Machinery in Eukaryotes. *Cold Spring Harbor perspectives in biology*, **10**.

Hilgers, V., Lemke, S.B. and Levine, M. (2012) ELAV mediates 3 ´ UTR extension in the Drosophila nervous system. *Genes & Development*, **26**, 2259-2264.

Hocine, S., Singer, R.H. and Grünwald, D. (2010) RNA processing and export. *Cold Spring Harbor perspectives in biology*, **2**, a000752.

Hofacker, I.L. (2003) Vienna RNA secondary structure server. *Nucleic acids research*, **31**, 3429-3431.

Hoffman, Y., Bublik, D.R., P. Ugalde, A., Elkon, R., Biniashvili, T., Agami, R., Oren, M. and Pilpel, Y. (2016) 3'UTR Shortening Potentiates MicroRNA-Based Repression of Pro-differentiation Genes in Proliferating Human Cells. *PLOS Genetics*, **12**, e1005879.

Hogg, J.R. and Goff, S.P. (2010) Upf1 senses 3'UTR length to potentiate mRNA decay. *Cell*, **143**, 379-389.

Hong, W., Ruan, H., Zhang, Z., Ye, Y., Liu, Y., Li, S., Jing, Y., Zhang, H., Diao, L., Liang, H. *et al.* (2019) APAatlas: decoding alternative polyadenylation across human tissues. *Nucleic Acids Research*, **48**, D34-D39.

Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J.Y., Yehia, G. and Tian, B. (2013) Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat Methods*, **10**, 133-139.

Howard, J.M. and Sanford, J.R. (2015) The RNAissance family: SR proteins as multifaceted regulators of gene expression. *Wiley Interdiscip Rev RNA*, **6**, 93-110.

Huang, K., Zhang, Y., Shi, X., Yin, Z., Zhao, W., Huang, L., Wang, F. and Zhou, X. (2023) Cell-type-specific alternative polyadenylation promotes oncogenic gene expression in non-small cell lung cancer progression. *Molecular Therapy - Nucleic Acids*, **33**, 816-831.

Hubisz, M.J., Pollard, K.S. and Siepel, A. (2010) PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in Bioinformatics*, **12**, 41-51.

Hug, N., Longman, D. and Cáceres, J.F. (2016) Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Research*, **44**, 1483-1495.

Hunter, K., Welch, D.R. and Liu, E.T. (2003) Genetic background is an important determinant of metastatic potential. *Nature genetics*, **34**, 23-24.

Hussen, B.M., Kheder, R.K., Abdullah, S.T., Hidayat, H.J., Rahman, H.S., Salihi, A., Taheri, M. and Ghafouri-Fard, S. (2022) Functional interplay between long non-coding RNAs and Breast CSCs. *Cancer Cell International*, **22**, 233.

Jain, R.A. and Gavis, E.R. (2008) The Drosophila hnRNP M homolog Rumpelstiltskin regulates nanos mRNA localization. *Development*, **135**, 973-982.

Jan, C.H., Friedman, R.C., Ruby, J.G. and Bartel, D.P. (2011) Formation, regulation and evolution of Caenorhabditis elegans 3 ´ UTRs. *Nature*, **469**, 97-101.

Jenal, M., Elkon, R., Loayza-Puch, F., van Haaften, G., Kühn, U., Menzies, Fiona M., Vrielink, Joachim A.F.O., Bos, Arnold J., Drost, J., Rooijers, K. *et al.* (2012) The Poly(A)-Binding Protein Nuclear 1 Suppresses Alternative Cleavage and Polyadenylation Sites. *Cell*, **149**, 538-553.

Ji, Z., Lee, J.Y., Pan, Z., Jiang, B. and Tian, B. (2009) Progressive lengthening of 3 ´ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences*, **106**, 7028-7033.

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

Jia, Q., Nie, H., Yu, P., Xie, B., Wang, C., Yang, F., Wei, G. and Ni, T. (2019) HNRNPA1-mediated 3' UTR length changes of HN1 contributes to cancer- and senescence-associated phenotypes. *Aging*, **11**, 4407-4437.

Jiang, C. and Pugh, B.F. (2009) Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics*, **10**, 161-172.

Jiang, Z., Generoso, S.F., Badia, M., Payer, B. and Carey, L.B. (2021) A conserved expression signature predicts growth rate and reveals cell & lineage-specific differences. *PLOS Computational Biology*, **17**, e1009582.

Jobbins, A.M., Haberman, N., Artigas, N., Amourda, C., Paterson, H.A.B., Yu, S., Blackford, S.J.I., Montoya, A., Dore, M., Wang, Y.-F. *et al.* (2022) Dysregulated RNA polyadenylation contributes to metabolic impairment in non-alcoholic fatty liver disease. *Nucleic Acids Research*, **50**, 3379-3393.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L. and Dreyfuss, G. (2010) U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, **468**, 664-668.

Kamieniarz-Gdula, K., Gdula, M.R., Panser, K., Nojima, T., Monks, J., Wiśniewski, J.R., Riepsaame, J., Brockdorff, N., Pauli, A. and Proudfoot, N.J. (2019) Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination. *Molecular cell*, **74**, 158-172.e159.

Kanehisa, M. (2019) Toward understanding the origin and evolution of cellular organisms. *Protein Science*, **28**, 1947-1951.

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. and Tanabe, M. (2020) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research*, **49**, D545-D551.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, **28**, 27-30.

Karousis, E.D. and Mühlemann, O. (2019) Nonsense-Mediated mRNA Decay Begins Where Translation Ends. *Cold Spring Harbor perspectives in biology*, **11**.

Kaufmann, I., Martin, G., Friedlein, A., Langen, H. and Keller, W. (2004) Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *Embo j*, **23**, 616-626.

Kaye, J.A., Rose, N.C., Goldsworthy, B., Goga, A. and Noelle, D. (2009) A 3’ UTR pumilio-binding element directs translational activation in olfactory sensory neurons. *Neuron*, **61**, 57-70.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. and Segal, E. (2007) The role of site accessibility in microRNA target recognition. *Nature genetics*, **39**, 1278-1284.

Kerwitz, Y., Kühn, U., Lilie, H., Knoth, A., Scheuermann, T., Friedrich, H., Schwarz, E. and Wahle, E. (2003) Stimulation of poly(A) polymerase through a direct interaction with the nuclear poly(A) binding protein allosterically regulated by RNA. *Embo j*, **22**, 3705-3714.

Kharchenko, P.V., Silberstein, L. and Scadden, D.T. (2014) Bayesian approach to single-cell differential expression analysis. *Nature Methods*, **11**, 740-742.

Kim, H., Huang, W., Jiang, X., Pennicooke, B., Park, P.J. and Johnson, M.D. (2010) Integrative genome analysis reveals an oncomir/oncogene cluster regulating glioblastoma survivorship. *Proceedings of the National Academy of Sciences*, **107**, 2183-2188.

Kim, N., Chung, W., Eum, H.H., Lee, H.-O. and Park, W.-Y. (2019) Alternative polyadenylation of single cells delineates cell types and serves as a prognostic marker in early stage breast cancer. *PloS one*, **14**, e0217196.

Kim, S., Yamamoto, J., Chen, Y., Aida, M., Wada, T., Handa, H. and Yamaguchi, Y. (2010) Evidence that cleavage factor Im is a heterotetrameric protein complex controlling alternative polyadenylation. *Genes to Cells*, **15**, 1003-1013.

Kim, Y.K., Furic, L., DesGroseillers, L. and Maquat, L.E. (2005) Mammalian Staufen1 Recruits Upf1 to Specific mRNA 3 ´ UTRs so as to Elicit mRNA Decay. *Cell*, **120**, 195-208.

King, P.H., Levine, T.D., Fremeau, R. and Keene, J. (1994) Mammalian homologs of Drosophila ELAV localized to a neuronal subset can bind in vitro to the 3'UTR of mRNA encoding the Id transcriptional repressor. *Journal of Neuroscience*, **14**, 1943-1952.

Kislauskis, E.H. and Singer, R.H. (1992) Determinants of mRNA localization. *Curr Opin Cell Biol*, **4**, 975-978.

Kislauskis, E.H., Zhu, X. and Singer, R.H. (1994) Sequences responsible for intracellular localization of beta-actin messenger RNA also affect cell phenotype. *J Cell Biol*, **127**, 441-451.

Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J. and Peterson, H. (2023) g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, **51**, W207-W212.

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N. and Sergushichev, A. (2021) Fast gene set enrichment analysis. *bioRxiv*, 060012.

Kowalski, M.H., Wessels, H.-H., Linder, J., Choudhary, S., Hartman, A., Hao, Y., Mascio, I., Dalgarno, C., Kundaje, A. and Satija, R. (2023) CPA-Perturb-seq: Multiplexed single-cell characterization of alternative polyadenylation regulators. *bioRxiv*, 2023.2002.2009.527751.

Krajewska, M., Dries, R., Grassetti, A.V., Dust, S., Gao, Y., Huang, H., Sharma, B., Day, D.S., Kwiatkowski, N., Pomaville, M. *et al.* (2019) CDK12 loss in cancer cells affects DNA damage response genes through premature cleavage and polyadenylation. *Nature Communications*, **10**, 1757.

Krause, M., Niazi, A.M., Labun, K., Cleuren, Y.N.T., Müller, F.S. and Valen, E. (2019) tailfindr: alignment-free poly (A) length measurement for Oxford Nanopore RNA and DNA sequencing. *RNA (New York, N.Y.)*, **25**, 1229-1241.

Kristjánsdóttir, K., Fogarty, E.A. and Grimson, A. (2015) Systematic analysis of the Hmga2 3' UTR identifies many independent regulatory sequences and a novel interaction between distal sites. *RNA (New York, N.Y.)*, **21**, 1346-1360.

Kühn, U., Gündel, M., Knoth, A., Kerwitz, Y., Rüdel, S. and Wahle, E. (2009) Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. *The Journal of biological chemistry*, **284**, 22803-22814.

Kumar, A., Clerici, M., Muckenfuss, L.M., Passmore, L.A. and Jinek, M. (2019) Mechanistic insights into mRNA 3'-end processing. *Curr Opin Struct Biol*, **59**, 143-150.

Kumar, A., Yu, C.W.H., Rodríguez-Molina, J.B., Li, X.H., Freund, S.M.V. and Passmore, L.A. (2021) Dynamics in Fip1 regulate eukaryotic mRNA 3' end processing. *Genes Dev*, **35**, 1510-1526.

Kurosaki, T., Popp, M.W. and Maquat, L.E. (2019) Quality and quantity control of gene expression by nonsense-mediated mRNA decay. *Nature Reviews Molecular Cell Biology*, **20**, 406-420.

Kwon, B., Fansler, M.M., Patel, N.D., Lee, J., Ma, W. and Mayr, C. (2022) Enhancers regulate 3' end processing activity to control expression of alternative 3'UTR isoforms. *Nature Communications*, **13**, 2709.

Kyburz, A., Friedlein, A., Langen, H. and Keller, W. (2006) Direct Interactions between Subunits of CPSF and the U2 snRNP Contribute to the Coupling of Pre-mRNA 3' End Processing and Splicing. *Molecular cell*, **23**, 195-205.

Lackford, B., Yao, C., Charles, G.M., Weng, L., Zheng, X., Choi, E.-A., Xie, X., Wan, J., Xing, Y., Freudenberg, J.M. *et al.* (2014) Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *The EMBO Journal*, **33**, 878-889.

Lai, D.-P., Tan, S., Kang, Y.-N., Wu, J., Ooi, H.-S., Chen, J., Shen, T.-T., Qi, Y., Zhang, X., Guo, Y. *et al.* (2015) Genome-wide profiling of polyadenylation sites reveals a link between selective polyadenylation and cancer metastasis. *Human Molecular Genetics*, **24**, 3410-3417.

Latham, V.M., Yu, E.H., Tullio, A.N., Adelstein, R.S. and Singer, R.H. (2001) A Rho-dependent signaling pathway operating through myosin localizes beta-actin mRNA in fibroblasts. *Curr Biol*, **11**, 1010-1016.

Lau, A.G., Irier, H.A., Gu, J., Tian, D., Ku, L., Liu, G., Xia, M., Fritsch, B., Zheng, J.Q., Dingledine, R. *et al.* (2010) Distinct 3 ´ UTRs differentially regulate activity-dependent translation of brain-derived neurotrophic factor (BDNF). *Proceedings of the National Academy of Sciences*, **107**, 15945-15950.

Lécuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T.R., Tomancak, P. and Krause, H.M. (2007) Global Analysis of mRNA Localization Reveals a Prominent Role in Organizing Cellular Architecture and Function. *Cell*, **131**, 174-187.

Lee, B.T., Barber, G.P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, C.M. *et al.* (2022) The UCSC Genome Browser database: 2022 update. *Nucleic Acids Res*, **50**, D1115-d1122.

Lee, S.-H. and Mayr, C. (2019) Gain of Additional BIRC3 Protein Functions through 3′-UTR-Mediated Protein Complex Formation. *Molecular cell*, **74**, 701-712.e709.

Lee, S.-H., Singh, I., Tisdale, S., Abdel-Wahab, O., Leslie, C.S. and Mayr, C. (2018) Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature*, **561**, 127-131.

Lee, Y.-B., Chen, H.-J., Peres, João N., Gomez-Deza, J., Attig, J., Štalekar, M., Troakes, C., Nishimura, Agnes L., Scotter, Emma L., Vance, C. *et al.* (2013) Hexanucleotide Repeats in ALS/FTD Form Length-Dependent RNA Foci, Sequester RNA Binding Proteins, and Are Neurotoxic. *Cell reports*, **5**, 1178-1186.

Levin, M., Zalts, H., Mostov, N., Hashimshony, T. and Yanai, I. (2020) Gene expression dynamics are a proxy for selective pressures on alternatively polyadenylated isoforms. *Nucleic Acids Research*, **48**, 5926-5938.

Li, H., Tong, S., Li, X., Shi, H., Ying, Z., Gao, Y., Ge, H., Niu, L. and Teng, M. (2011) Structural basis of pre-mRNA recognition by the human cleavage factor Im complex. *Cell Res*, **21**, 1039-1051.

Li, J., Yen, C., Liaw, D., Podsypanina, K., Bose, S., Wang, S.I., Puc, J., Miliaresis, C., Rodgers, L., McCombie, R. *et al.* (1997) PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science*, **275**, 1943-1947.

Li, L., Huang, K.-L., Gao, Y., Cui, Y., Wang, G., Elrod, N.D., Li, Y., Chen, Y.E., Ji, P., Peng, F. *et al.* (2021) An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nature genetics*, **53**, 994-1005.

Li, L. and Neaves, W.B. (2006) Normal Stem Cells and Cancer Stem Cells: The Niche Matters. *Cancer Research*, **66**, 4553-4557.

Li, W., Park, J.Y., Zheng, D., Hoque, M., Yehia, G. and Tian, B. (2016) Alternative cleavage and polyadenylation in spermatogenesis connects chromatin regulation with post-transcriptional control. *BMC Biology*, **14**, 6.

Li, W., You, B., Hoque, M., Zheng, D., Luo, W., Ji, Z., Park, J.Y., Gunderson, S.I., Kalsotra, A., Manley, J.L. *et al.* (2015) Systematic Profiling of Poly(A)+ Transcripts Modulated by Core 3' End Processing and Splicing Factors Reveals Regulatory Rules of Alternative Cleavage and Polyadenylation. *PLOS Genetics*, **11**, e1005166.

Li, W.V., Zheng, D., Wang, R. and Tian, B. (2021) MAAPER: model-based analysis of alternative polyadenylation using 3 ´ end-linked reads. *Genome Biology*, **22**, 222.

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

Liao, Y., Smyth, G.K. and Shi, W. (2013) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923-930.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P. and Tamayo, P. (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*, **1**, 417-425.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P. and Mesirov, J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739-1740.

Lin, Y., Li, Z., Ozsolak, F., Kim, S.W., Arango-Argoty, G., Liu, T.T., Tenenbaum, S.A., Bailey, T., Monaghan, A.P., Milos, P.M. *et al.* (2012) An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Research*, **40**, 8460-8471.

Liu, D., Brockman, J.M., Dass, B., Hutchins, L.N., Singh, P., McCarrey, J.R., MacDonald, C.C. and Graber, J.H. (2007) Systematic variation in mRNA 3'-processing signals during mouse spermatogenesis. *Nucleic Acids Res*, **35**, 234-246.

Liu, Y., Nie, H., Liu, H. and Lu, F. (2019) Poly (A) inclusive RNA isoform sequencing (PAIso− seq) reveals wide-spread non-adenosine residues within RNA poly (A) tails. *Nature communications*, **10**, 5292.

Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) ViennaRNA Package 2.0. *Algorithms for molecular biology*, **6**, 1-14.

Lou, H., Neugebauer, K.M., Gagel, R.F. and Berget, S.M. (1998) Regulation of alternative polyadenylation by U1 snRNPs and SRp20. *Mol Cell Biol*, **18**, 4977-4985.

Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550.

Lun, A., McCarthy, D. and Marioni, J. (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor [version 2; peer review: 3 approved, 2 approved with reservations]. *F1000Res*, **5**.

Lun, A.T.L., Riesenfeld, S., Andrews, T., Dao, T.P., Gomes, T., Marioni, J.C. and participants in the 1st Human Cell Atlas, J. (2019) EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biology*, **20**, 63.

Luo, Y., Na, Z. and Slavoff, S.A. (2018) P-Bodies: Composition, Properties, and Functions. *Biochemistry*, **57**, 2424-2431.

Luzzi, K.J., MacDonald, I.C., Schmidt, E.E., Kerkvliet, N., Morris, V.L., Chambers, A.F. and Groom, A.C. (1998) Multistep Nature of Metastatic Inefficiency: Dormancy of Solitary Cells after Successful Extravasation and Limited Survival of Early Micrometastases. *The American Journal of Pathology*, **153**, 865-873.

Lykke-Andersen, J. and Wagner, E. (2005) Recruitment and activation of mRNA decay enzymes by two ARE-mediated decay activation domains in the proteins TTP and BRF-1. *Genes Dev*, **19**, 351-361.

Ma, W. and Mayr, C. (2018) A Membraneless Organelle Associated with the Endoplasmic Reticulum Enables 3′UTR-Mediated Protein-Protein Interactions. *Cell*, **175**, 1492-1506.e1419.

Ma, Y., Zhu, Y., Shang, L., Qiu, Y., Shen, N., Wang, J., Adam, T., Wei, W., Song, Q., Li, J. *et al.* (2023) LncRNA XIST regulates breast cancer stem cells by activating proinflammatory IL-6/STAT3 signaling. *Oncogene*, **42**, 1419-1437.

Ma, Z., Zhu, P., Shi, H., Guo, L., Zhang, Q., Chen, Y., Chen, S., Zhang, Z., Peng, J. and Chen, J. (2019) PTC-bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS components. *Nature*, **568**, 259-263.

MacDonald, C.C. (2019) Tissue-specific mechanisms of alternative polyadenylation: Testis, brain, and beyond (2018 update). *WIREs RNA*, **10**, e1526.

Macdonald, P.M. and Kerr, K. (1997) Redundant RNA recognition events in bicoid mRNA localization. *RNA (New York, N.Y.)*, **3**, 1413-1420.

Macdonald, P.M., Kerr, K., Smith, J.L. and Leask, A. (1993) RNA regulatory element BLE1 directs the early steps of bicoid mRNA localization. *Development*, **118**, 1233-1243.

Macdonald, P.M. and Struhl, G. (1988) cis-acting sequences responsible for anterior localization of bicoid mRNA in Drosophila embryos. *Nature*, **336**, 595-598.

Macosko, Evan Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, Allison R., Kamitaki, N., Martersteck, Emily M. *et al.* (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, **161**, 1202-1214.

Mandel, C.R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J.L. and Tong, L. (2006) Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature*, **444**, 953-956.

Mansfield, K.D. and Keene, J.D. (2012) Neuron-specific ELAV/Hu proteins suppress HuR mRNA during neuronal differentiation by alternative polyadenylation. *Nucleic Acids Research*, **40**, 2734-2746.

Martin, G., Gruber, Andreas R., Keller, W. and Zavolan, M. (2012) Genome-wide Analysis of Pre-mRNA 3' End Processing Reveals a Decisive Role of Human Cleavage Factor I in the Regulation of 3' UTR Length. *Cell reports*, **1**, 753-763.

Martin, K.C. and Ephrussi, A. (2009) mRNA localization: gene expression in the spatial dimension. *Cell*, **136**, 719-730.

Martinson, H.G. (2011) An active role for splicing in 3'-end formation. *WIREs RNA*, **2**, 459-470.

Masamha, C.P. and Wagner, E.J. (2017) The contribution of alternative polyadenylation to the cancer phenotype. *Carcinogenesis*, **39**, 2-10.

Masamha, C.P., Xia, Z., Yang, J., Albrecht, T.R., Li, M., Shyu, A.-B., Li, W. and Wagner, E.J. (2014) CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature*, **510**, 412-416.

Mauger, D.M., Lin, C. and Garcia-Blanco, M.A. (2008) hnRNP H and hnRNP F complex with Fox2 to silence fibroblast growth factor receptor 2 exon IIIc. *Molecular and cellular biology*, **28**, 5403-5419.

Mayr, C. (2017) Regulation by 3'-untranslated regions. *Annual review of genetics*, **51**, 171-194.

Mayr, C. (2019) What Are 3' UTRs Doing? *Cold Spring Harbor perspectives in biology*, **11**.

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

Mayya, V.K. and Duchaine, T.F. (2015) On the availability of microRNA-induced silencing complexes, saturation of microRNA-binding sites and stoichiometry. *Nucleic Acids Research*.

Mazan-Mamczarz, K., Galbán, S., López de Silanes, I., Martindale, J.L., Atasoy, U., Keene, J.D. and Gorospe, M. (2003) RNA-binding protein HuR enhances p53 translation in response to ultraviolet light irradiation. *Proc Natl Acad Sci U S A*, **100**, 8354-8359.

Mbita, Z., Meyer, M., Skepu, A., Hosie, M., Rees, J. and Dlamini, Z. (2012) De-regulation of the RBBP6 isoform 3/DWNN in human cancers. *Mol Cell Biochem*, **362**, 249-262.

McGeary, S.E., Lin, K.S., Shi, C.Y., Pham, T.M., Bisaria, N., Kelley, G.M. and Bartel, D.P. (2019) The biochemical basis of microRNA targeting efficacy. *Science*, **366**.

McIlwain, D.R., Pan, Q., Reilly, P.T., Elia, A.J., McCracken, S., Wakeham, A.C., Itie-Youten, A., Blencowe, B.J. and Mak, T.W. (2010) Smg1 is required for embryogenesis and regulates diverse genes via alternative splicing coupled to nonsense-mediated mRNA decay. *Proceedings of the National Academy of Sciences*, **107**, 12186-12191.

Merino, D., Weber, T.S., Serrano, A., Vaillant, F., Liu, K., Pal, B., Di Stefano, L., Schreuder, J., Lin, D., Chen, Y. *et al.* (2019) Barcoding reveals complex clonal behavior in patient-derived xenografts of metastatic triple negative breast cancer. *Nature Communications*, **10**, 766.

Miles, W.O., Lembo, A., Volorio, A., Brachtel, E., Tian, B., Sgroi, D., Provero, P. and Dyson, N. (2016) Alternative Polyadenylation in Triple-Negative Breast Tumors Allows NRAS and c-JUN to Bypass PUMILIO Posttranscriptional Regulation. *Cancer Res*, **76**, 7231-7241.

Millevoi, S., Decorsière, A., Loulergue, C., Iacovoni, J., Bernat, S., Antoniou, M. and Vagner, S. (2009) A physical and functional link between splicing factors promotes pre-mRNA 3´ end processing. *Nucleic Acids Research*, **37**, 4672-4683.

Millevoi, S., Loulergue, C., Dettwiler, S., Karaa, S.Z., Keller, W., Antoniou, M. and Vagner, S. (2006) An interaction between U2AF 65 and CF Im links the splicing and 3´ end processing machineries. *The EMBO Journal*, **25**, 4854-4864.

Millevoi, S. and Vagner, S. (2010) Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic Acids Res*, **38**, 2757-2774.

Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J.O. and Lai, E.C. (2013) Widespread and extensive lengthening of 3´ UTRs in the mammalian brain. *Genome Research*.

Mohanan, N.K., Shaji, F., Koshre, G.R. and Laishram, R.S. (2022) Alternative polyadenylation: An enigma of transcript length variation in health and disease. *WIREs RNA*, **13**, e1692.

Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E. *et al.* (2003) PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, **34**, 267-273.

Motadi, L.R., Bhoola, K.D. and Dlamini, Z. (2011) Expression and function of retinoblastoma binding protein 6 (RBBP6) in human lung cancer. *Immunobiology*, **216**, 1065-1073.

Movassat, M., Crabb, T.L., Busch, A., Yao, C., Reynolds, D.J., Shi, Y. and Hertel, K.J. (2016) Coupling between alternative polyadenylation and alternative splicing is limited to terminal introns. *RNA biology*, **13**, 646-655.

Mukherjee, N., Corcoran, D.L., Nusbaum, J.D., Reid, D.W., Georgiev, S., Hafner, M., Ascano, M., Jr., Tuschl, T., Ohler, U. and Keene, J.D. (2011) Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Molecular cell*, **43**, 327-339.

Müller-McNicoll, M., Botti, V., de Jesus Domingues, A.M., Brandl, H., Schwich, O.D., Steiner, M.C., Curk, T., Poser, I., Zarnack, K. and Neugebauer, K.M. (2016) SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. *Genes & Development*, **30**, 553-566.

Nam, J.W., Rissland, O.S., Koppstein, D., Abreu-Goodger, C., Jan, C.H., Agarwal, V., Yildirim, M.A., Rodriguez, A. and Bartel, D.P. (2014) Global analyses of the effect of different cellular contexts on microRNA targeting. *Molecular cell*, **53**, 1031-1043.

Nanavaty, V., Abrash, E.W., Hong, C., Park, S., Fink, E.E., Li, Z., Sweet, T.J., Bhasin, J.M., Singuri, S., Lee, B.H. *et al.* (2020) DNA Methylation Regulates Alternative Polyadenylation via CTCF and the Cohesin Complex. *Molecular cell*, **78**, 752-764.e756.

Neve, J., Burger, K., Li, W., Hoque, M., Patel, R., Tian, B., Gullerova, M. and Furger, A. (2016) Subcellular RNA profiling links splicing and nuclear DICER1 to alternative cleavage and polyadenylation. *Genome Research*, **26**, 24-35.

Neve, J., Patel, R., Wang, Z., Louey, A. and Furger, A.M. (2017) Cleavage and polyadenylation: Ending the message expands gene regulation. *RNA biology*, **14**, 865-890.

Niwa, M., Rose, S.D. and Berget, S.M. (1990) In vitro polyadenylation is stimulated by the presence of an upstream intron. *Genes Dev*, **4**, 1552-1559.

Ogorodnikov, A., Levin, M., Tattikota, S., Tokalov, S., Hoque, M., Scherzinger, D., Marini, F., Poetsch, A., Binder, H., Macher-Göppinger, S. *et al.* (2018) Transcriptome 3´ end organization by PCF11 links alternative polyadenylation to formation and neuronal differentiation of neuroblastoma. *Nature Communications*, **9**, 5331.

Oh, J.-M., Di, C., Venters, C.C., Guo, J., Arai, C., So, B.R., Pinto, A.M., Zhang, Z., Wan, L., Younis, I. *et al.* (2017) U1 snRNP telescripting regulates a size–function-stratified human genome. *Nature Structural & Molecular Biology*, **24**, 993-999.

Oktaba, K., Zhang, W., Lotz, T.S., Jun, D.J., Lemke, S.B., Ng, S.P., Esposito, E., Levine, M. and Hilgers, V. (2015) ELAV links paused Pol II to alternative polyadenylation in the Drosophila nervous system. *Molecular cell*, **57**, 341-348.

Oleynikov, Y. and Singer, R.H. (2003) Real-time visualization of ZBP1 association with beta-actin mRNA during transcription and localization. *Curr Biol*, **13**, 199-207.

Orkin, S.H., Cheng, T.C., Antonarakis, S.E. and Kazazian, H.H., Jr. (1985) Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene. *Embo j*, **4**, 453-456.

Pa, M., Naizaer, G., Seyiti, A. and Kuerbang, G. (2017) Long Noncoding RNA MALAT1 Functions as a Sponge of MiR-200c in Ovarian Cancer. *Oncology Research*.

Pan, Z., Zhang, H., Hague, L.K., Lee, J.Y., Lutz, C.S. and Tian, B. (2006) An intronic polyadenylation site in human and mouse CstF-77 genes suggests an evolutionarily conserved regulatory mechanism. *Gene*, **366**, 325-334.

Parisi, M. and Lin, H. (1999) The Drosophila pumilio gene encodes two functional protein isoforms that play multiple roles in germline development, gonadogenesis, oogenesis and embryogenesis. *Genetics*, **153**, 235-250.

Park, H.J., Ji, P., Kim, S., Xia, Z., Rodriguez, B., Li, L., Su, J., Chen, K., Masamha, C.P., Baillat, D. *et al.* (2018) 3' UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk. *Nature genetics*, **50**, 783-789.

Passmore, L.A. and Coller, J. (2022) Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. *Nat Rev Mol Cell Biol*, **23**, 93-106.

Patrick, R., Humphreys, D.T., Janbandhu, V., Oshlack, A., Ho, J.W.K., Harvey, R.P. and Lo, K.K. (2020) Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *Genome Biology*, **21**, 167.

Peiris-Pagès, M., Martinez-Outschoorn, U.E., Pestell, R.G., Sotgia, F. and Lisanti, M.P. (2016) Cancer stem cell metabolism. *Breast Cancer Research*, **18**, 55.

Piqué, M., López, J.M., Foissac, S., Guigó, R. and Méndez, R. (2008) A Combinatorial Code for CPE-Mediated Translational Control. *Cell*, **132**, 434-448.

Polenkowski, M., Allister, A.B., Burbano de Lara, S., Soltau, M., Kendre, G. and Tran, D.D.H. (2023) Mapping alternative polyadenylation in human cells using direct RNA sequencing technology. *STAR Protocols*, **4**, 102420.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. and Siepel, A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, **20**, 110-121.

Pouremamali, F., Vahedian, V., Hassani, N., Mirzaei, S., Pouremamali, A., Kazemzadeh, H., Faridvand, Y., Jafari-gharabaghlou, D., Nouri, M. and Maroufi, N.F. (2022) The role of SOX family in cancer stem cell maintenance: With a focus on SOX2. *Pathology - Research and Practice*, **231**, 153783.

Proudfoot, N.J. (2011) Ending the message: poly(A) signals then and now. *Genes Dev*, **25**, 1770-1782.

Proudfoot, N.J. and Brownlee, G.G. (1976) 3' Non-coding region sequences in eukaryotic messenger RNA. *Nature*, **263**, 211-214.

Prudencio, M., Belzil, V.V., Batra, R., Ross, C.A., Gendron, T.F., Pregent, L.J., Murray, M.E., Overstreet, K.K., Piazza-Johnston, A.E. and Desaro, P. (2015) Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS. *Nature neuroscience*, **18**, 1175-1182.

Pullmann, R., Kim, H.H., Abdelmohsen, K., Lal, A., Martindale, J.L., Yang, X. and Gorospe, M. (2007) Analysis of Turnover and Translation Regulatory RNA-Binding Protein Expression through Binding to Cognate mRNAs. *Molecular and Cellular Biology*, **27**, 6265-6278.

Quenault, T., Lithgow, T. and Traven, A. (2011) PUF proteins: repression, activation and mRNA localization. *Trends in Cell Biology*, **21**, 104-112.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F. and Manke, T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, **44**, W160-W165.

Rana, T.M. (2007) Illuminating the silence: understanding the structure and function of small RNAs. *Nature Reviews Molecular Cell Biology*, **8**, 23-36.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H. and Vilo, J. (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, **47**, W191-W198.

Richard, P. and Manley, J.L. (2009) Transcription termination by nuclear RNA polymerases. *Genes Dev*, **23**, 1247-1269.

Robinson, D.R., Wu, Y.M., Lonigro, R.J., Vats, P., Cobain, E., Everett, J., Cao, X., Rabban, E., Kumar-Sinha, C., Raymond, V. *et al.* (2017) Integrative clinical genomics of metastatic cancer. *Nature*, **548**, 297-303.

Ross, A.F., Oleynikov, Y., Kislauskis, E.H., Taneja, K.L. and Singer, R.H. (1997) Characterization of a beta-actin mRNA zipcode-binding protein. *Mol Cell Biol*, **17**, 2158-2165.

Rozenblatt-Rosen, O., Nagaike, T., Francis, J.M., Kaneko, S., Glatt, K.A., Hughes, C.M., LaFramboise, T., Manley, J.L. and Meyerson, M. (2009) The tumor suppressor Cdc73 functionally associates with CPSF and CstF 3' mRNA processing factors. *Proceedings of the National Academy of Sciences*, **106**, 755-760.

Salmena, L., Poliseno, L., Tay, Y., Kats, L. and Pandolfi, P.P. (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, **146**, 353-358.

Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A. and Burge, C.B. (2008) Proliferating cells express mRNAs with shortened 3'untranslated regions and fewer microRNA target sites. *Science*, **320**, 1643-1647.

Sanjana, N.E., Shalem, O. and Zhang, F. (2014) Improved vectors and genome-wide libraries for CRISPR screening. *Nat Meth*, **11**, 783-784.

Sartini, B.L., Wang, H., Wang, W., Millette, C.F. and Kilpatrick, D.L. (2008) Pre-messenger RNA cleavage factor I (CFIm): potential role in alternative polyadenylation during spermatogenesis. *Biology of reproduction*, **78**, 472-482.

Savage, P., Pacis, A., Kuasne, H., Liu, L., Lai, D., Wan, A., Dankner, M., Martinez, C., Muñoz-Ramos, V., Pilon, V. *et al.* (2020) Chemogenomic profiling of breast cancer patient-derived xenografts reveals targetable vulnerabilities for difficult-to-treat tumors. *Commun Biol*, **3**, 310.

Saxton, R.A. and Sabatini, D.M. (2017) mTOR Signaling in Growth, Metabolism, and Disease. *Cell*, **168**, 960-976.

Schäfer, P., Tüting, C., Schönemann, L., Kühn, U., Treiber, T., Treiber, N., Ihling, C., Graber, A., Keller, W., Meister, G. *et al.* (2018) Reconstitution of mammalian Cleavage Factor II involved in 3' processing of mRNA precursors. *RNA (New York, N.Y.)*.

Schmidt, M., Kluge, F., Sandmeir, F., Kühn, U., Schäfer, P., Tüting, C., Ihling, C., Conti, E. and Wahle, E. (2022) Reconstitution of 3' end processing of mammalian pre-mRNA reveals a central role of RBBP6. *Genes Dev*, **36**, 195-209.

Schwerdtfeger, M., Desiderio, V., Kobold, S., Regad, T., Zappavigna, S. and Caraglia, M. (2021) Long non-coding RNAs in cancer stem cells. *Translational Oncology*, **14**, 101134.

Schwich, O.D., Blümel, N., Keller, M., Wegener, M., Setty, S.T., Brunstein, M.E., Poser, I., Mozos, I.R.L., Suess, B., Münch, C. *et al.* (2021) SRSF3 and SRSF7 modulate 3'UTR length through suppression or activation of proximal polyadenylation sites and regulation of CFIm levels. *Genome Biol*, **22**, 82.

Scialdone, A., Natarajan, K.N., Saraiva, L.R., Proserpio, V., Teichmann, S.A., Stegle, O., Marioni, J.C. and Buettner, F. (2015) Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, **85**, 54-61.

Semotok, J.L., Cooperstock, R.L., Pinder, B.D., Vari, H.K., Lipshitz, H.D. and Smibert, C.A. (2005) Smaug Recruits the CCR4/POP2/NOT Deadenylase Complex to Trigger Maternal Transcript Localization in the Early Drosophila Embryo. *Current Biology*, **15**, 284-294.

Semotok, J.L., Luo, H., Cooperstock, R.L., Karaiskakis, A., Vari, H.K., Smibert, C.A. and Lipshitz, H.D. (2008) Drosophila maternal Hsp83 mRNA destabilization is directed by multiple SMAUG recognition elements in the open reading frame. *Mol Cell Biol*, **28**, 6757-6772.

Sharova, L.V., Sharov, A.A., Nedorezov, T., Piao, Y., Shaik, N. and Ko, M.S. (2009) Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Res*, **16**, 45-58.

Shen, C., Yang, C., Xia, B. and You, M. (2021) Long non-coding RNAs: Emerging regulators for chemo/immunotherapy resistance in cancer stem cells. *Cancer Letters*, **500**, 244-252.

Shen, Z.-J., Esnault, S. and Malter, J.S. (2005) The peptidyl-prolyl isomerase Pin1 regulates the stability of granulocyte-macrophage colony-stimulating factor mRNA in activated eosinophils. *Nature immunology*, **6**, 1280-1287.

Shepard, P.J., Choi, E.-A., Lu, J., Flanagan, L.A., Hertel, K.J. and Shi, Y. (2011) Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA (New York, N.Y.)*, **17**, 761-772.

Sheridan, R.M., Fong, N., D'Alessandro, A. and Bentley, D.L. (2019) Widespread Backtracking by RNA Pol II Is a Major Effector of Gene Activation, 5' Pause Release, Termination, and Transcription Elongation Rate. *Molecular cell*, **73**, 107-118.e104.

Shi, Y., Di Giammartino, D.C., Taylor, D., Sarkeshik, A., Rice, W.J., Yates, J.R., 3rd, Frank, J. and Manley, J.L. (2009) Molecular architecture of the human pre-mRNA 3' processing complex. *Molecular cell*, **33**, 365-376.

Shulman, E.D. and Elkon, R. (2019) Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. *Nucleic Acids Research*, **47**, 10027-10039.

Singh, I., Lee, S.-H., Sperling, A.S., Samur, M.K., Tai, Y.-T., Fulciniti, M., Munshi, N.C., Mayr, C. and Leslie, C.S. (2018) Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nature Communications*, **9**, 1716.

Singh, P., Alley, T.L., Wright, S.M., Kamdar, S., Schott, W., Wilpan, R.Y., Mills, K.D. and Graber, J.H. (2009) Global Changes in Processing of mRNA 3' Untranslated Regions Characterize Clinically Distinct Cancer Subtypes. *Cancer Research*, **69**, 9422-9430.

Smith, T., Heger, A. and Sudbery, I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res*, **27**, 491-499.

So, B.R., Di, C., Cai, Z., Venters, C.C., Guo, J., Oh, J.-M., Arai, C. and Dreyfuss, G. (2019) A Complex of U1 snRNP with Cleavage and Polyadenylation Factors Controls Telescripting, Regulating mRNA Transcription in Human Cells. *Molecular cell*, **76**, 590-599.e594.

Sommerkamp, P., Altamura, S., Renders, S., Narr, A., Ladel, L., Zeisberger, P., Eiben, P.L., Fawaz, M., Rieger, M.A., Cabezas-Wallscheid, N. *et al.* (2020) Differential Alternative Polyadenylation Landscapes Mediate Hematopoietic Stem Cell Activation and Regulate Glutamine Metabolism. *Cell Stem Cell*, **26**, 722-738.e727.

Sommerkamp, P., Cabezas-Wallscheid, N. and Trumpp, A. (2021) Alternative Polyadenylation in Stem Cell Self-Renewal and Differentiation. *Trends in Molecular Medicine*, **27**, 660-672.

Song, M.S., Salmena, L. and Pandolfi, P.P. (2012) The functions and regulation of the PTEN tumour suppressor. *Nat Rev Mol Cell Biol*, **13**, 283-296.

Sowd, G.A., Serrao, E., Wang, H., Wang, W., Fadel, H.J., Poeschla, E.M. and Engelman, A.N. (2016) A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proceedings of the National Academy of Sciences*, **113**, E1054-E1063.

Spies, N., Burge, C.B. and Bartel, D.P. (2013) 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome research*, **23**, 2078-2090.

Spies, N., Nielsen, C.B., Padgett, R.A. and Burge, C.B. (2009) Biased Chromatin Signatures around Polyadenylation Sites and Exons. *Molecular cell*, **36**, 245-254.

Srikantan, S., Tominaga, K. and Gorospe, M. (2012) Functional interplay between RNA-binding protein HuR and microRNAs. *Curr Protein Pept Sci*, **13**, 372-379.

Steeg, P.S. (2006) Tumor metastasis: mechanistic insights and clinical challenges. *Nature medicine*, **12**, 895-904.

Subramanian, A., Hall, M., Hou, H., Mufteev, M., Yu, B., Yuki, K.E., Nishimura, H., Sathaseevan, A., Lant, B., Zhai, B. *et al.* (2021) Alternative polyadenylation is a determinant of oncogenic Ras function. *Sci Adv*, **7**, eabh0562.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**, 15545-15550.

Suh, N., Crittenden, S.L., Goldstrohm, A., Hook, B., Thompson, B., Wickens, M. and Kimble, J. (2009) FBF and its dual control of gld-1 expression in the Caenorhabditis elegans germline. *Genetics*, **181**, 1249-1260.

Sun, M., Ding, J., Li, D., Yang, G., Cheng, Z. and Zhu, Q. (2017) NUDT21 regulates 3'-UTR length and microRNA-mediated gene silencing in hepatocellular carcinoma. *Cancer Letters*, **410**, 158-168.

Szkop, K.J. and Nobeli, I. (2017) Untranslated Parts of Genes Interpreted: Making Heads or Tails of High-Throughput Transcriptomic Data via Computational Methods. *BioEssays*, **39**, 1700090.

Takagaki, Y. and Manley, J.L. (2000) Complex protein interactions within the human polyadenylation machinery identify a novel component. *Mol Cell Biol*, **20**, 1515-1525.

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

Takagaki, Y. and Manley, J.L. (1997) RNA recognition by the human polyadenylation factor CstF. *Mol Cell Biol*, **17**, 3907-3914.

Takagaki, Y., Ryner, L.C. and Manley, J.L. (1989) Four factors are required for 3'-end cleavage of pre-mRNAs. *Genes Dev*, **3**, 1711-1724.

Takagaki, Y., Seipelt, R.L., Peterson, M.L. and Manley, J.L. (1996) The Polyadenylation Factor CstF-64 Regulates Alternative Processing of IgM Heavy Chain Pre-mRNA during B Cell Differentiation. *Cell*, **87**, 941-952.

Taliaferro, J.M., Lambert, N.J., Sudmant, P.H., Dominguez, D., Merkin, J.J., Alexis, M.S., Bazile, C.A. and Burge, C.B. (2016) RNA sequence context effects measured in vitro predict in vivo protein binding and regulation. *Molecular cell*, **64**, 294-306.

Taliaferro, J.M., Vidaki, M., Oliveira, R., Olson, S., Zhan, L., Saxena, T., Wang, E.T., Graveley, B.R., Gertler, F.B., Swanson, M.S. *et al.* (2016) Distal Alternative Last Exons Localize mRNAs to Neural Projections. *Molecular cell*, **61**, 821-833.

Tan, S., Zhang, M., Shi, X., Ding, K., Zhao, Q., Guo, Q., Wang, H., Wu, Z., Kang, Y., Zhu, T. *et al.* (2021) CPSF6 links alternative polyadenylation to metabolism adaption in hepatocellular carcinoma progression. *J Exp Clin Cancer Res*, **40**, 85.

Tan, S., Zhang, M., Shi, X., Ding, K., Zhao, Q., Guo, Q., Wang, H., Wu, Z., Kang, Y., Zhu, T. *et al.* (2021) CPSF6 links alternative polyadenylation to metabolism adaption in hepatocellular carcinoma progression. *Journal of Experimental & Clinical Cancer Research*, **40**, 85.

Tang, H.W., Hu, Y., Chen, C.L., Xia, B., Zirin, J., Yuan, M., Asara, J.M., Rabinow, L. and Perrimon, N. (2018) The TORC1-Regulated CPA Complex Rewires an RNA Processing Network to Drive Autophagy and Metabolic Reprogramming. *Cell Metab*, **27**, 1040-1054.e1048.

Tay, Y., Kats, L., Salmena, L., Weiss, D., Tan, S.M., Ala, U., Karreth, F., Poliseno, L., Provero, P., Di Cunto, F. *et al.* (2011) Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell*, **147**, 344-357.

Tellier, M., Zaborowska, J., Neve, J., Nojima, T., Hester, S., Fournier, M., Furger, A. and Murphy, S. (2022) CDK9 and PP2A regulate RNA polymerase II transcription termination and coupled RNA maturation. *EMBO Rep*, **23**, e54520.

Terns, M.P. and Jacob, S.T. (1989) Role of Poly(A) Polymerase in the Cleavage and Polyadenylation of mRNA Precursor. *Molecular and Cellular Biology*, **9**, 1435-1444.

Thivierge, C., Tseng, H.-W., Mayya, V.K., Lussier, C., Gravel, S.-P. and Duchaine, T.F. (2018) Alternative polyadenylation confers Pten mRNAs stability and resistance to microRNAs. *Nucleic Acids Research*, **46**, 10340-10352.

Tian, B., Lutz, C.S., Zhang, H. and Hu, J. (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, **33**, 201-212.

Tian, B. and Manley, J.L. (2017) Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology*, **18**, 18.

Tian, B., Pan, Z. and Lee, J.Y. (2007) Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res*, **17**, 156-165.

Tian, L., Fang, Y.-X., Xue, J.-l. and Chen, J.-Z. (2013) Four MicroRNAs Promote Prostate Cell Proliferation with Regulation of PTEN and Its Downstream Signals In Vitro. *PloS one*, **8**, e75885.

To, K.K., Robey, R.W., Knutsen, T., Zhan, Z., Ried, T. and Bates, S.E. (2009) Escape from hsa-miR-519c enables drug-resistant cells to maintain high expression of ABCG2. *Mol Cancer Ther*, **8**, 2959-2968.

Tranter, M., Helsley, R.N., Paulding, W.R., McGuinness, M., Brokamp, C., Haar, L., Liu, Y., Ren, X. and Jones, W.K. (2011) Coordinated post-transcriptional regulation of Hsp70.3 gene expression by microRNA and alternative polyadenylation. *The Journal of biological chemistry*, **286**, 29828-29837.

Tseng, H.-W., Mota-Sydor, A., Leventis, R., Jovanovic, P., Topisirovic, I. and Duchaine, Thomas F. (2022) Distinct, opposing functions for CFIm59 and CFIm68 in mRNA alternative polyadenylation of Pten and in the PI3K/Akt signalling cascade. *Nucleic Acids Research*, **50**, 9397-9412.

Tushev, G., Glock, C., Heumüller, M., Biever, A., Jovanovic, M. and Schuman, E.M. (2018) Alternative 3'UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments. *Neuron*, **98**, 495-511.e496.

Usuki, F., Yamashita, A., Shiraishi, T., Shiga, A., Onodera, O., Higuchi, I. and Ohno, S. (2013) Inhibition of SMG-8, a subunit of SMG-1 kinase, ameliorates nonsense-mediated mRNA decay-exacerbated mutant phenotypes without cytotoxicity. *Proceedings of the National Academy of Sciences*, **110**, 15037-15042.

Vagner, S., Vagner, C. and Mattaj, I.W. (2000) The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF 65 to couple 3'-end processing and splicing. *Genes Dev*, **14**, 403-413.

Vasudevan, K.M., Gurumurthy, S. and Rangnekar, V.M. (2004) Suppression of PTEN expression by NF-kappa B prevents apoptosis. *Mol Cell Biol*, **24**, 1007-1021.

Velten, L., Anders, S., Pekowska, A., Järvelin, A.I., Huber, W., Pelechano, V. and Steinmetz, L.M. (2015) Single-cell polyadenylation site mapping reveals 3 ´ isoform choice variability. *Molecular systems biology*, **11**, 812.

Venet, D., Dumont, J.E. and Detours, V. (2011) Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. *PLOS Computational Biology*, **7**, e1002240.

Veraldi, K.L., Arhin, G.K., Martincic, K., Chung-Ganster, L.H., Wilusz, J. and Milcarek, C. (2001) hnRNP F influences binding of a 64-kilodalton subunit of cleavage stimulation factor to mRNA precursors in mouse B cells. *Mol Cell Biol*, **21**, 1228-1238.

Virolle, T., Adamson, E.D., Baron, V., Birle, D., Mercola, D., Mustelin, T. and de Belle, I. (2001) The Egr-1 transcription factor directly activates PTEN during irradiation-induced signalling. *Nat Cell Biol*, **3**, 1124-1128.

Wahl, M.C., Will, C.L. and Lührmann, R. (2009) The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell*, **136**, 701-718.

Wang, L., Hu, X., Wang, P. and Shao, Z.M. (2016) The 3'UTR signature defines a highly metastatic subgroup of triple-negative breast cancer. *Oncotarget*, **7**, 59834-59844.

Wang, L., Hu, X., Wang, P. and Shao, Z.M. (2018) Integrative 3' Untranslated Region-Based Model to Identify Patients with Low Risk of Axillary Lymph Node Metastasis in Operable Triple-Negative Breast Cancer. *The Oncologist*, **24**, 22-30.

Wang, R., Zheng, D., Wei, L., Ding, Q. and Tian, B. (2019) Regulation of Intronic Polyadenylation by PCF11 Impacts mRNA Expression of Long Genes. *Cell reports*, **26**, 2766-2778.e2766.

Wang, R., Zheng, D., Yehia, G. and Tian, B. (2018) A compendium of conserved cleavage and polyadenylation events in mammalian genes. *Genome Research*, **28**, 1427-1441.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57-63.

Ward, Patrick S. and Thompson, Craig B. (2012) Metabolic Reprogramming: A Cancer Hallmark Even Warburg Did Not Anticipate. *Cancer cell*, **21**, 297-308.

Wei, L. and Lai, E.C. (2022) Regulation of the Alternative Neural Transcriptome by ELAV/Hu RNA Binding Proteins. *Frontiers in Genetics*, **13**.

Wei, L., Lee, S., Majumdar, S., Zhang, B., Sanfilippo, P., Joseph, B., Miura, P., Soller, M. and Lai, E.C. (2020) Overlapping Activities of ELAV/Hu Family RNA Binding Proteins Specify the Extended Neuronal 3' UTR Landscape in Drosophila. *Molecular Cell*, **80**, 140-155.e146.

Wei, L., Lee, S., Majumdar, S., Zhang, B., Sanfilippo, P., Joseph, B., Miura, P., Soller, M. and Lai, E.C. (2020) Overlapping Activities of ELAV/Hu Family RNA Binding Proteins Specify the Extended Neuronal 3 ´ UTR Landscape in Drosophila. *Molecular cell*, **80**, 140-155.e146.

West, S. and Proudfoot, N.J. (2008) Human Pcf11 enhances degradation of RNA polymerase II-associated nascent RNA and transcriptional termination. *Nucleic Acids Res*, **36**, 905-914.

Whitelaw, E. and Proudfoot, N. (1986) Alpha-thalassaemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to 3' end processing in the human alpha 2 globin gene. *Embo j*, **5**, 2915-2922.

Wickens, M., Bernstein, D.S., Kimble, J. and Parker, R. (2002) A PUF family portrait: 3 ´ UTR regulation as a way of life. *Trends in Genetics*, **18**, 150-157.

Wickens, M. and Stephenson, P. (1984) Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3' end formation. *Science*, **226**, 1045-1051.

Wickham, H. (2016) *ggplot2: elegant graphics for data analysis*. Springer-Verlag New York.

Wu, M., Lin, Z., Li, X., Xin, X., An, J., Zheng, Q., Yang, Y. and Lu, D. (2016) HULC cooperates with MALAT1 to aggravate liver cancer stem cells growth through telomere repeat-binding factor 2. *Scientific reports*, **6**, 36045.

Wu, X. and Bartel, D.P. (2017) Widespread Influence of 3'-End Structures on Mammalian mRNA Processing and Stability. *Cell*, **169**, 905-917.e911.

Wu, X., Liu, T., Ye, C., Ye, W. and Ji, G. (2020) scAPAtrap: identification and quantification of alternative polyadenylation sites from single-cell RNA-seq data. *Briefings in Bioinformatics*, **22**.

Xia, Z., Donehower, L.A., Cooper, T.A., Neilson, J.R., Wheeler, D.A., Wagner, E.J. and Li, W. (2014) Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nature Communications*, **5**, 5274.

Xiang, Y., Ye, Y., Lou, Y., Yang, Y., Cai, C., Zhang, Z., Mills, T., Chen, N.-Y., Kim, Y., Muge Ozguc, F. *et al.* (2017) Comprehensive Characterization of Alternative Polyadenylation in Human Cancer. *JNCI: Journal of the National Cancer Institute*, **110**, 379-389.

Xiao, C., Srinivasan, L., Calado, D.P., Patterson, H.C., Zhang, B., Wang, J., Henderson, J.M., Kutok, J.L. and Rajewsky, K. (2008) Lymphoproliferative disease and autoimmunity in mice with increased miR-17-92 expression in lymphocytes. *Nature Immunology*, **9**, 405-414.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S. and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3 ´ UTRs by comparison of several mammals. *Nature*, **434**, 338-345.

Yamashita, A., Ohnishi, T., Kashima, I., Taya, Y. and Ohno, S. (2001) Human SMG-1, a novel phosphatidylinositol 3-kinase-related protein kinase, associates with components of the mRNA surveillance complex and is involved in the regulation of nonsense-mediated mRNA decay. *Genes Dev*, **15**, 2215-2228.

Yang, E., van Nimwegen, E., Zavolan, M., Rajewsky, N., Schroeder, M., Magnasco, M. and Darnell, J.E., Jr. (2003) Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res*, **13**, 1863-1872.

Yang, Q., Coseno, M., Gilmartin, G.M. and Doublié, S. (2011) Crystal Structure of a Human Cleavage Factor CFIm25/CFIm68/RNA Complex Provides an Insight into Poly(A) Site Recognition and RNA Looping. *Structure*, **19**, 368-377.

Yang, Q., Gilmartin, G.M. and Doublié, S. (2010) Structural basis of UGUA recognition by the Nudix protein CFI(m)25 and implications for a regulatory role in mRNA 3' processing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 10062-10067.

Yang, Q., Gilmartin, G.M. and Doublié, S. (2011) The structure of human cleavage factor I(m) hints at functions beyond UGUA-specific RNA binding: a role in alternative polyadenylation and a potential link to 5' capping and splicing. *RNA biology*, **8**, 748-753.

Yang, W., Hsu, P.L., Yang, F., Song, J.E. and Varani, G. (2018) Reconstitution of the CstF complex unveils a regulatory role for CstF-50 in recognition of 3'-end processing signals. *Nucleic Acids Res*, **46**, 493-503.

Yang, Y., Li, W., Hoque, M., Hou, L., Shen, S., Tian, B. and Dynlacht, B.D. (2016) PAF Complex Plays Novel Subunit-Specific Roles in Alternative Cleavage and Polyadenylation. *PLOS Genetics*, **12**, e1005794.

Yang, Y., Paul, A., Bach, T.N., Huang, Z.J. and Zhang, M.Q. (2021) Single-cell alternative polyadenylation analysis delineates GABAergic neuron types. *BMC Biology*, **19**, 144.

Yao, C., Biesinger, J., Wan, J., Weng, L., Xing, Y., Xie, X. and Shi, Y. (2012) Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. *Proc Natl Acad Sci U S A*, **109**, 18773-18778.

Yao, C., Choi, E.A., Weng, L., Xie, X., Wan, J., Xing, Y., Moresco, J.J., Tu, P.G., Yates, J.R., 3rd and Shi, Y. (2013) Overlapping and distinct functions of CstF64 and CstF64τ in mammalian mRNA 3' processing. *RNA (New York, N.Y.)*, **19**, 1781-1790.

Ye, C., Long, Y., Ji, G., Li, Q.Q. and Wu, X. (2018) APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics*, **34**, 1841-1849.

Ye, W., Lian, Q., Ye, C. and Wu, X. (2022) A Survey on Methods for Predicting Polyadenylation Sites from DNA Sequences, Bulk RNA-seq, and Single-cell RNA-seq. *Genomics, Proteomics & Bioinformatics*.

Yi, Z., Sanjeev, M. and Singh, G. (2021) The Branched Nature of the Nonsense-Mediated mRNA Decay Pathway. *Trends in Genetics*, **37**, 143-159.

Youn, J.Y., Dunham, W.H., Hong, S.J., Knight, J.D.R., Bashkurov, M., Chen, G.I., Bagci, H., Rathod, B., MacLeod, G., Eng, S.W.M. *et al.* (2018) High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Molecular cell*, **69**, 517-532.e511.

Yu, M., Yang, W., Ni, T., Tang, Z., Nakadai, T., Zhu, J. and Roeder, R.G. (2015) RNA polymerase II-associated factor 1 regulates the release and phosphorylation of paused RNA polymerase II. *Science*, **350**, 1383-1386.

Yuan, F., Hankey, W., Wagner, E.J., Li, W. and Wang, Q. (2021) Alternative polyadenylation of mRNA and its role in cancer. *Genes Dis*, **8**, 61-72.

Yudin, D., Hanz, S., Yoo, S., Iavnilovitch, E., Willis, D., Gradus, T., Vuppalanchi, D., Segal-Ruder, Y., Ben-Yaakov, K., Hieda, M. *et al.* (2008) Localized Regulation of Axonal RanGTPase Controls Retrograde Injury Signaling in Peripheral Nerve. *Neuron*, **59**, 241-252.

Zaborowska, J., Egloff, S. and Murphy, S. (2016) The pol II CTD: new twists in the tail. *Nature Structural & Molecular Biology*, **23**, 771-777.

Zaessinger, S., Busseau, I. and Simonelig, M. (2006) Oskar allows nanos mRNA translation in Drosophila embryos by preventing its deadenylation by Smaug/CCR4. *Development*, **133**, 4573-4583.

Zanoni, M., Bravaccini, S., Fabbri, F. and Arienti, C. (2022) Emerging Roles of Aldehyde Dehydrogenase Isoforms in Anti-cancer Therapy Resistance. *Frontiers in Medicine*, **9**.

Zhang, H., Lee, J.Y. and Tian, B. (2005) Biased alternative polyadenylation in human tissues. *Genome Biology*, **6**, R100.

Zhang, H.L., Eom, T., Oleynikov, Y., Shenoy, S.M., Liebelt, D.A., Dictenberg, J.B., Singer, R.H. and Bassell, G.J. (2001) Neurotrophin-induced transport of a beta-actin mRNP complex increases beta-actin levels and stimulates growth cone motility. *Neuron*, **31**, 261-275.

Zhang, Q. and Tian, B. (2023) The emerging theme of 3 ´ UTR mRNA isoform regulation in reprogramming of cell metabolism. *Biochemical Society Transactions*.

Zhang, X., Powell, K. and Li, L. (2020) Breast Cancer Stem Cells: Biomarkers, Identification and Isolation Methods, Regulating Mechanisms, Cellular Origin, and Beyond. *Cancers (Basel)*, **12**.

Zhang, Y., Sun, Y., Shi, Y., Walz, T. and Tong, L. (2020) Structural Insights into the Human Pre-mRNA 3'-End Processing Machinery. *Molecular cell*, **77**, 800-809.e806.

Zhao, D., Duan, H., Kim, Y.-C. and Jefcoate, C.R. (2005) Rodent StAR mRNA is substantially regulated by control of mRNA stability through sites in the 3'-untranslated region and through coupling to ongoing transcription. *The Journal of Steroid Biochemistry and Molecular Biology*, **96**, 155-173.

Zhao, W., Li, Y. and Zhang, X. (2017) Stemness-Related Markers in Cancer. *Cancer Transl Med*, **3**, 87-95.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, **8**, 14049.

Zheng, Q., Zhang, M., Zhou, F., Zhang, L. and Meng, X. (2021) The Breast Cancer Stem Cells Traits and Drug Resistance. *Frontiers in Pharmacology*, **11**.

Zhou, J., Chen, Q., Zou, Y., Chen, H., Qi, L. and Chen, Y. (2019) Stem Cells and Cellular Origins of Breast Cancer: Updates in the Rationale, Controversies, and Therapeutic Implications. *Frontiers in Oncology*, **9**.

Zhu, H., Zhou, H.-L., Hasman, R.A. and Lou, H. (2007) Hu Proteins Regulate Polyadenylation by Blocking Sites Containing U-rich Sequences*. *Journal of Biological Chemistry*, **282**, 2203-2210.

Zhu, Y., Wang, X., Forouzmand, E., Jeong, J., Qiao, F., Sowd, G.A., Engelman, A.N., Xie, X., Hertel, K.J. and Shi, Y. (2018) Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation. *Molecular cell*, **69**, 62-74.e64.

Appendix 1: Supplemental information to chapter 2

Hsin-Wei Tseng[1,2], Anthony Mota-Sydor[1,2], Rania Leventis[1,2], Predrag Jovanovic[2,3,4,5], Ivan Topisirovic[2,3,4,5], Thomas F. Duchaine[1,2,*]

[1] Rosalind and Morris Goodman Cancer Institute, McGill University, Montréal, H3G1Y6, Canada.

[2] Department of Biochemistry, McGill University, Montréal, H3G1Y6, Canada.

[3] Lady Davis Institute for Medical Research, Montréal, H3T1E2, Canada

[4] Gerald Bronfman Department of Oncology, McGill University, Montréal, H4A3T2, Canada

[5] Department of Medicine, Division of Experimental Medicine, McGill University, Montréal, H4A3J1

* Correspondence: thomas.duchaine@mcgill.ca

***Nucleic Acids Research**, 9 September 2022, doi: 10.1093/nar/gkac704*

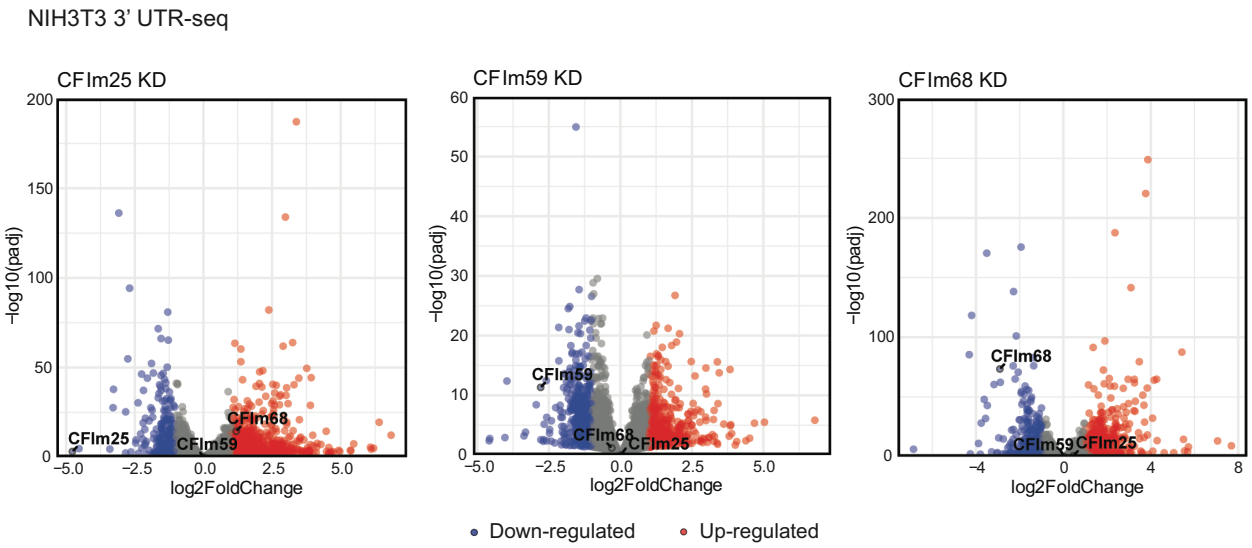


Figure A1.1: Total RNA-seq analyses of NIH3T3 CFIm KD. Datasets of 3'UTR- seq were analyzed as total RNA-seq datasets by collapsing all mRNA isoforms of each gene and tabulating the total count. Up-regulated and down-regulated genes are shown as red and blue points, with CFIm components labeled. Fold change cut-off was set at 2-fold with false discovery rate adjusted P-value (padj) threshold of 0.05.

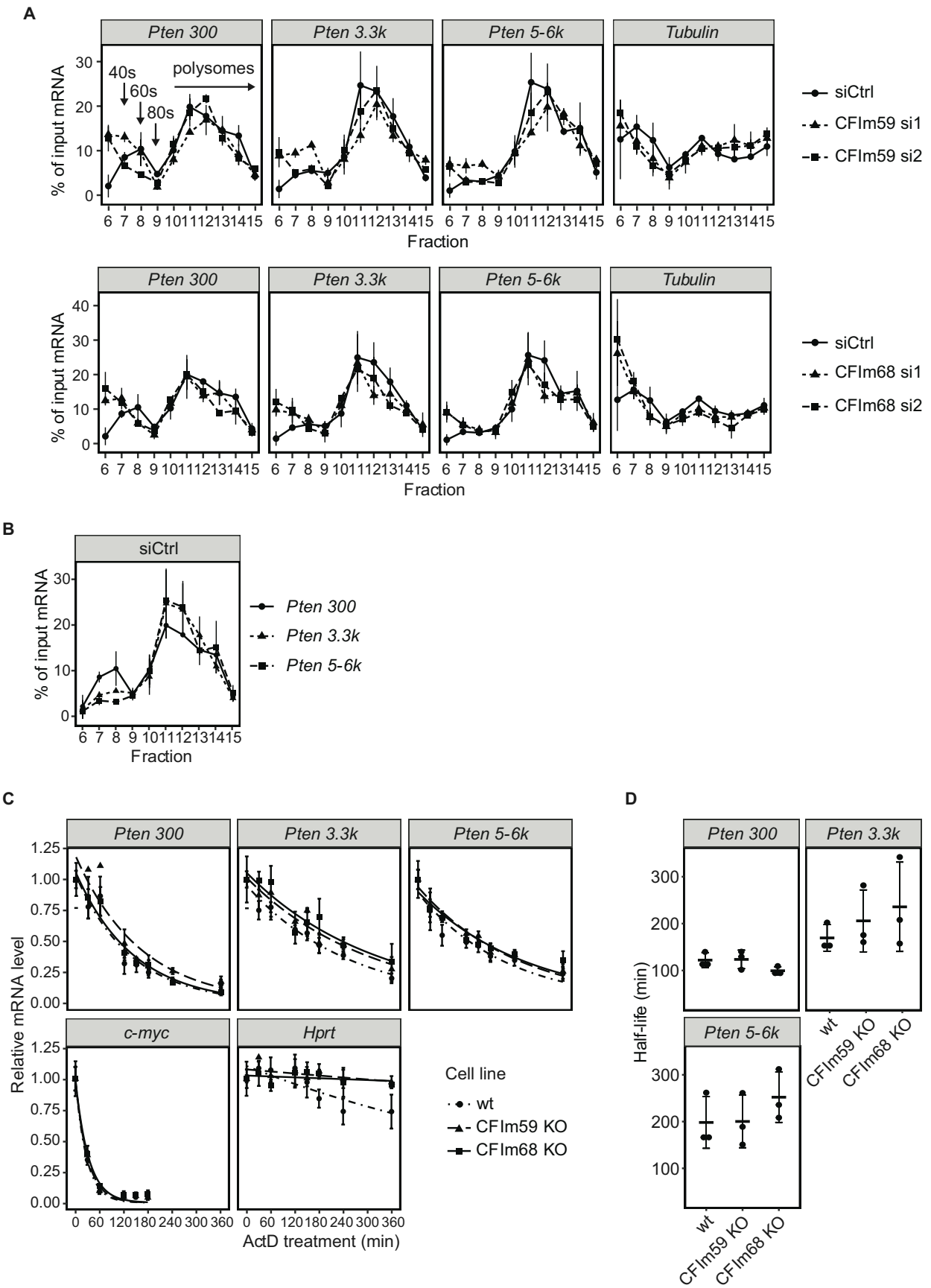

A
Pten 300
Pten 3.3k
Pten 5-6k
Tubulin
% of input mRNA
40s
60s
80s
polysomes
Fraction
siCtrl
CFIm59 si1
CFIm59 si2
CFIm68 si1
CFIm68 si2
B
siCtrl
Pten 300
Pten 3.3k
Pten 5-6k
C
c-myc
Hprt
Relative mRNA level
ActD treatment (min)
Cell line
wt
CFIm59 KO
CFIm68 KO
D
Half-life (min)

Figure A1.2: Pten mRNA translatability and stability upon CFIm59 and -68 depletion. (A) Polysome profiling of *Pten* 300 nt, 3.3k, and 5-6k isoforms upon CFIm59 KD (top) and -68 KD (bottom). Expression of each transcript was quantified by RT-qPCR across fractions as % of input mRNA. Tubulin was assayed as a control. (B) Polysome profiling of control KD as in (A) represented with different *Pten* 3'UTR isoforms superimposed. (C) *Pten* mRNA stability assay for individual *Pten* 3'UTR isoform upon CFim59 and -68 KO, as quantified by RT-qPCR. *c-Myc* was assayed as a positive control, while *Hprt* as a negative control for the Actinomycin-D treatment. (D) Half-life of different *Pten* isoforms quantified from (C). Error bars are represented as the range covered by two independent replicates in the polysome profiling experiments (A, B), and as mean ± standard deviation across three biological replicates of the mRNA stability assay (C, D).

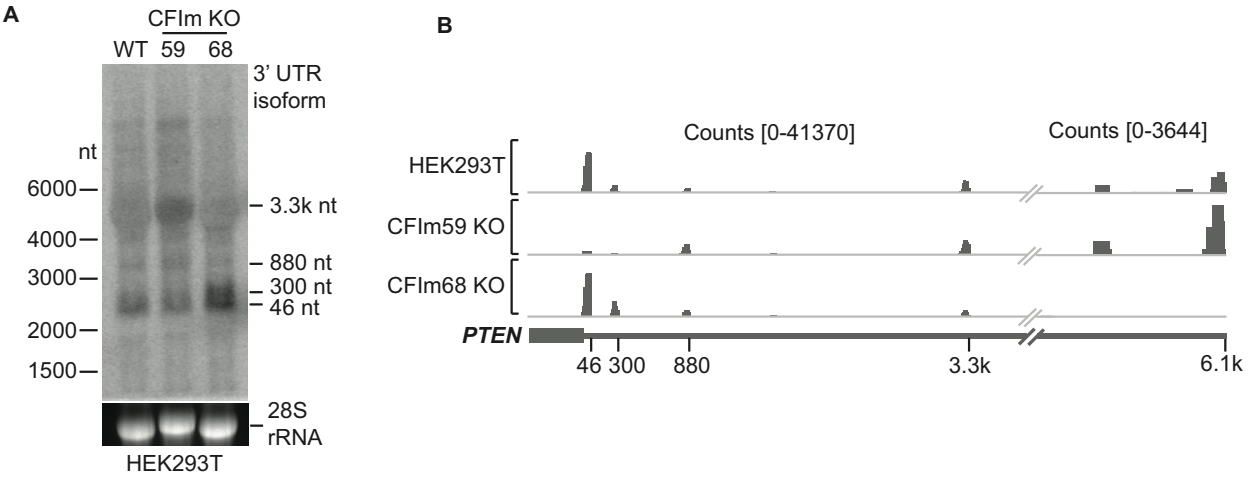


Figure A1.3: Human PTEN APA upon CFIm59 and CFIm68 KO. (A) Northern blotting of endogenous human *PTEN* in HEK293T and isogenic cell lines CFIm59 and -68 KO. Size markers are shown on the left and the identified 3'UTR isoforms on the right. (B) PAS-seq tracks of HEK293T and isogenic CFIm59 and -68 KO cell lines. Schematic of *PTEN* PAS appears on the bottom. Note that tracks are divided into two parts with distinct scales of counts (y axis) indicated on top to show all isoforms.

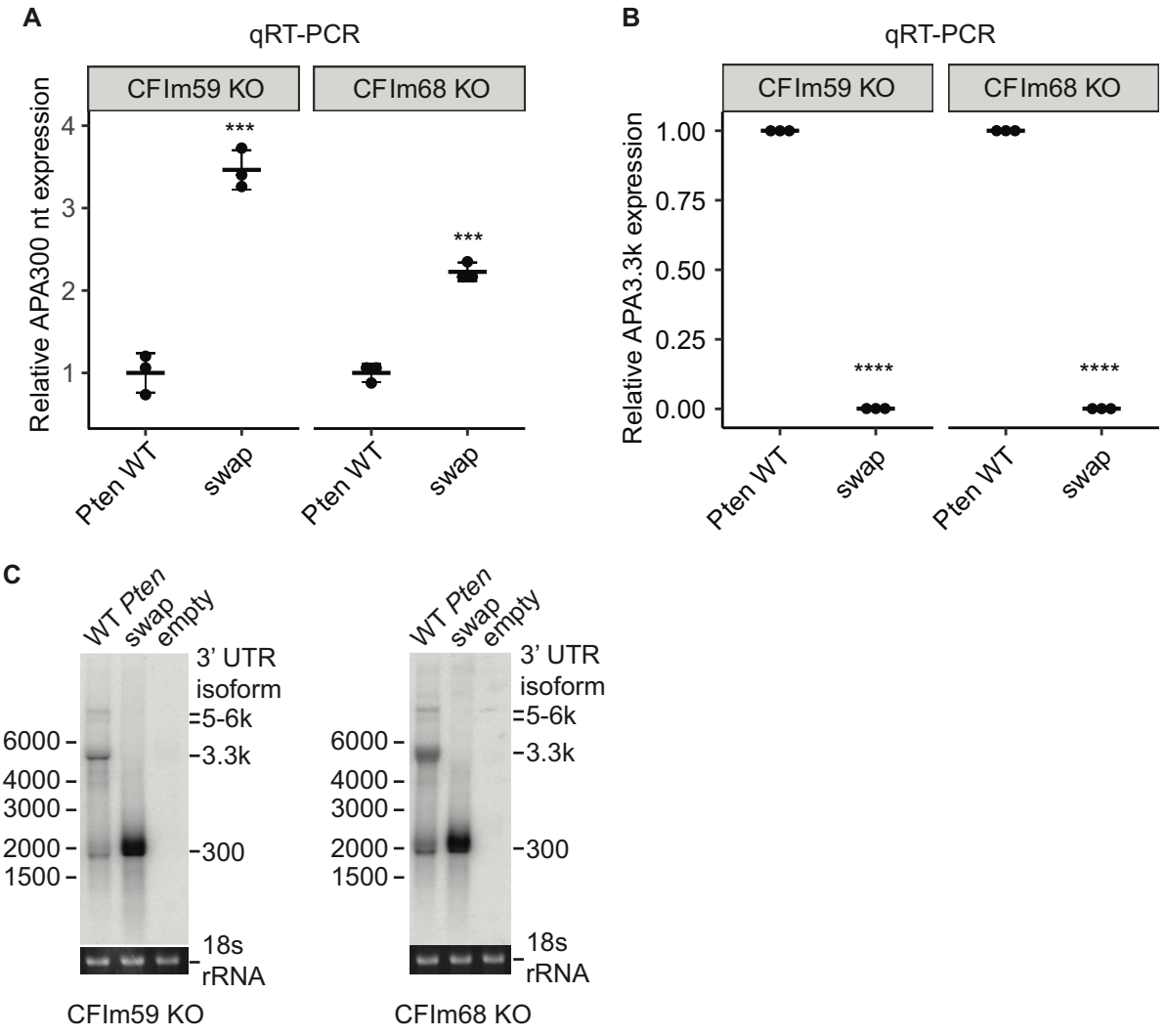


Figure A1.4: Pten swap constructs expression in CFIm59 and CFIm68 KO cells. (A, B) *Pten* APA 300 nt (A) and 3.3k (B) are quantified by RT-qPCR and normalized against *Pten* WT transfection. (C) Mouse *Pten*-specific northern blotting of samples in (A) and (B). P values were calculated with Student's *t*-test (***$P \leq 0.001$ and ****$P \leq 0.0001$).

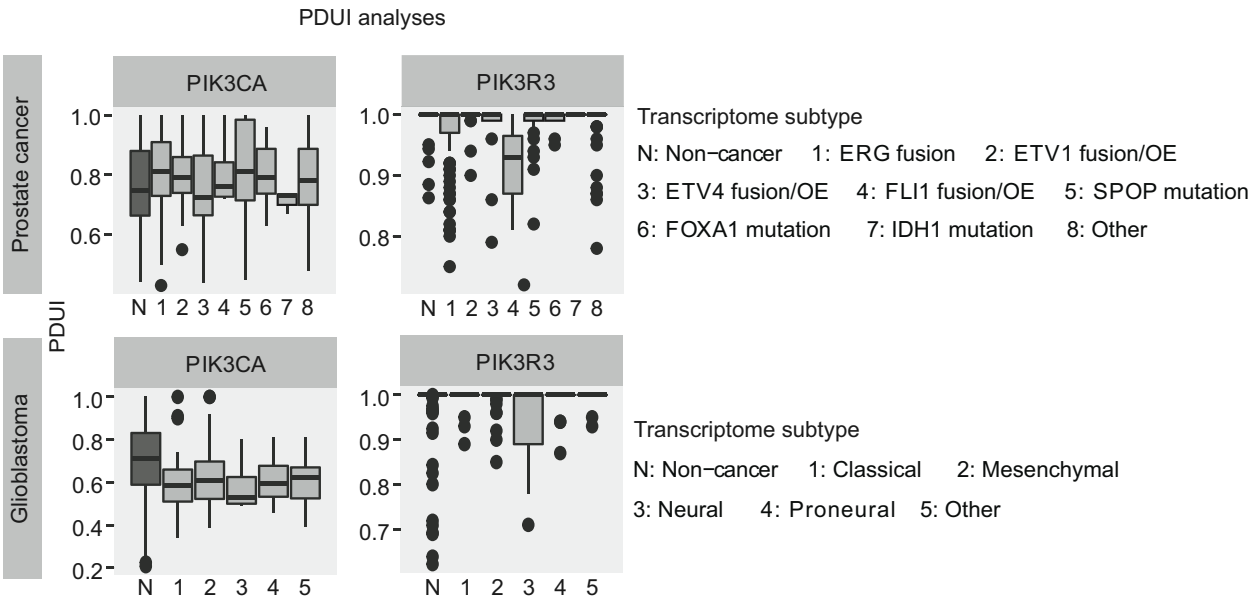


Figure A1.5: PDUI analyses of PI3K subunits in cancer. APA changes of PIK3CA and PIK3R3 in prostate cancer and glioblastoma are profiled. Cancer PDUI datasets are taken from TC3A and normal tissue PDUI taken from APAatlas. Transcriptomic subtypes of each cancer are defined by the respective TCGA projects.

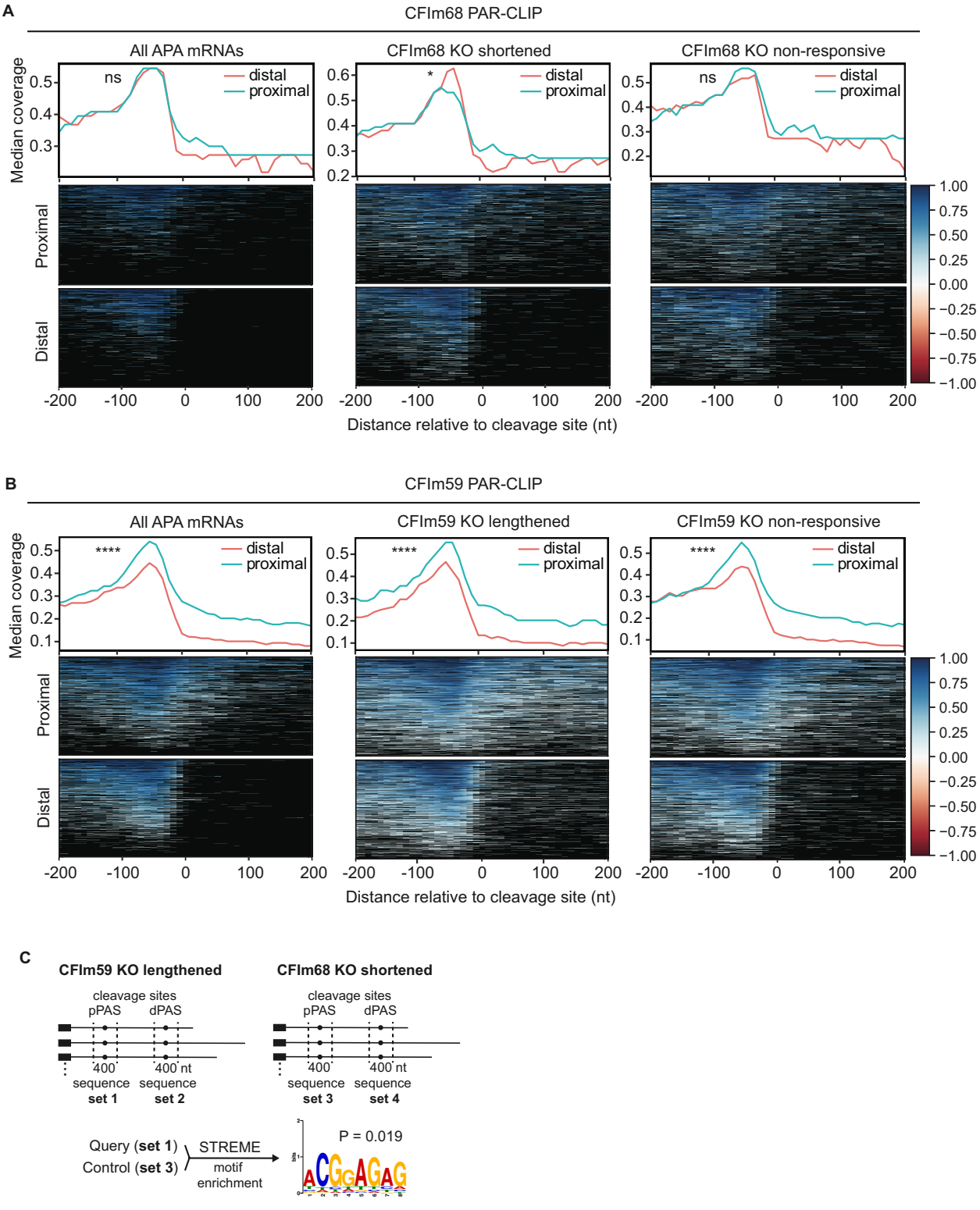


Figure A1.6: CFIm59 and -68 PAR-CLIP on APA genes and motif enrichment surrounding responsive PAS. (A) CFIm68 PAR-CLIP signals within 400 nt centered around the PAS cleavage

site for all APA mRNAs, CFIm68 KO shortened, and CFIm68 KO non-responsive mRNAs. Median coverage was traced across all nucleotides for distal and proximal PAS (top). Signal density corresponding each gene set of interest are also represented as heatmaps (bottom). (B) Same as (A) but for CFIm59 PAR-CLIP for all APA mRNAs, CFIm59 KO lengthened, and CFIm59 KO non-responsive mRNAs. (C) Motif enrichment analysis performed with sequence sets surrounding distal or proximal PAS cleavage sites of CFIm59 KO lengthened mRNAs and CFIm68 KO shortened mRNAs. The analysis gave one significant hit; a motif enriched nearby proximal PAS of CFIm59 KO lengthened mRNAs over CFIm68 KO shortened mRNAs.

Table 1.1: Oligos used.

ID	Use	Sequence
TDO679	*Pten* isoform-specific RT-qPCR: Reverse transcription	GCGAGCTCCGCGGCCGCGTTTTTTTTTTT
TDO3108	*Pten* isoform-specific RT-qPCR: PCR1 forward primer for APA 3.3k	TATGACAGTATTCACGATTAGCC
TDO886	*Pten* isoform-specific RT-qPCR: PCR1 forward primer for APA 300 nt; *Pten* ORF RT-qPCR forward primer	GCGTGCAGATAATGACAAGG
TDO680	*Pten* isoform-specific RT-qPCR: PCR1 reverse primer	CCAGTGAGCAGAGTGACG
TDO2197	*Pten* isoform-specific RT-qPCR: qPCR2 forward primer for APA 300 nt	TGGCAATAGGACATTGTGTCA
TDO3734	*Pten* isoform-specific RT-qPCR: qPCR2 reverse primer for APA 300 nt	CAA GTG TCA AAA CCC TGT GG
TDO3764	*Pten* isoform-specific RT-qPCR: qPCR2 forward primer for APA 3.3k	ACCTGCCAGCTCAAAAGTTC
TDO3765	*Pten* isoform-specific RT-qPCR: qPCR2 reverse primer for APA 3.3k	TGCTGCACAGCACAAGAGTA
TDO1610	RT-qPCR: *Hprt1* forward primer	AAGCTTGCTGGTGAAAAGGA

TDO1611	RT-qPCR: *Hprt1* reverse primer	TTGCGCTCATCTTAGGCTTT
TDO887	*Pten* ORF RT-qPCR reverse primer	TCTGGATTTGATGGCTCCTC
TDO3768	*Pten* APA 5-6k RT-qPCR forward primer	GCTCAGCAAATGCGTACCTA
TDO3769	*Pten* APA 5-6k RT-qPCR reverse primer	ACAAGTCACAGAAGCACACA
TDO7301	*Pten* APA 300 nt UGUA1 mutation forward primer	AAGGTTGAGAAGCTGTGTCATGTATATACC
TDO7340	*Pten* APA 300 nt UGUA1 mutation reverse primer	TTTTTAACTGGACAACAAGTGTCAAA
TDO7302	*Pten* APA 300 nt UGUA2 mutation forward primer	AGCTGTGTCAAGAATATACCTTTTTGTGTCAA
TDO7303	*Pten* APA 300 nt UGUA2 mutation reverse primer	ACACAACCTTTTTTTAACTGGACAAC
TDO7304	*Pten* APA 300 nt UGUA3 mutation forward primer	GGCTGATGAGAATACGCAGGAGTT

TDO7305	*Pten* APA 300 nt UGUA3 mutation reverse primer	TGTCTCCACTTTTTATAAAAC TGGAAT
TDO7306	*Pten* APA 3.3k UGUA4 mutation forward primer	GCTCTGTGAGAAAATGCTATG CACT
TDO7307	*Pten* APA 3.3k UGUA4 mutation reverse primer	CACTGCTGCACAGCACAAGA
TDO7308	*Pten* APA 3.3k UGUA5 mutation forward primer	AAATATGACGAGAACAGGAT AATGCCTC
TDO7309	*Pten* APA 3.3k UGUA5 mutation reverse primer	GTGTATCCTCAGTGCATAGCA T

Appendix 2: Supplemental information to chapter 3

Hsin-Wei Tseng[1,2], Thomas F. Duchaine[1,2,*]

[1] Rosalind and Morris Goodman Cancer Institute, McGill University, Montréal, H3G1Y6, Canada.

[2] Department of Biochemistry, McGill University, Montréal, H3G1Y6, Canada.

* Correspondence: thomas.duchaine@mcgill.ca

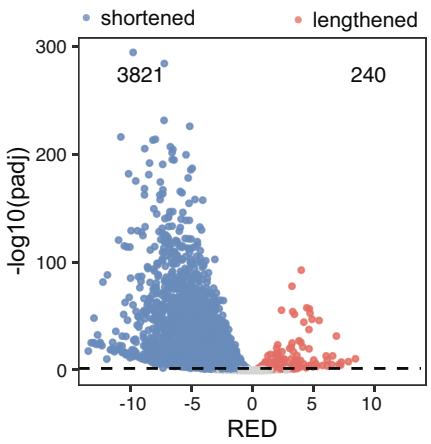


Figure A2.1: CFIm68 KO results in global 3'UTR shortening. Every point represents a transcription unit (TU). Blue represents shortened TUs while red represents lengthened TUs as determined by RED scores passing padj < 0.05 threshold. The numbers of TUs lengthened or shortened are indicated.

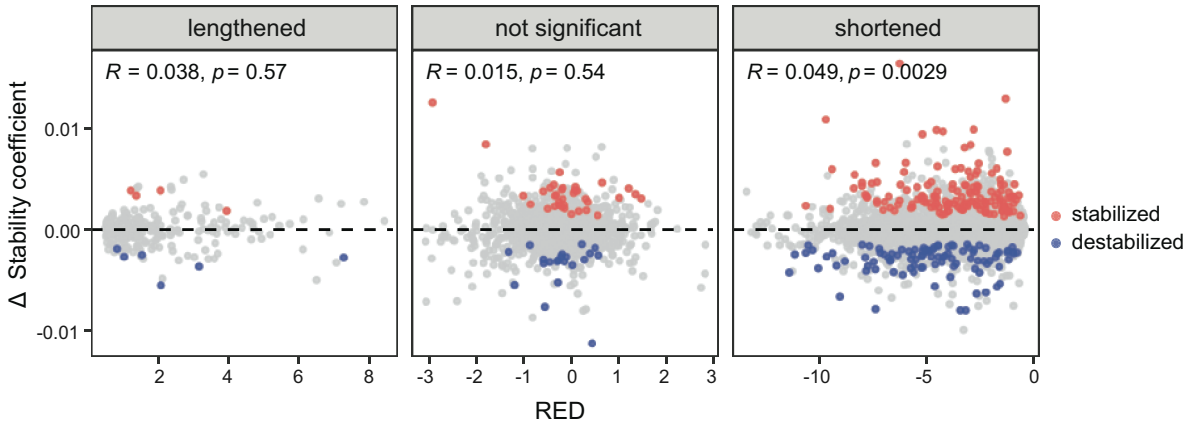


Figure A2.2: Changes in TU stability are not correlated with the magnitude nor the direction of APA change in response to CFIm68 KO. For each transcription unit (TU) that is lengthened, shortened, or not significantly changed in APA, changes in the stability coefficient are plotted against the RED scores. Red points represent TUs that are significantly stabilized upon CFIm68 KO, while blue points represent TUs that are significantly destabilized. Pearson's *R* and *p*-values are indicated.

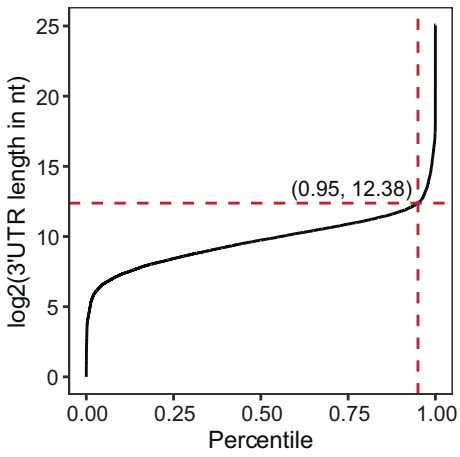


Figure A2.3: The distribution of 3'UTR lengths in wildtype cells. All transcripts are ranked and plotted according to their 3'UTR length. The 95th percentile transcript is indicated, whose 3'UTR is 2^12.38 ≈ 5330 nt in length.

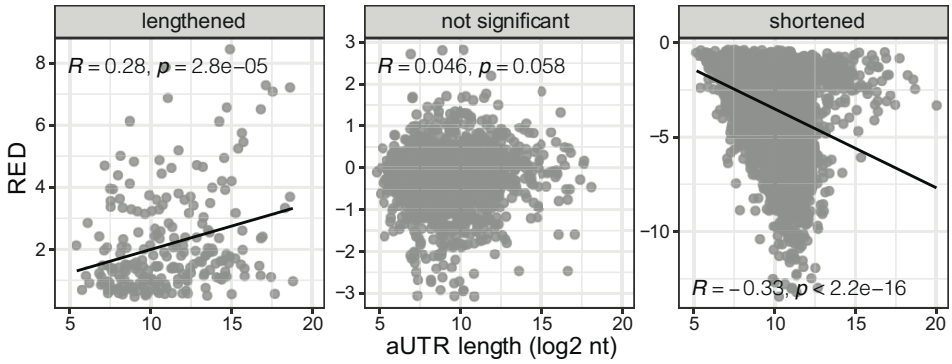


Figure A2.4: aUTR lengths correlate with the magnitude of APA shift for lengthened and shortened TUs in response to CFIm68 KO. The APA change (RED scores) and aUTR length are plotted for each transcription unit (TU) that are significantly lengthened, shortened, or not significantly changed in APA. Pearson's *R* and *p*-values are indicated.

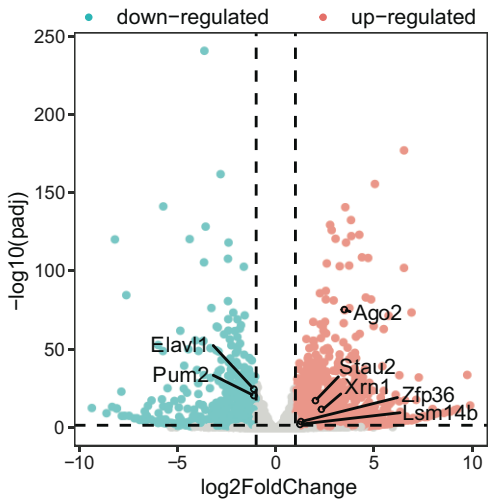


Figure A2.5: CFIm68 KO affects mRNA expression of genes in different mRNA decay pathways. Green points represent mRNAs downregulated by CFIm68 KO, while red represents upregulated mRNAs. Significantly changed mRNAs were filtered by padj < 0.05 and log2FoldChange ≥ 1.

Appendix 3: Supplemental information to chapter 4

Hsin-Wei Tseng[1,2], Phuong U. Le[4], Rima Ezzeddine[1,2], Charlotte Girondel[1], Marco Biondini[1,3], Matthew Dankner[1,3], Kevin Petrecca[4], Peter M. Siegel[1,3], Thomas F. Duchaine[1,2,*]

[1] Rosalind and Morris Goodman Cancer Institute, McGill University, Montréal, QC, Canada.

[2] Department of Biochemistry, McGill University, Montréal, QC, Canada.

[3] Department of Medicine, McGill University, Montréal, QC, Canada

[4] Department of Neurosciences, Montreal Neurological Institute-Hospital, McGill University, Montréal, QC, Canada

* Correspondence: thomas.duchaine@mcgill.ca

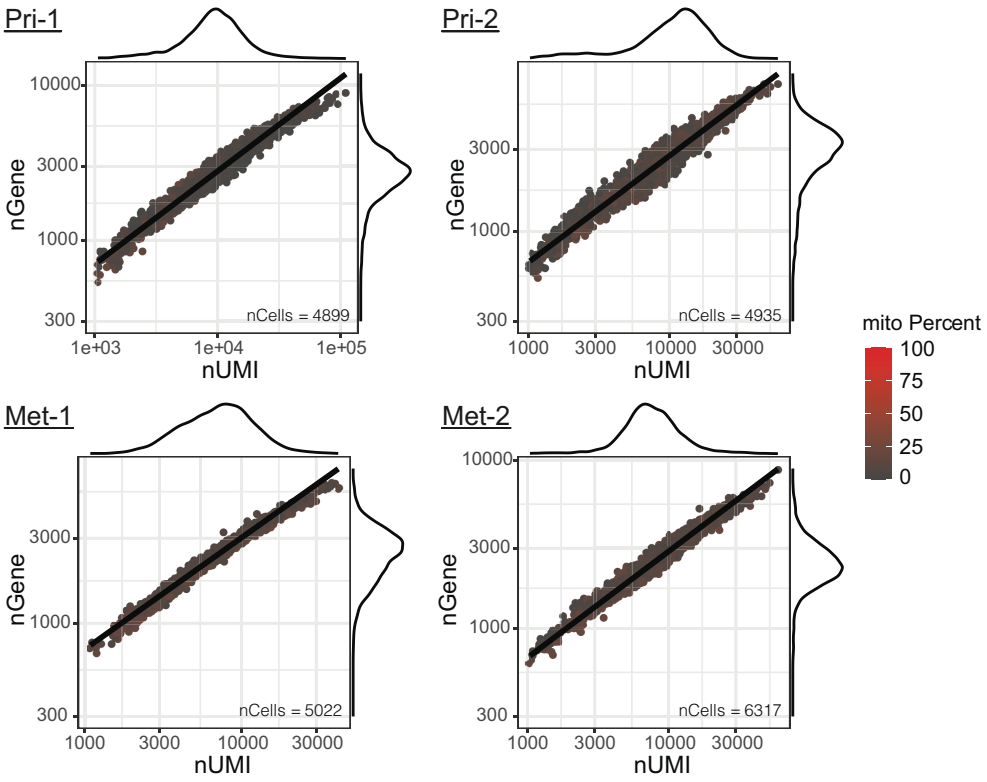


Figure A3.1: Cells passing quality control. Scatter plots for the number of genes (nGene) and the number of mRNA molecules (nUMI, unique molecular identifier) identified in each cell for each sample. Cells are colored by the percentage of genes mapped to mitochondrial genes. The number of cells (nCells) retained after quality control is indicated for each sample.

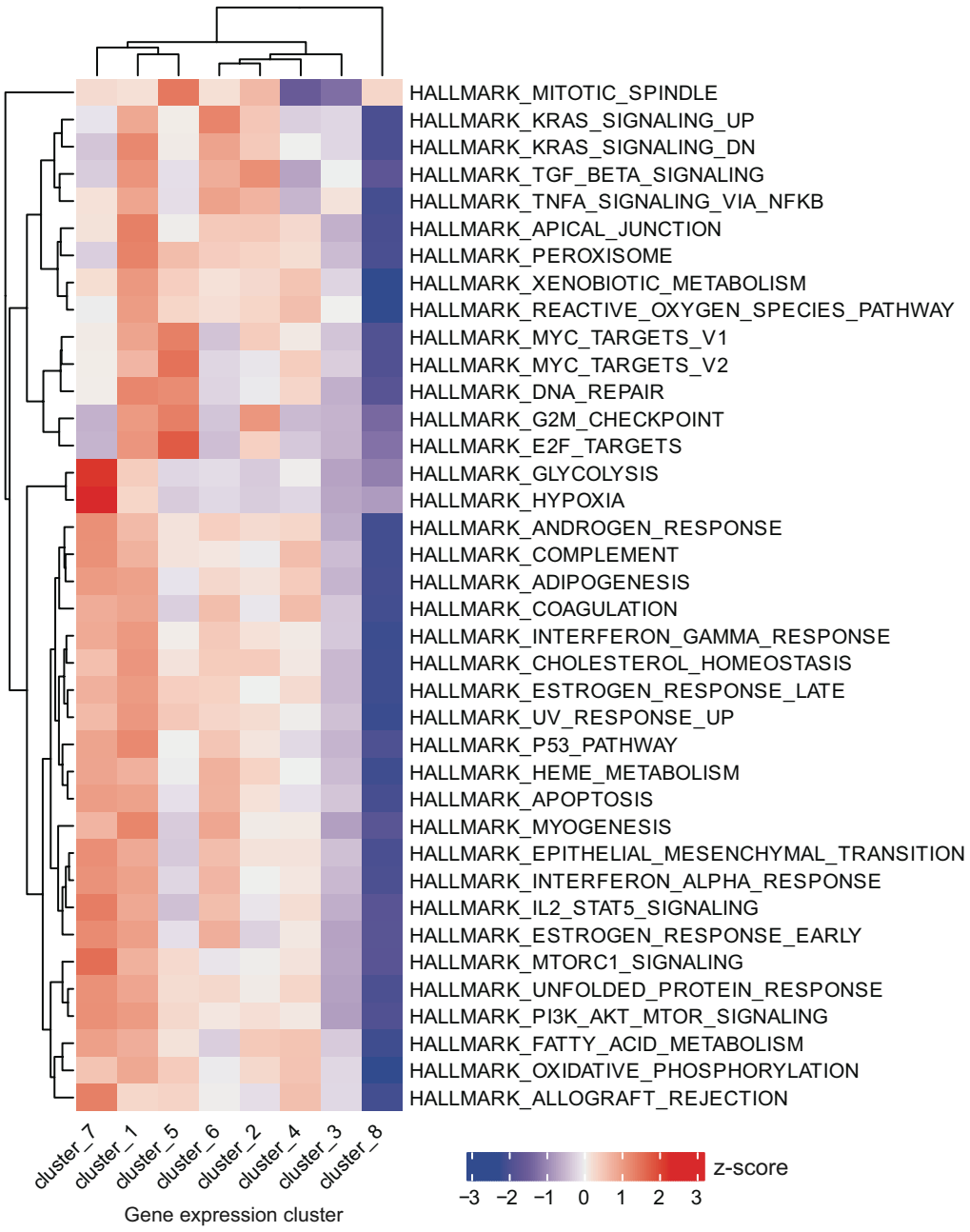


Figure A3.2: Gene set enrichment analysis for all gene expression clusters. The level of expression for genes in each of the hallmark gene set is plotted in a heatmap for every gene expression cluster (1-8).

A

GO:MF		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
electron transfer activity	GO:0009055	5.616×10^{-9}	
protein binding	GO:0005515	1.578×10^{-8}	
oxidoreductase activity	GO:0016491	2.035×10^{-7}	
oxidoreduction-driven active transmembrane transporter activity	GO:0015453	5.004×10^{-7}	
primary active transmembrane transporter activity	GO:0015399	6.798×10^{-7}	
NADH dehydrogenase (ubiquinone) activity	GO:0008137	7.429×10^{-5}	
NADH dehydrogenase (quinone) activity	GO:0050136	8.840×10^{-5}	
NADH dehydrogenase activity	GO:0003954	1.235×10^{-4}	
NAD(P)H dehydrogenase (quinone) activity	GO:0003955	1.450×10^{-4}	
oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	GO:0016655	7.636×10^{-4}	
active transmembrane transporter activity	GO:0022804	3.249×10^{-3}	
oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	GO:0016616	3.359×10^{-3}	
proton transmembrane transporter activity	GO:0015078	5.438×10^{-3}	
oxidoreductase activity, acting on CH-OH group of donors	GO:0016614	6.103×10^{-3}	
ATP-dependent protein folding chaperone	GO:0140662	8.880×10^{-3}	
protein folding chaperone	GO:0044183	1.459×10^{-2}	
oxidoreductase activity, acting on NAD(P)H	GO:0016651	1.636×10^{-2}	

B

KEGG		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
Oxidative phosphorylation	KEGG:00190	3.055×10^{-9}	
Huntington disease	KEGG:05016	1.257×10^{-8}	
Chemical carcinogenesis - reactive oxygen species	KEGG:05208	1.386×10^{-8}	
Prion disease	KEGG:05020	6.562×10^{-8}	
Parkinson disease	KEGG:05012	3.000×10^{-7}	
Alzheimer disease	KEGG:05010	8.341×10^{-7}	
Thermogenesis	KEGG:04714	1.514×10^{-6}	
Pathways of neurodegeneration - multiple diseases	KEGG:05022	1.975×10^{-6}	
Diabetic cardiomyopathy	KEGG:05415	1.025×10^{-5}	
Amyotrophic lateral sclerosis	KEGG:05014	4.223×10^{-5}	
Metabolic pathways	KEGG:01100	7.366×10^{-4}	
Non-alcoholic fatty liver disease	KEGG:04932	7.896×10^{-4}	
Carbon metabolism	KEGG:01200	1.712×10^{-2}	

C

REAC		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
The citric acid (TCA) cycle and respiratory electron tran...	REAC:R-HSA-1...	4.367×10^{-9}	
Respiratory electron transport, ATP synthesis by chemi...	REAC:R-HSA-1...	1.915×10^{-7}	
Respiratory electron transport	REAC:R-HSA-6...	1.336×10^{-5}	
Mitochondrial protein import	REAC:R-HSA-1...	1.898×10^{-3}	
Metabolism	REAC:R-HSA-1...	5.326×10^{-3}	
Complex I biogenesis	REAC:R-HSA-6...	7.791×10^{-3}	

Figure A3.3: Gene set enrichment analysis for mRNAs that are significantly shortened in the met-specific super-cluster compared to the pri-specific super-cluster. Different gene sets were obtained from Gene Ontology (GO) molecular functions (MF) (A), Kyoto Encyclopedia of Genes and Genomes (KEGG) (B), and Reactome (REAC) (C).

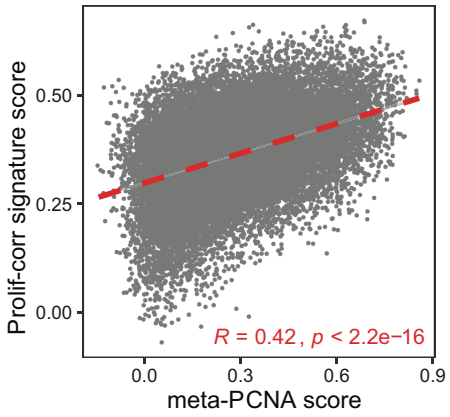


Figure A3.4: Proliferation-correlated signature scores significantly correlate with meta-PCNA scores. Each point represents a cell. Pearson's *R* and *p*-values are indicated.

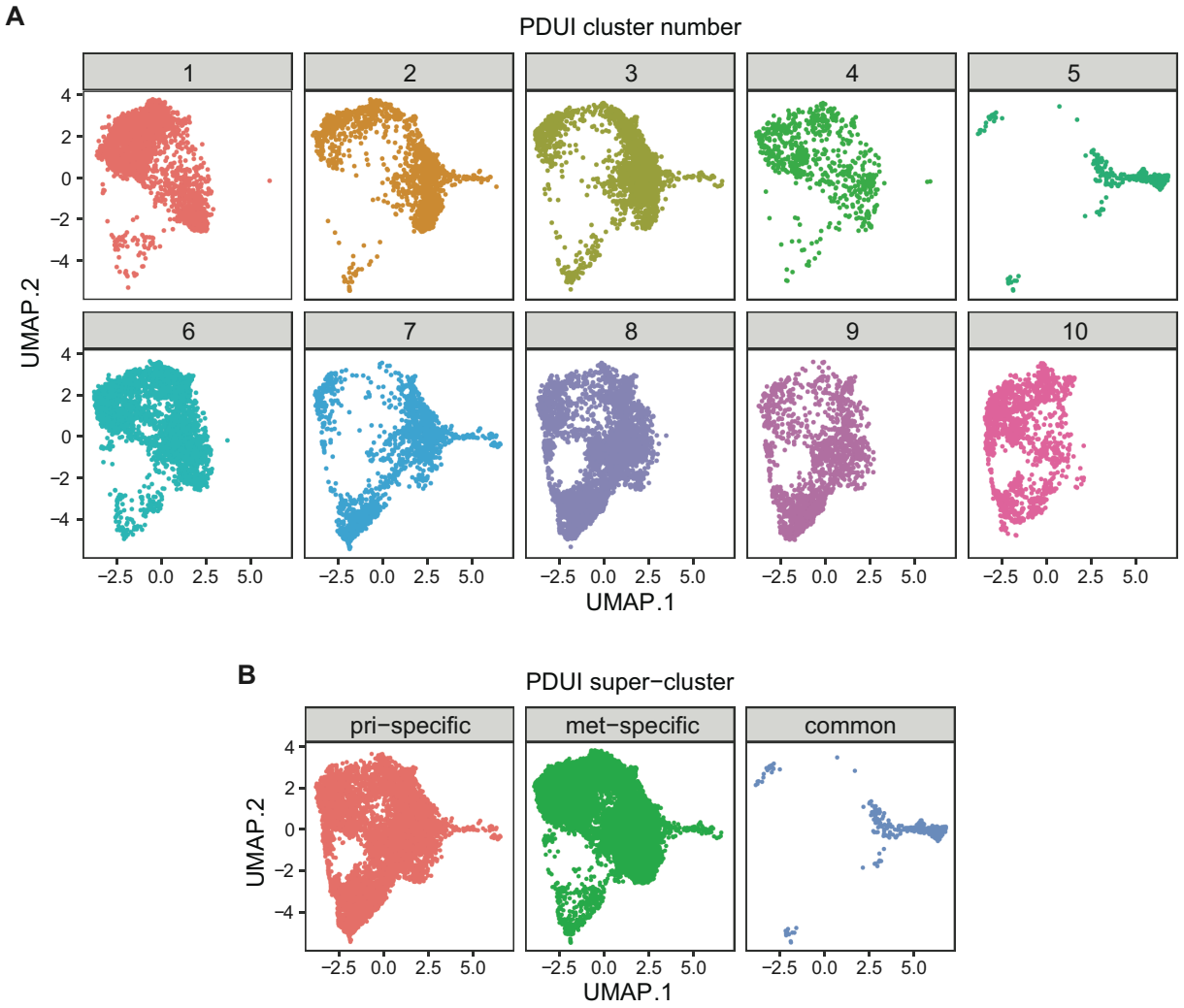


Figure A3.5: Overlaps between gene expression clusters and PDUI (super-)clusters. (A, B) UMAP plots of cells clustered by gene expression and colored according to their PDUI cluster number (as indicated between 1 and 10) (A), or colored according to their PDUI super-cluster membership (pri-specific, met-specific, or common) (B).

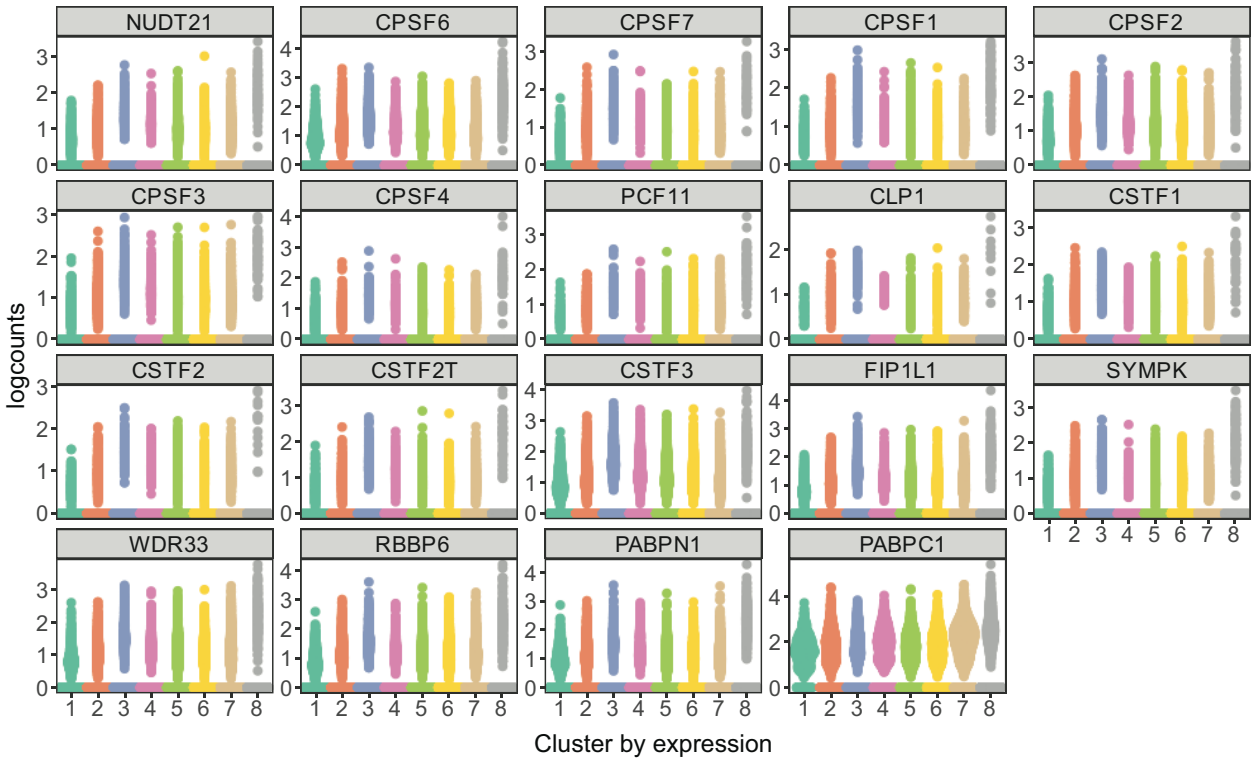


Figure A3.6: mRNA expression of cleavage and polyadenylation machinery components for cells in each gene expression cluster.

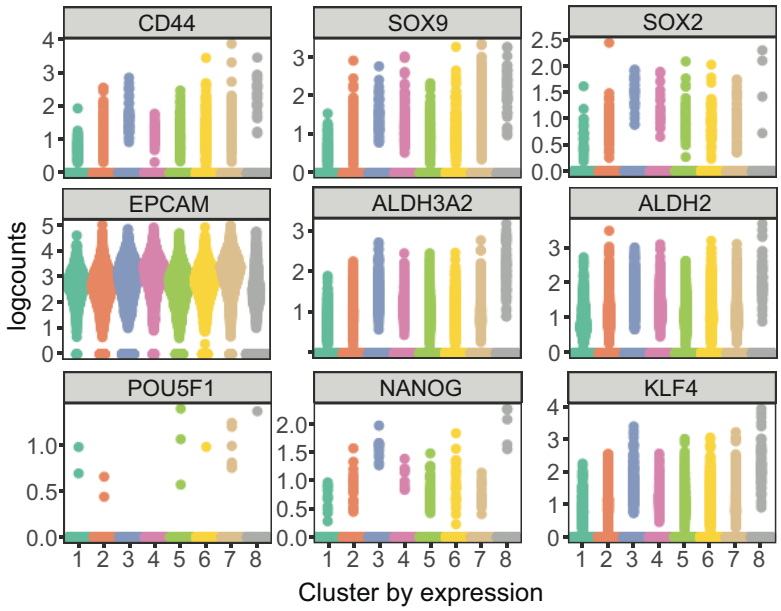


Figure A3.7: mRNA expression of stem cell markers for cells in each gene expression cluster.