

Investigating the Contribution of Alternative Polyadenylation to the Regulation of Gene Expression



Jonathan Neve

Pembroke College, University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity Term 2016

ABSTRACT

Investigating the Contribution of Alternative Polyadenylation to the Regulation of Gene Expression

Jonathan Neve

Pembroke College

Doctor of Philosophy in Medical Sciences

Trinity 2016

Around 70% of human genes have been found to contain multiple cleavage and polyadenylation (pA) sites. The alternative usage of these different pA sites is referred to as alternative cleavage and polyadenylation (APA). A wide range of studies have described switches in APA profiles when comparing different cell types and different cell states, notably general shifts to shorter APA isoforms upon the onset of cancer. These shifts have been described at both the individual gene and transcriptome levels. However, the extent to which these changes in APA profiles are contributing to changes in gene expression, and thereby altering the observed phenotype, is still under debate.

This thesis develops a novel subcellular fractionation approach to isolate those genes where the alternative use of different pA sites results in APA isoforms that differ in either the cytoplasmic stability or nuclear export rates. These differences would materialise as a differential representation of the resulting APA isoforms in the cytoplasmic and nuclear RNA fractions. This approach was applied to 12 human cell lines using a combination of publicly available generic poly-A⁺ RNA-sequencing data and 3' end sequencing performed here. Eleven out of twelve of these cell lines showed a significant bias towards an overrepresentation of the shorter, proximal APA isoform in the cytoplasm, which fits with the general model that 3' untranslated regions (3'UTRs) are mainly repressive and promote

cytoplasmic instability. Comparisons of the isolated genes for each cell line show that differential regulation of APA isoforms tends to be largely cell type specific.

Application of this subcellular fractionation approach to a cell-line based model for colorectal cancer progression highlights a small cohort of genes where APA is likely to have an impact on gene expression, potentially contributing to colorectal cancer progression. This analysis provides an ideal starting point to study these APA events at the individual gene level to elucidate fully whether APA is having a significant impact on gene expression.

Furthermore, post-transcriptional regulatory pathways governing these differential regulation events were investigated, including the contribution of micro-RNAs (miRNAs) to the differential regulation of APA isoforms. Unexpectedly, depletion of the endoribonuclease III, DICER1, reveals that it does not only shape the steady-state cytoplasmic APA profile through its well-known function in miRNA biogenesis, but also influences pA site choice in the nucleus.

ACKNOWLEDGEMENTS

I would firstly like to thank my supervisor, André Furger, who conceptualised this project and helped me every step of the way with his invaluable comments and unrivalled enthusiasm. I would also like to thank Nicholas Crump, who throughout both my Master's and Ph.D. studies provided precious support in and out of the lab. And thank you to the rest of the Furger Lab for creating such a pleasant and constructive workplace. I would also like to thank our collaborating labs of Monika Gullerova and Bin Tian for their significant inputs, without which, this project would have not got off the ground.

My thanks to my family, Joanna, Patrick, Shona and Hannah, who have all unwaveringly supported (and put up with) me for the past 26 years. I would also like to thank Meghan Campbell, Geoffrey Swenson, Michael King, Rens Krijgsman and Francesca Ward for making my return to Oxford not just about biochemistry. And most of all I would like to thank my wife-to-be, Alice, for supporting me every step of the way right from the decision to leave life as a sugar factory worker in Nottingham, back to being a student, and now on to London. It has been an interesting journey so far, but it is only the start...

LIST OF PUBLICATIONS

Neve, J., Burger, K., Li, W., Hogue, M., Patel, R., Tian, B., Gullerova, M., & Furger, A. (2016). Subcellular RNA profiling links splicing and nuclear DICER1 to alternative cleavage and polyadenylation. *Genome Research*, 26, 24–35. <http://doi.org/10.1101/gr.193995.115>

Neve, J., & Furger, A. (2014). Alternative polyadenylation: less than meets the eye? *Biochemical Society Transactions*, 42(4), 1190–5. <http://doi.org/10.1042/BST20140054>

STATEMENT OF AUTHORSHIP

The experiments below were used in the thesis and were performed by the following:

Experiment	Performed by
3'READS library preparation	Mainul Hoque Tian Lab Rutgers University
All chromatin immunoprecipitation and quantitative reverse transcriptase experiments in Chapter 5	Kaspar Burger Gullerova Lab Sir William Dunn School of Pathology University of Oxford

All data produced here were produced in collaborative projects and permissions to use these data in this thesis has been obtained.

GLOSSARY

3'RACE	3' rapid amplification of cDNA ends
3'READS	3'region extraction and deep sequencing
APA	Alternative polyadenylation
ARE	AU-rich element
aUTR	Alternative UTR
CARE	CA-rich element
cDNA	Complementary DNA
ChIP	Chromatin Immunoprecipitation
CPE	Cytoplasmic polyadenylation element
CPSF	Cleavage and polyadenylation specificity factor
CR-APA	Coding region APA
CstF	Cleavage stimulation factor
CURE	CU-rich element
cUTR	Common UTR
DaPars	Dynamic analysis of alternative polyadenylation by RNA-sequencing
DREME	Discriminative Regular Expression Motif Elicitation
DRS	Direct RNA sequencing
DSE	Downstream element
dsRNA	Double stranded RNA
ERCC	External RNA Controls Consortium
FDR	False discovery rate
FISH	Fluorescence in situ hybridisation
GRE	GU-rich element
GO	Gene Ontology
hnRNP	Heterogeneous Nuclear Ribonucleoprotein
lncRNA	Long non-coding RNA
MEFs	Mouse embryonic fibroblast cells
mESCs	Mouse embryonic stem cells
miRISC	miRNA-induced silencing complex
miRNA	Micro RNA
mRNA	Messenger RNA
OPMD	Oculopharyngeal muscular dystrophy
pA site	Cleavage and polyadenylation site
PABP	Poly(A) binding protein
PANTHER	Protein ANalysis THrough Evolutionary Relationships
PAS	Polyadenylation signal
PASS	PolyA site supporting
PDUI	Percentage distal usage index
PLATA	Probe level alternative transcript analysis
Pre-miRNA	Precursor hairpin miRNA
pre-mRNA	Precursor messenger RNA
Pre-rRNA	Precursor ribosomal RNA
Pri-miRNA	Primary precursor miRNA

PTB	Polypyrimidine tract binding protein
RIN	RNA integrity number
RNA	Ribonucleic acid
RNAa	RNA activation
RNAPII	RNA polymerase II
rpm	Reads per million
rRNA	Ribosomal RNA
RT-PCR	Reverse transcription PCR
RT-qPCR	Reverse transcription quantitative PCR
SAAP	Significance Analysis of Alternative Polyadenylation
shRNA	Short hairpin RNA
SI-NETS	Small intestinal neuroendocrine tumours
siRNAs	Small interfering RNAs
sUTR	Single UTR
UTR	Untranslated region
UTR-APA	Untranslated region APA
Δ PDUI	Change in proximal/distal usage index

LIST OF FIGURES

Figure 1: Types of polyadenylation	7
Figure 2: Sequencing profiles over a gene with differentially regulated APA isoforms.....	37
Figure 3: Schematic of 3'READS library preparation.....	38
Figure 4: Schematic of 3'READS bioinformatics analysis pipeline	53
Figure 5: QuantSeq 3'mRNA sequencing library preparation	55
Figure 6: Subcellular Fractionation Controls for HEK293 cells	60
Figure 7: Correlation plots of 3'READS data on HEK293 subcellular fractions.....	62
Figure 8: Classification of pA site positioning	64
Figure 9: Positioning of pA sites in HEK293 cells.....	65
Figure 10: Number of pA sites per gene.....	66
Figure 11: Frequency of polyadenylation signal usage	69
Figure 12: Correlation between read number and PAS usage	70
Figure 13: Frequency of polyadenylation signal usage in mouse.....	72
Figure 14: Positional dependency of polyadenylation signal frequencies.....	74
Figure 15: Distribution of No PAS pA sites	74
Figure 16: Differential representation of UTR-APA isoforms in subcellular HEK293 fractions	82
Figure 17: Differential representation of CR-APA isoforms in subcellular HEK293 fractions.....	84
Figure 18: Differentially represented APA isoforms controlled for sequencing depth.....	87
Figure 19: Confirmation of mRNA distribution as measured by 3'READS	89
Figure 20: Northern blot analysis of <i>CIAOI</i> APA isoforms.....	91
Figure 21: DaPars analysis of RNA-seq from subcellular fractions.....	94
Figure 22: Cell-type specificity of differentially regulated APA isoforms	97
Figure 23: Screenshot of ENCODE RNA-seq data of <i>HEMK1</i>	98
Figure 24: Gene ontology analysis of genes that show a differential representation of APA isoforms	99
Figure 25: Screenshot of RNA-seq data of the <i>HNRNPA2B1</i> gene.....	100
Figure 26: Motif analysis of aUTR regions of differentially regulated APA isoforms	103
Figure 27: Enrichment of miRNA target sites in the alternative UTRs of differentially regulated APA isoforms	105
Figure 28: Stability experiment for <i>CIAOI</i> APA isoforms	107
Figure 29: Overrepresentation of <i>CIAOI</i> distal APA isoform in the nucleus is coupled to upstream intron retention.....	108
Figure 30: Intron retention is seen at multiple genes the display differential subcellular APA representation ..	110
Figure 31: microRNA biogenesis pathway.....	122
Figure 32: Quality control experiments for DICER1-depletion	126
Figure 33: Effect of DICER1-depletion of subcellular UTR- APA profiles	128

Figure 34: Effect of DICER1-depletion of subcellular CR-APA profiles	129
Figure 35: APA events upon DICER1-depleted cells controlled for sequencing depth.....	130
Figure 36: Changes in gene expression of known effectors of pA site choice, upon DICER1 depletion	131
Figure 37: Confirmation of 3'READS results at the <i>ETNK1</i> locus upon <i>DICER1</i> -depletion	133
Figure 38: DICER1 localisation to the <i>ETNK1</i> locus	135
Figure 39: Dependence of DICER1 recruitment on dsRNA	137
Figure 40: Chromatin modifications across <i>ETNK1</i> locus	139
Figure 41: RNA Polymerase II occupancy across <i>ETNK1</i> locus.....	141
Figure 42: Model for DICER1-dependent termination at the proximal pA site of <i>ETNK1</i>	143
Figure 43: APA analysis of mouse RNA-seq data before and after Dicer1 ablation	145
Figure 44: Subcellular fractionation controls of 3 colorectal cell lines.....	154
Figure 45: Isolation of reads mapping to true pA sites	155
Figure 46: Reproducibility of 3'-end sequencing data from 1CT, SW480 and SW620 subcellular fractions	157
Figure 47: Heat map of genes which show variable expression between 1CT, SW480 and SW620 cell lines ..	159
Figure 48: APA analysis of the subcellular fractions of colorectal cancer cell lines	160
Figure 49: Screenshot of 3'end sequencing reads mapping to the <i>TOLLIP</i> locus.....	163
Figure 50: Genes that show both differential regulation of APA isoforms and a shift in nuclear APA profiles through colorectal cancer progression.....	165
Figure 51: Changes in gene expression levels upon shifts in nuclear APA profiles between SW480s and 1CTs	168
Figure 52: Changes in gene expression levels upon shifts in nuclear APA profiles between SW620s and SW480s	169
Figure 53: Gene Ontology analysis of Differentially Regulated Genes in Colorectal Cell lines	172
Figure 54: Examples of genes which display varying levels of differential regulation.....	175
Figure 55: Motif analysis and miRNA-target site enrichment analysis in the aUTRs of differentially regulated genes in colorectal cell line	177
Figure 56: CR-APA analysis of the subcellular fractions of colorectal cancer cell lines.....	179

LIST OF TABLES

Table 1 - <i>Cis</i> -elements in the 3'UTR	2
Table 2 - pA-factors known to influence pA efficiency	13
Table 3 – Other RNA-binding factors known to influence pA efficiency.....	14
Table 4 - Other factors that are known to influence pA site choice	14
Table 5 - Examples of genes producing differentially regulated UTR-APA isoforms.....	20
Table 6 - Methods of 3'end targeted deep sequencing.....	34
Table 7 - Chemicals and kits used in this study.....	42
Table 8 - Primary antibodies used in this study.....	44
Table 9 - Composition of buffers used in this study.....	45
Table 10 - Components of RT-PCR reaction.....	47
Table 11 - Cell lines used for cytoplasmic-nuclear comparison of APA profiles	92
Table 12 - Most significantly overexpressed genes in SW620 relative to 1CT cells	158

TABLE OF CONTENTS

Abstract.....	i
Acknowledgements.....	iii
List of Publications.....	iv
Statement of Authorship	v
Glossary	vi
List of Figures.....	viii
List of Tables	x
Table of Contents	xi
1. Chapter 1 – Introduction	1
1.1. RNA Polyadenylation.....	1
1.1.1. The 3'UTR and Post-Transcriptional Regulation.....	1
1.1.2. RNA stability.....	2
1.1.3. RNA translation.....	3
1.1.4. RNA localisation.....	4
1.2. Alternative 3'-End Processing.....	5
1.3. Large-scale Shifts of APA Profiles.....	8
1.3.1. APA in Cellular Proliferation and Activation.....	8
1.3.2. APA in Development.....	9
1.3.3. APA in Distinct Tissues	9
1.3.4. APA in Cancer.....	10
1.3.5. Polyadenylation in Disease	10
1.4. Factors Influencing Polyadenylation Site Choice.....	11
1.4.1. Overview	11
1.4.2. Cis-acting sequence elements	12
1.4.3. Trans-acting 3'end processing factors.....	13
1.4.4. Crosstalk between APA and transcription	15
1.4.5. Crosstalk between APA and splicing.....	16
1.4.6. Crosstalk between APA with chromatin	17
1.4.7. Overview	17
1.5. Impact of APA in a Cellular Context.....	18
1.5.1. Effects of UTR-APA on Transcript Stability	18
1.5.2. Effects of UTR-APA on Translational Efficiency.....	21

1.5.3.	Effects of UTR-APA on RNA localisation	22
1.5.4.	Coding Region APA in Impact on the Transcriptome and Proteome.....	23
1.6.	Conservation of APA	24
1.7.	Issues Surrounding the Global Analysis of APA	26
1.8.	Techniques to study Single Gene APA.....	29
1.9.	Techniques to study Global APA.....	30
1.9.1.	EST analysis from Sanger sequencing.....	30
1.9.2.	Microarrays.....	31
1.9.3.	RNA-seq.....	31
1.9.4.	3'-end targeted deep sequencing	32
1.10.	Strategy	34
1.10.1.	Aims of the project.....	34
1.10.2.	The subcellular fraction approach versus previous approaches	35
1.10.3.	Choosing pA mapping method	38
1.10.4.	Identifying mechanisms responsible differential APA isoform regulation.....	41
1.10.5.	Individual gene studies.....	41
2.	Chapter 2 – Materials and Methods	42
2.1.	Introduction	42
2.2.	Chemicals and Kits	42
2.3.	Antibodies	44
2.4.	Buffers and Compositions.....	45
2.5.	Cell Culture	46
2.6.	Subcellular Fractionation	46
2.7.	Reverse-Transcription PCR.....	47
2.8.	Denaturing Agarose RNA Gel Electrophoresis and Northern Blotting	48
2.9.	Chromatin Immunoprecipitation and quantitative RT-PCR.....	48
2.10.	Reverse-Transcription quantitative PCR (RT-qPCR)	49
2.11.	Western Blotting.....	50
2.12.	3'Region Extraction and Deep Sequencing (3'READS).....	50
2.13.	Bioinformatic Analysis of 3'READS Data.....	51
2.14.	Significance Analysis of Alternative Polyadenylation.....	53
2.15.	Quant-Seq APA Analysis	54
2.16.	Dynamic analysis of APA from RNA-seq (DaPars)	56
2.17.	Motif Analysis.....	56
2.18.	Gene Ontology Analysis	57
3.	Chapter 3 – Global Mapping of Polyadenylation Sites	58
3.1.	Introduction	58
3.2.	Results	58

3.2.1.	Establishing a clean subcellular fractionation protocol	58
3.2.2.	Reproducibility of 3'READS Results.....	61
3.2.3.	Classification of a polyadenylation site	62
3.2.4.	Location of pA sites throughout HEK293 cells.....	63
3.2.5.	Number of pA sites per gene	65
3.2.6.	Polyadenylation Signal Hexamer usage.....	67
3.2.7.	Positional effect on hexamer usage.....	73
3.3.	Discussion	75
3.3.1.	Advantages and disadvantages of 3'READS technique.....	75
3.3.2.	Potential reasons for having multiple pA sites per gene	76
3.3.3.	Distribution of hexamer usage	77
3.3.4.	Conclusions.....	79
4.	Chapter 4 – Differential Representation of APA Isoforms in Subcellular Fractions.....	80
4.1.	Introduction	80
4.2.	Results	81
4.2.1.	Differential Subcellular APA Isoform Representation	81
4.2.2.	Reproducibility of subcellular APA distribution.....	86
4.2.3.	Corroboration of 3'READS data using northern blotting and RT-PCR.....	87
4.2.4.	DaPars analysis to assess subcellular APA isoform distribution.....	92
4.2.5.	Isolated genes in which APA results in differentially regulated APA isoforms.....	95
4.2.6.	Conservation of Subcellular Distribution of APA Isoforms Across Different Cell Types	95
4.2.7.	Gene ontology enrichment of genes which show differential regulation of APA isoforms.....	98
4.2.8.	Motif discovery in the aUTRs of genes that show 3'UTR shortening in the cytoplasm	101
4.2.9.	miRNA target sites in aUTRs that show shortening.....	103
4.2.10.	Intron retention in CIAO1 APA isoforms.....	106
4.2.11.	Intron retention-coupled to specific APA isoforms.....	109
4.3.	Discussion	110
4.3.1.	Methods for the identification of differentially regulated APA isoforms.....	110
4.3.2.	The isolation of differentially regulated APA isoforms	111
4.3.3.	Differential regulation of APA isoforms via RNA-binding proteins	113
4.3.4.	Differential regulation of APA isoforms via miRNAs	114
4.3.5.	Differential regulation of APA isoforms via retained introns	116
4.3.6.	Coding region APA	118
4.3.7.	Conclusions	118
5.	Chapter 5 – The Impact of <i>DICER1</i> Knockdown on Subcellular APA Profiles	120
5.1.	Introduction	120
5.1.1.	Methods of compromising miRNA-targeted degradation	120

5.1.2.	DICER1 inside of the nucleus	121
5.1.3.	Overview of findings	123
5.2.	Results	124
5.2.1.	DICER1 Knockdown Using a Tet-Inducible System	124
5.2.2.	The Effect of DICER1 knockdown on Subcellular APA Profiles.....	127
5.2.3.	Changes in levels of polyA factors upon DICER1 knockdown	131
5.2.4.	DICER1 ChIP in the ETNK1 Locus	132
5.2.5.	Double-stranded RNA at the ETNK1 locus.....	136
5.2.6.	Chromatin marks associated with DICER1 recruitment.....	138
5.2.7.	Dependence of pA site selection in ETNK1 on RNAPII processivity.....	140
5.2.8.	Effect of Dicer1 ablation in mouse ES cells	144
5.3.	Discussion	146
5.3.1.	Differentially regulated APA isoforms mediated by miRNAs	146
5.3.2.	Nuclear function of DICER1	147
5.3.3.	Influence of DICER1 on pA site choice	148
5.3.4.	Conclusions and future directions	149
6.	Chapter 6 – Subcellular APA Isoform Analysis in Colorectal Cancer Cell Lines	
	150	
6.1.	Introduction	150
6.1.1.	Cancerous versus non-cancerous colorectal cell lines.....	151
6.1.2.	3'end sequencing approach	153
6.2.	Results	153
6.2.1.	Subcellular fractionation of 1CT, SW480 and SW620 cells.....	153
6.2.2.	1CTs, SW480s and SW620s as a model for colorectal cancer.....	156
6.2.3.	Comparison of Differential Subcellular APA isoform distributions	160
6.2.4.	Comparison of Nuclear APA profiles in 1CT, SW480 and SW620 cells	162
6.2.5.	Comparison of Cytoplasmic APA profiles in SW620 and 1CT cells	163
6.2.6.	Overlap of genes that show differential regulation and pA site shifts	164
6.2.7.	Gene ontology analysis of APA genes in colorectal cell lines	171
6.2.8.	Genes that are differentially regulated in SW480s and SW620s but not 1CTs	173
6.2.9.	Motif analysis of aUTRs in genes that show differential regulation of APA isoforms	176
6.2.10.	Coding region APA in colorectal cancer progression	178
6.3.	Discussion	181
6.3.1.	Introduction.....	181
6.3.2.	Use of 1CTs, SW480s and SW620s as a model for colorectal cancer	181
6.3.3.	General shifts in APA profiles through colorectal cancer progression	182
6.3.4.	Gene expression changes upon changes in pA site choice	184
6.3.5.	Motif analysis of genes that show differential regulation of APA isoforms	184

6.3.6.	APA events of interest isolated by the subcellular fraction approach	186
7.	Chapter 7 – Discussion	190
7.1.	Introduction	190
7.2.	The Prevalence of APA as a Method of Regulating Gene Expression	191
7.3.	The Subcellular Fractionation Approach	195
7.4.	Mechanisms involved in differential regulation of APA isoforms	197
7.5.	Obtaining Specificity in Regulating pA Site Choice.....	198
7.6.	Regulation of gene expression via post-transcriptional mechanisms .	201
7.7.	Conclusions and Future Directions	202
8.	Chapter 8 – Appendices.....	203
9.	References.....	207

1. CHAPTER 1 – INTRODUCTION

1.1. RNA POLYADENYLATION

Upon transcription by RNA polymerase II (RNAPII), protein-coding primary transcripts must undergo extensive processing to form stable messenger RNAs (mRNAs) that are capable of being exported from the nucleus and subsequently translated. In mammals, this processing includes the endonucleolytic cleavage of the precursor mRNA (pre-mRNA), and subsequent addition of usually 150-200 adenosines to the 3'-end of the upstream cleavage product (Proudfoot, 2011).

Despite the relative simplicity of this reaction, upwards of 16 proteins have been identified in the human cleavage and polyadenylation complex (Xiang et al., 2014). This complex binds to distinct cleavage and polyadenylation (pA) sites present on the pre-mRNA, which is defined by the surrounding sequence elements (Tian and Graber, 2012). The canonical pA site is an A(A/U)UAAA hexamer, known as the polyadenylation signal (PAS), ~30 nucleotides upstream of the cleavage site, and a GU-rich region downstream element (DSE). In protein-coding genes, these pA sites are primarily located downstream of a translational stop codon, and the region between this stop codon and the pA site constitutes the 3'untranslated region (3'UTR).

1.1.1. The 3'UTR and Post-Transcriptional Regulation

Traditionally, it was viewed that the regulation of gene expression was achieved solely through the control of transcription rates. However, as our knowledge of the RNA world has progressed it has come to light that regulation of mRNA stability, localisation and translational control play a significant role in the modulation of gene expression (Moore, 2005). This post-transcriptional control is largely directed by *cis*-acting elements located in

the 3'UTR. The bias towards *cis*-elements being located in the UTRs is a reflection of the restricted variation that the coding region can withstand before the protein function is compromised. Conversely, the UTRs have no such constraint, and essentially provides a platform that can freely evolve regulatory elements, and subsequently mediate post-transcriptional regulation (Lee and Gorospe, 2011).

Cis-elements found in the 3'UTR can regulate several aspects of RNA metabolism, including mRNA stability, nuclear/cytoplasmic export rates, subcellular localisation and translational efficiency, which collectively influence the final level of synthesised protein. These *cis*-elements primarily act through the binding of *trans*-factors, which are typically RNA-binding proteins, or RNAs themselves. These regulatory elements have the potential to trigger a number of different post-transcriptional regulatory pathways, some of which are summarised in Table 1.

Table 1 - *Cis*-elements in the 3'UTR

Class of <i>cis</i>-elements	Sequence element	Overview	References
AU-rich elements (AREs)	AUUUA	These are present in 5-8% of all genes and can trigger mRNA destabilisation and translational repression. This is triggered by the binding of ARE-binding proteins (ARE-BPs), including TTP.	(Bakheet et al., 2006; Beisang and Bohjanen, 2012)
GU-rich elements (GREs)	GUUUG	Contained in at least 5% of human mRNAs and triggers mRNA deadenylation and degradation. Acts through binding of proteins from the CELF family.	(Rattenbacher et al., 2010)
CU-rich elements (CUREs)	(C/U)CCAN _x CCC (U/A)Py _x UC(C/U)CC	PTB is the best-characterised CURE-binding protein and can affect translational repression, polyadenylation and mRNA stability.	(Sawicka et al., 2008)
CA-rich elements (CAREs)	(CA) _n	A stabilising dinucleotide repeat, which acts primarily via hnRNP L binding, which alters the susceptibility of the mRNA to endo- and exonucleases.	(Lee et al., 2009)
microRNA target sites	NNNNNNN	By far the most common destabilising element and target sites are present in > 60% of all genes. Regulation is primarily done via destabilisation of target mRNA (>84%), rather than translational inhibition.	(Friedman et al., 2009; Guo et al., 2010)

1.1.2. RNA stability

The most prevalent mechanism of regulating mRNA stability is through miRNA-mediated regulation, whereby > 60% of human genes contain miRNA-target sites (Friedman et al., 2009), with over 2500 individual mature miRNAs described on miRBase. In animals,

the targeting of mRNA by these ~22 nt mature miRNAs is determined by nucleotides 2 to 8 from the 5'-end of the miRNA, known as the “seed region” (Kim et al., 2009). These seed regions bind through complementary base pairing to the target sites on mRNA alongside the AGO proteins. Once bound, a miRNA-induced silencing complex (miRISC) is formed which triggers either translational inhibition or mRNA degradation (Bartel et al., 2004). mRNA degradation is primarily through mRNA deadenylation, mediated through binding of GW182 to the polyA-binding protein (PABP) where it recruits the deadenylases CCR4 and CAF1 (Eulalio et al., 2009).

The transcriptional regulation of pre-miRNAs is like that of other RNAPII transcribed genes, whereby complex feedback pathways act to modulate the exact miRNA expression levels in a spatiotemporal manner (Krol et al., 2010). As the target sites for miRNAs only involve short sequence signatures, it suits a combinatorial mechanism, by which a single mRNA can be targeted by a large number of distinct miRNAs. Conversely, miRNAs themselves can target sequences present in hundreds of different mRNAs. For example, disruption of miR-223 in mouse neutrophils has previously been shown to significantly affect the levels of hundreds of proteins (Baek et al., 2008). Over the last couple of decades, the true extent to which miRNAs affect protein output is becoming clearer. Despite the pervasiveness of regulation by miRNAs, along with the other post-transcriptional regulatory mechanisms, they are primarily thought of as “fine-tuning” mechanisms that act to complement the “on-off” switches that are characteristic of transcriptional control (P. Yao et al., 2012).

1.1.3. RNA translation

Governing the translational efficiency of RNA through its 3'UTR is an effective way of generating a rapid response to extracellular cues. This control is usually exerted through RNA-binding proteins, which may or may not bind their target RNA, dependent on their

stimulus controlled post-translational modifications. One of the most well-characterised examples of this is the interferon [IFN]- γ -activated inhibitor of translation (GAIT) complex. The binding of this complex to the stem-loop element located in its target 3'UTRs prevents the association of the small ribosomal subunit, thereby inhibiting the translation of that RNA. Stimulation by IFN- γ causes the necessary phosphorylation of components of the GAIT complex that are required to form the complex, which subsequently allows its binding to the target 3'UTRs, and ultimately the inhibition of the translation of those RNAs (Sampath et al., 2004).

Initially, miRNA-mRNA interactions in animals were also thought to significantly affect translational efficiencies (Yekta et al., 2004). However, a recent ribosomal profiling study in human cells showed that in miRNA-mediated regulation, >84% of the reduced protein production is through mRNA stability and that translational control has a relatively minor impact (Guo et al., 2010).

1.1.4. RNA localisation

The level of cytoplasmic mRNA is largely dependent on how much the nucleus pumps out. This, itself, is determined through a large number of complex steps, including transcriptional rates, pre-mRNA processing and the nuclear export of the mature mRNA, each of which can be regulated (reviewed in Maniatis and Reed, 2002). Controlling the nuclear export of mRNA was first suggested in a study where ~30% of polyA⁺ RNA was found to localise entirely to the nucleus (Herman et al., 1976). Subsequent RNA-fluorescence in-situ hybridisation (RNA-FISH) experiments with an oligo-dT probe showed that these nuclear polyA⁺ RNAs were locating to specific foci in the nucleus (Carter et al., 1991). The identity of these loci has been described in recent years and is now known to include Cajal bodies, nuclear speckles and paraspeckles (Mao et al., 2011).

In *Saccharomyces cerevisiae*, nuclear foci are shown to bind transcripts and retain them, subsequently releasing them slowly for cytoplasmic translation (Kallehauge et al., 2012). Interestingly, in this same study, they show that forced export of these nuclear-retained transcripts results in translational incapacity and is potentially toxic to the cell. Nuclear retention of translationally competent RNAs can therefore potentially act as a method to regulate the level of cytoplasmic levels of RNAs, thereby controlling the final level of gene expression (Kallehauge et al., 2012).

Mammalian examples of regulation of nuclear/cytoplasmic shuttling have also been seen. The ubiquitous RNA-binding protein HuR binds to mRNA via AU-rich elements (AREs), and the resulting complexes then undergo nuclear/cytoplasmic export (Fan and Steitz, 1998). Phosphorylation of specific sites on HuR in response to extracellular cues can subsequently coordinate the mRNA binding and nuclear/cytoplasmic shuttling of HuR and the ARE-containing bound mRNAs (Doller et al., 2010).

RNA localisation extends beyond nuclear/cytoplasmic distribution. Localisation of RNA to specific sites for translation is an effective way to react rapidly to extracellular stimuli, as it can avoid the requirement to localise the resulting protein. This concept is particularly important in highly polarised cells, such as neurons (Andreassi and Riccio, 2009). These localisation elements can range from as little as 11 nucleotides to several kilobases (Jambhekar and Derisi, 2007) and act through the binding of *trans*-factors.

1.2. ALTERNATIVE 3'-END PROCESSING

Alternative polyadenylation (APA) occurs when a single gene has multiple pA sites, compared to constitutive polyadenylation, whereby one gene has one sole pA site. Like alternative splicing, APA results in multiple distinct RNA transcripts being produced from a single gene. Global analysis has found that ~70% of all human genes contain multiple pA

sites (Derti et al., 2012), and the relative positions of these pA sites in a gene will determine the sequence elements contained in the different transcript isoforms.

Untranslated region APA (UTR-APA) occurs when multiple pA sites occur within the same 3'UTR of the gene. This, therefore, does not alter the coding capacity of the transcript, but solely acts to remove a portion of the 3'UTR in the shorter transcript isoform. This removed portion of the 3'UTR will be referred to as the alternative UTR (aUTR), relative to the common UTR (cUTR), which is present in all APA isoforms.

Coding region APA (CR-APA) occurs when multiple pA sites occur at positions in which the coding region, and thus sequence of the final protein product, is altered. CR-APA can further be categorised into 3 separate classes: exonic CR-APA, whereby a pA site is located within the coding region itself, and therefore lacks a translational stop codon; composite-intronic CR-APA, where the pA site is located within an intron and can arise when a 5'-splice site becomes silent; skipped-intronic CR-APA, where alternative splicing results in the inclusion of a separate terminal exon containing a pA site (Edwalds-Gilbert et al., 1997). One final classification of pA site are the upstream antisense pA sites which arise during the termination of short antisense transcripts which derive from bidirectional promoters (Almada et al., 2013). These classifications of APA are summarised in Figure 1.

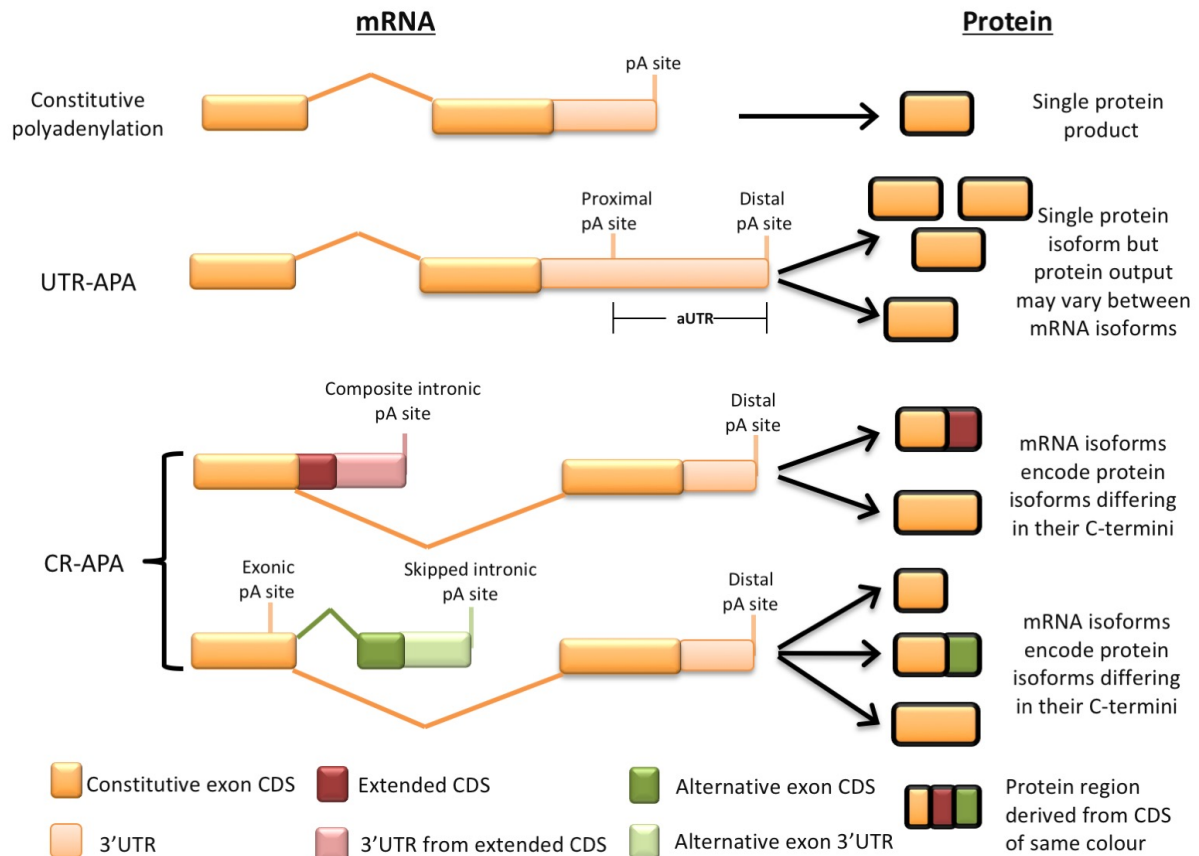


Figure 1: Types of polyadenylation

Top: Constitutive polyadenylation, which occurs when a gene contains a single cleavage and polyadenylation (pA) site.

Middle: Untranslated region alternative cleavage and polyadenylation (UTR-APA), which occurs when multiple pA sites occur within the same 3'untranslated region (3'UTR) resulting in several APA mRNA isoforms. Since these mRNA isoforms only differ in their 3'UTR length, they encode the same protein but may have altered mRNA stability or translational efficiency, potentially resulting in a different level of protein output. The region located between the proximal and distal pA sites is known as the alternative UTR (aUTR).

Bottom: Coding-region APA (CR-APA), which occurs when a pA site is located upstream of the terminal 3'UTR. This can occur upon intron retention, in the case of composite-intronic CR-APA, or within an alternatively spliced exon in the case of skipped-intronic CR-APA. Exonic CR-APA occurs when a pA site is located within the coding sequence (CDS) of a gene. CR-APA results in either C-terminus truncation of the resulting protein in the case of exonic APA, or a different C-terminus, in both composite and skipped-intronic CR-APA. Adapted from Neve and Furger, 2014.

1.3. LARGE-SCALE SHIFTS OF APA PROFILES

Advances in next generation sequencing have allowed cellular APA profiles to be analysed on a transcriptome scale. This has resulted in a huge number of studies looking at general shifts in APA profiles upon a change in cellular state (e.g. cellular activation) or a change in cell type (e.g. different tissues). These are generally described as shifts between upstream promoter-proximal sites and downstream promoter-distal sites, which will be referred to as proximal and distal pA sites, respectively, throughout this work. An overview of these studies and their findings will be given in this section.

1.3.1. *APA in Cellular Proliferation and Activation*

A general pattern of shifts towards the proximal pA sites has been seen upon increased cellular proliferation (Elkon et al., 2012; Sandberg et al., 2008). In the case of murine T-lymphocyte activation, the UTR-APA events show a much more obvious pattern of a switch to proximal pA sites, compared to CR-APA, where there appeared to be equal movements to both proximal and distal pA sites (Sandberg et al., 2008). Comparison with an extensive range of both murine and human cell lines showed a similar pattern whereby proliferation rate was inversely related to 3'UTR length (Sandberg et al., 2008).

Furthermore, activation of neurons by depolarising agents results in shifts to proximal pA sites in a specific subset of genes (Flavell et al., 2008). Many of these cases are CR-APA, which results in truncated proteins that may have alternative functions to their full-length counterparts (Flavell et al., 2008). In this same study, analysis of microarray data from various cells stimulated with other extracellular factors, such as epidermal growth factor, also showed pA site switching upon stimulation (Flavell et al., 2008).

1.3.2. *APA in Development*

APA profiles have also been found to vary as a function of differentiation status. This was first described in mouse embryonic development, where the differentiation of C2C12 myoblast cells to myotubes showed a general shift to distal pA sites (Ji et al., 2009). Conversely, if this is reversed, and induced pluripotent stem cells are generated from fully differentiated somatic cells, a general 3'UTR shortening is observed (Ji and Tian, 2009). However, more recent studies looking at spermatogenesis in mouse find general shifts to proximal pA sites as spermatogenesis progresses (Li et al., 2016) demonstrating that the role of APA in development or differentiation cannot be considered under a single umbrella, and each process has to be considered individually.

1.3.3. *APA in Distinct Tissues*

Characteristic APA profiles for specific tissues have also been observed. The most striking examples of this are in the mammalian and invertebrate testes, where proximal pA sites are favoured, and the brain, where distal pA sites are favoured (Miura et al., 2013; Smibert et al., 2012; Wang et al., 2008). Individual examples of neural-specific 3'UTR lengthening have been documented for many years using northern blot analysis (Wang et al., 1993). This pattern for preference of distal pA sites in the brain has been shown to be a widespread phenomenon, firstly by analysis of complementary DNA (cDNA) data (Zhang et al., 2005), then by RNA-seq analysis (Miura et al., 2013) and also specific 3'-end sequencing (Lianoglou et al., 2013).

In addition to simple switches between common pA sites, there are several examples of unique pA sites in testis (Liu et al., 2007). These sites appear to have a lower incidence of the AAUAAA canonical hexamer, as well as having unique upstream and downstream elements (Liu et al., 2007). Since these sites are not used efficiently in somatic cells, it has

been suggested that testis-specific pA factors may be responsible for this pA site switching (MacDonald and McMahon, 2010), such as the CstF-64 variant, τ CstF-64 (Dass et al., 2002).

1.3.4. *APA in Cancer*

Several studies have observed shifts to proximal pA sites upon cellular transformation and carcinogenesis (Fu et al., 2011; Lin et al., 2012; Masamha et al., 2014; Mayr and Bartel, 2009; Morris et al., 2012; Sandberg et al., 2008; Singh et al., 2009). It is thought that shifts to these proximal pA sites in cancer functions to exclude instability elements present in the aUTR, thus resulting in increased mRNA and protein levels of specific oncogenes, such as *CCND1* (Mayr and Bartel, 2009).

In the cancer cell lines where general 3'UTR shortening is observed, there are still a significant number of genes in which 3'UTR lengthening is seen (Fu et al., 2011; Lin et al., 2012; Singh et al., 2009), showing that the observed APA patterns in transformed cells are highly complex. In addition to potentially contributing to gene regulation, changes in APA profiles may also prove a valuable indicator of both cancer subtype (Singh et al., 2009) and prognosis (Lembo et al., 2012). Indeed, APA profiles have previously been proposed as biomarkers for cancer progression (Morris et al., 2012).

1.3.5. *Polyadenylation in Disease*

Mutations in the *cis*-elements of pA sites have been shown to cause a number of different human diseases. These include both mutations in the PAS, such as in thalassemia and IPEX syndrome (Bennett et al., 2001; Orkin et al., 1985; Whitelaw and Proudfoot, 1986) and mutations to other *cis*-elements surrounding the pA site (Yasuda et al., 2003). In these genes, pathogenesis generally results from inefficient 3'end processing at these genes, ultimately causing reduced protein levels. Conversely, mutations that result in the

introduction of new pA sites can also contribute to pathogenesis. Point mutations in the *CCND1* gene that introduce pA sites causing 3'UTR truncation, ultimately removing destabilising elements. This has been shown to lead to the overexpression of CCND1 in mantle cell lymphomas, resulting in increased carcinogenesis (Wiestner et al., 2007).

Pathogenesis as a result of mutations in proteins implicated in regulating APA has been described in several cases including myotonic dystrophy (Batra et al., 2014) and oculopharyngeal muscular dystrophy (OPMD) (de Klerk and Venema, 2012; Jenal et al., 2012). Triplet repeat expansion of an alanine repeat in PABPN1 leads to OPMD (Brais et al., 1998). As discussed below, PABPN1 is a known regulator of APA and its depletion, or in the case of OPMD, loss of function, leads to a general shift to proximal pA sites in a subset of genes (Jenal et al., 2012).

Similar to OPMD, myotonic dystrophy is also a result of triplet expansion which leads to the loss of function of the MBNL protein and results in large scale bidirectional APA shifts (Batra et al., 2014). However, the exact pathogenesis of both OMPD and myotonic dystrophy are unknown, and whether it is these shifts in APA profiles are causing the disease phenotype, or simply a side-effect to the change in cellular state, remains unknown.

1.4. FACTORS INFLUENCING POLYADENYLATION SITE CHOICE

1.4.1. Overview

Despite this large number of observed shifts between pA sites, the precise mechanism of defining pA site choice and the regulatory mechanisms that are responsible for switching between them remains ill defined. This is likely because it is a balance of a huge number of factors, including the inherent strength of the competing pA sites, defined by the surrounding *cis*-elements, as well as the availability of the plethora of *trans*-factors that may either promote or inhibit cleavage and polyadenylation at the different sites.

1.4.2. *Cis-acting sequence elements*

The efficiency of polyadenylation at any given site is governed by three main factors:

- 1) The intrinsic strength of the pA site, governed by its position in the gene, and the surrounding *cis*-elements.
- 2) The concentration of *trans*-factors that can influence the efficiency of the cleavage and polyadenylation reaction.
- 3) The chromatin environment throughout the gene, which covers both accessibility to *trans*-factors and RNAPII speed.

When multiple pA sites exist within a gene, these factors combine to determine the relative usage of each site.

The key *cis*-elements consist of a bipartite sequence. Between 0-40 nt upstream of the site of cleavage site lies the PAS, which typically consists of the canonical hexamer (A(A/U)UAAA), which binds CPSF160 (Murthy and Manley, 1995). Directly downstream of the pA site lies less well-defined U-rich and/or GU-rich DSE, which recruits the cleavage stimulating factor (CstF) through its 64-kDa subunit (Pérez Cañadillas and Varani, 2003).

The presence of this upstream canonical hexamer and DSE is not essential, and the degeneration of these elements can be compensated for by other *cis*-elements. *In vitro* studies looking at the effect of point mutations of the canonical hexamer have shown that single nucleotide alterations still allow cleavage and polyadenylation, albeit generally at a lower efficiency (Sheets et al., 1990). When looking at the frequency of these variant PASs in global pA site mapping studies, they are found at around 20-25% of all pA sites in humans (Wang et al., 2013). In addition to hexamer identity, the position of the DSE relative to the cleavage site can also influence efficiency (Gil and Proudfoot, 1987).

The concept of compensation of PAS degeneration has been previously demonstrated where the presence of a strong DSE allows efficient cleavage and polyadenylation with solely an A-rich upstream region and no defined PAS (Nunes et al., 2010). Further upstream, compensatory U-rich sequences can bind additional factors including hFip1 and CPSF30 (Barabino et al., 1997; Kaufmann et al., 2004). Distal G-rich boxes, both upstream and

downstream also have the potential to enhance cleavage and polyadenylation, mediated through the binding of hnRNP H, H', and F (Arhin et al., 2002; Dalziel et al., 2007; Millevoi et al., 2009; Veraldi and Arhin, 2001).

1.4.3. *Trans-acting 3'end processing factors*

As the *cis*-elements influencing polyadenylation efficiency do not change, the switching between pA sites has to be done through the fluctuation of *trans*-factors. Several different classes of *trans*-factors can influence pA site choice, including pA factors themselves (Table 2), splicing factors and other RNA-binding proteins (Table 3). Other factors that influence pA site choice that do not fall into any of these categories are also shown in Table 4. Transplant of large chromosomal segments from *Debaryomyces hansenii* into *Saccharomyces cerevisiae* result in the APA profile adopting that of the host demonstrating the effect that the *trans*-factors can have on the establishment of the APA profile (Moqtaderi et al., 2013).

Table 2 - pA-factors known to influence pA efficiency

Factor	Motif bound	Proposed model	References
CF _{im}	(UGUA) _n	At high levels, CF _{im} interacts with suboptimal CF _{im} binding sites preventing the interaction of CPSF with these proximal pA sites and promotes usage of distal pA sites. Depletion of CF _{im} allows the interaction of CPSF with proximal pA sites, resulting in 3'UTR shortening.	(Kubo et al., 2006; Martin et al., 2012; Masamha et al., 2014)
CstF64	U-rich	Co-depletion of CstF64 and CstF64 τ leads to APA shifts in a small number of genes primarily to the distal pA site, which is thought to be reflective of the general higher efficiency of distal pA sites. Furthermore, CstF64 has been found to promote usage of weaker pA sites containing the downstream GUKKU motif.	(Hwang et al., 2016; Shell et al., 2005; Takagaki et al., 1996; C. Yao et al., 2012)
CstF77		High levels of CstF77 result in activation of the pA site in intron 3 of CstF77 gene resulting in a negative feedback loop. Additionally, it results in changes in APA profiles of cell cycle genes, specifically where U-rich regions surround the pA sites.	(Luo et al., 2013)
hFip1		A component of CPSF complex. Regulation of APA by Fip1 is dependent on the distance between pA sites. When far apart, low-levels of Fip1 result in reduced pA efficiency and decreased use of weaker, proximal pA sites. When close together, Fip1 blocks CstF binding at the proximal site and therefore results in distal pA site usage.	(Lackford et al., 2014; Li et al., 2015)
Pcf11		Enhances use of proximal pA sites through direct binding to pre-mRNA and is a component of the pA complex.	(Li et al., 2015; West and Proudfoot, 2008)
RBBP6		Competes with iso3 for binding with the core pA machinery and when RBBP6 is bound it enhances pA efficiency and promotes the use of weaker proximal pA sites	(Di Giammartino et al., 2014)
β CstF-64		A neuronal splice variant of CstF64 that associates with the CstF complex and stimulates pA thereby activating weaker pA sites.	(Shankarling and Macdonald, 2013)

Table 3 – Other RNA-binding factors known to influence pA efficiency

Factor	Motif bound	Proposed model	References
Cirbp and Rbm3	GNNGNNG	Upon cold-shock, these factors are upregulated and, through 3'UTR binding, inhibit the use of proximal pA sites	(Liu et al., 2013)
CPEB1	CPE	CPEB1 shuttles to the nucleus binding cytoplasmic polyadenylation elements (CPEs) and enhances pA at nearby pA sites, as well as preventing U2AF65 binding, thereby inhibiting splicing.	(Bava et al., 2013)
ELAV (<i>Drosophila</i>)		In the neuronal tissues, ELAV is recruited to promoter-paused RNAPII complex. Upon resuming transcription, ELAV is deposited near proximal pA sites, inhibiting pA, thereby resulting in extended 3'UTRs.	(Oktaba et al., 2014; Slobodin and Agami, 2015)
FUS	UGGUU	FUS binds to RNA directly, and when bound downstream of a pA site it enhances pA efficiency via CPSF160 recruitment. Conversely, when a binding site is located upstream of a pA site, it results in premature RNAPII termination and downregulation of that pA site.	(Masuda et al., 2015)
hnRNP F	G-rich DSE	Competes with CstF-64 by binding to G-rich motifs near pA sites.	(Alkan et al., 2006)
hnRNP H1	Auxiliary DSE	Depletion results in a general shift to distal pA sites, with hnRNP H1 binding sites surrounding proximal sites.	(Katz et al., 2010)
hnRNP H2	G-rich	Binds near pA sites and enhances binding of CstF-64	(Bagga et al., 1998)
hnRNP K	UCCCUU	Competes with CFI for binding the pre-mRNA, reducing pA efficiency and a lower level of usage of that pA site.	(Naganuma et al., 2012)
hnRNP L	CA-rich elements	Functions as a splicing regulator, so altering levels of hnRNPL can sway the balance between competing splicing and intronic pA events	(Hung et al., 2008)
HuR (<i>ELAVL1</i>)	AREs	Through competition for binding sites on pre-mRNAs HuR can influence both APA and splicing, including its autoregulation.	(Barnhart et al., 2013; Dai et al., 2011)
Mbnl proteins	R/YGCY	Mbnl can either inhibit pA site usage if it binds close to pA site or enhance pA site usage if it binds further upstream. Inhibition is thought to occur through steric hindrance. Dysregulation of this process is involved in myotonic dystrophy.	(Batra et al., 2014; Taliaferro et al., 2016)
Nova	YCAY	Neural-specific factor that binds YCAY elements in the 3'UTR and, depending on the location of these motifs can influence pA site choice.	(Licatalosi et al., 2008)
PABPC1		No obvious <i>cis</i> -element enriched around pA sites regulated by PABPC1, but, like PABPN1, promotes distal pA site usage.	(Li et al., 2015)
PABPN1		Promotes the use of distal pA sites by inhibiting pA at weaker proximal pA site through competition with CPSF for binding to the PAS.	(Beaulieu et al., 2012; de Klerk and Venema, 2012; Jenal et al., 2012; Simonelig, 2012)
PCBP	C-rich	PCBP binds C-rich motifs in proximity to pA sites and enhances their usage by acting as an upstream sequence element enhancer	(Ji et al., 2013)
PTB	G-rich USE	Competes with CstF64 to bind the DSE, thereby inhibiting pA. However, it can also aid recruitment of HNRNP H1, which stimulates pA.	(Castelo-Branco et al., 2004)
SRm160	Unknown	Enhances pA through the association with CPSF	(McCracken et al., 2002)
U1 snRNP	AGguragu	Suppresses cryptic pA sites in the gene body, which is essential for the formation of full-length transcripts.	(Berg et al., 2012; Kaida et al., 2010; Lou and Neugebauer, 1998)
U1A	AUGCN(1-3)C	U1A binds GU-rich regions downstream of pA sites inhibiting the binding of CstF64, thus inhibiting pA. U1A can also bind to the polyA polymerase inhibiting the polyadenylation reaction itself.	(Gunderson et al., 1994; Phillips et al., 2004)
U2	3'SS	U2 interacts with CPSF and enhances pA efficiency.	(Kyburz et al., 2006)
U2AF	Pyrimidine tract	U2AF interacts with CFI stimulating pA.	(Millevoi et al., 2006)

Table 4 - Other factors that are known to influence pA site choice

Factor	Motif bound	Proposed model	References
CpG islands		The methylation status of CpG islands influences pA site selection in the murine imprinted gene <i>H13</i> . This is thought to act via an unidentified DNA methylation sensitive pA factor.	(Wood et al., 2008)
E2F	TTGGCGG	Through enhanced proliferation, increased levels of the transcription factor E2F result in the increased use of proximal pA sites.	(Elkon et al., 2012)
Nucleosome positioning		High nucleosome occupancy directly upstream of proximal pA sites generally correlates with increased proximal pA usage.	(Huang et al., 2013; Spies et al., 2009)
Transcription rate		Slow transcription rates result in a longer time between when the proximal and distal pA sites are transcribed thereby causing in an increased probability of proximal pA site utilisation.	(Pinto et al., 2011)

As can be seen from Tables 2, 3 and 4, there is a plethora of *trans*-factors that, when depleted or compromised, cause shifts in APA profiles. However, a key point here is, even though these factors have been found to **influence** pA site choice, it does not mean that all of these *trans*-factors are actively **regulating** the APA profile. This is a point that will be returned to later in the thesis

1.4.4. Crosstalk between APA and transcription

It has been long known that pre-mRNA processing and transcription are tightly coupled (Proudfoot et al., 2002). Cleavage and polyadenylation are essential for transcription termination (West and Proudfoot, 2009), as well as the recycling of transcriptional factors and the reinitiation at upstream promoters (Mapendano et al., 2010). The phosphorylated C-terminal domain of RNAPII also provides an essential platform for the recruitment of several pA factors to the pre-mRNA (Proudfoot et al., 2002). Specifically, CPSF and CstF, together with PcfII, are brought to the pre-mRNA through the interaction with RNAPII (Hsin and Manley, 2012). The pausing of RNAPII during transcription, and the subsequent recruitment of CstF to the transcription elongation complex, are decisive in pA site choice (Fusby et al., 2015). Even after the RNAPII has passed the pA site, the RNA tether between the two is necessary for the assembly of the mature pA complex, and thus efficient cleavage and polyadenylation (Rigo et al., 2005).

The influence of transcription on cleavage and polyadenylation is not limited to the 3'-end of the gene. The presence or absence of specific transcriptional activators can stimulate transcription-coupled pA in yeast through PAF1c, which is thought to facilitate the formation of the pA factors in the elongation complex (Nagaike et al., 2011). Furthermore, the promoters of specific genes in *Drosophila*, have been shown to be essential in the

recruitment of ELAV to RNAPII, which subsequently acts at the 3'ends of those genes to inhibit proximal pA site, thus resulting in 3'UTR extension (Oktaba et al., 2014).

1.4.5. *Crosstalk between APA and splicing*

The splicing of the terminal intron of a pre-mRNA is tightly coupled with pA in the 3'-exon, and both of these events are strongly enhanced by one another (Niwa et al., 1992). This coupling involves extensive interactions between the splicing machinery and the pA machinery (as reviewed in Martinson, 2011).

In contrast to pA sites located in the 3'UTR, 5'splice sites upstream of pA sites located in intronic regions can actively suppress cleavage and polyadenylation at these pA sites (Tian et al., 2007). This is because the upstream 5'splice site effectively competes with the pA site. This can result in 3 separate outcomes: 1) splicing proceeds unhindered and the intronic pA site is excised out; 2) the pA site outcompetes the upstream 5'SS, and pairs with the upstream 3'SS, thereby defining the terminal exon, resulting in a composite-intronic pA event (Figure 1); 3) the pA site pairs with a separate, cryptic 3'SS from within the intron resulting in a new terminal exon and a skipped-intronic pA event (Figure 1).

The outcome of this competitive interplay between splicing and pA is a balance between a huge number of factors including the relative strength of the 5'SS and pA site (Peterson and Perry, 1989), the distances between the sites (Peterson et al., 1994) and the levels *trans*-factors, including both splice factors and pA factors (Li et al., 2015; Takagaki et al., 1996). As a result, a large number of splice factors have been shown to affect pA site choice (see Table 3). Most notably, U1 snRNP actively inhibits pA usage in introns, thereby protecting the cell from premature pA events (Kaida et al., 2010). Physiologically, U1 levels are known to be very high which may be to ensure levels never fall below the threshold that would result in the activation of these cryptic pA sites (Elkon et al., 2013).

1.4.6. Crosstalk between APA with chromatin

Analogous to alternative splicing, APA is also governed by the kinetic aspect of transcription, including RNAPII transcriptional rates, chromatin structure, and histone modifications. Mutant *Drosophila* strains that have reduced RNAPII transcription elongation rates have shown that this can result in the preferential use of proximal pA sites in several alternatively polyadenylated genes (Pinto et al., 2011). The proposed mechanism for this is similar to that seen for exon skipping, in that a faster RNAPII reduces the latency period between when the proximal and distal pA sites are transcribed, thereby reducing the probability of utilising the, generally weaker, proximal pA site.

Furthermore, the conformation of chromatin has also been linked to APA. Studies looking at nucleosome positioning surrounding pA sites show a general depletion of nucleosomes immediately downstream of high-usage pA sites, and then an enrichment past ~100bp downstream of the pA site, when compared to low-usage pA sites (Huang et al., 2013; Spies et al., 2009).

1.4.7. Overview

In summary, there are a huge number of factors that come together to create the observed APA profile. These variable factors include 1) the core pA machinery that is required for processing of all pre-mRNAs; 2) auxiliary *trans*-factors which alter the pA efficiency at a subset of pA sites; 3) transcription rates, chromatin modifications and splicing events which can all alter the kinetics of the pA reaction.

The precise APA profile of a cell is essentially a balancing act between all these factors. It is, therefore, unsurprising that when we take a cell of different identity or a cell that has been perturbed in some way, we can see a change in their APA profiles. The major

challenge is going to be finding which of these events are physiologically significant and whether these large-scale shifts in APA profiles are key to the cell's identity and function.

1.5. IMPACT OF APA IN A CELLULAR CONTEXT

1.5.1. *Effects of UTR-APA on Transcript Stability*

In UTR-APA, the lengthening or shortening of the 3'UTR through APA can significantly impact upon the post-transcriptional regulation of the transcript. As 3'UTRs are thought to have an overall repressive role (Morris et al., 2012), cleavage and polyadenylation of a transcript at a pA site more proximal to the stop codon is generally considered to result in the removal of potentially repressive *cis*-elements in the 3'UTR, such as summarised in Table 1. This can, therefore, stabilise the transcript, or promote its translation, thus increasing the overall expression of that gene.

This principle has been demonstrated directly in quiescent muscle stem cells, where 3'-UTR shortening by APA in the myogenic regulator *PAX3* mRNA results in the removal of a *miR-206* target site and consequently enables *PAX3* to escape *miR-206*-mediated repression (Boutet et al., 2012). Conversely, 3' UTR lengthening in the DNA repair enzyme *MGMT* in glioblastomas serves as an example of where APA results in the inclusion of several miRNA-target sites and the subsequent repression of *MGMT* expression (Kreth et al., 2013).

Mayr *et. al.* confirmed that, in the cases of *CCND1*, *CCND2* and *IMP-1*, the shorter transcripts have an increased stability relative to the longer transcripts. In addition, reporter luciferase assays were used to compare protein levels when using solely the proximal pA site versus solely the distal pA site. In the case of *CCND2*, it was found that use of solely the proximal pA site, protein expression was increased up to 42-fold. (Mayr and Bartel, 2009). Subsequent mutation of all known miRNA target sites in the aUTR of *CCND1*, *CCND2* and *IMP-1* confirmed that miRNA-mediated repression accounted for some, but not all of the

reduction in protein expression. The remainder of this may be due to other, unknown instability elements, or altered translational rates. Finally, forced usage of solely the proximal pA site in the oncogene *IMP-I*, has shown that this alone can enhance transformation when compared to the use of the distal pA site (Mayr and Bartel, 2009). Together, this shows that differential use of pA sites can alter the stability of transcripts, affect the overall protein expression, as well as having significant impacts on major biochemical processes, such as cellular transformation.

When considering the global analysis of *cis*-elements located in the aUTRs, comparative analysis has shown that there are ~30% more miRNA target sites in aUTRs relative to constitutive UTRs (cUTRs) or single UTRs (sUTRs) (Ji et al., 2009). The miRNA target sites in aUTRs are also located in regions with higher AU content. This reduces the possibility that these target sites are involved in secondary structures, providing a better target for miRNAs (Ji et al., 2009). In addition, the lengthening of 3'UTRs by APA can potentially increase the distance of a target site to the 3'-end, a factor, which is correlated with a higher potency of the target site (Grimson et al., 2007). Combined, these factors indicate that aUTRs are more likely to be effectively targeted by miRNAs than either cUTRs or sUTRs.

Furthermore, when looking at the positioning of AREs within mature mRNAs, most AREs are found immediately upstream of distal pA sites (Gautheret et al., 1998). These AREs would, therefore, be present solely the transcript isoform in which the distal pA site is used, and result in specific destabilisation of this distal APA isoform. To support this, it was found that AREs are one of the most enriched motifs that are found in aUTRs, whereby the various transcript isoforms are subject to differential regulation (Legendre et al., 2006). Many examples of APA resulting in the inclusion or exclusion of diverse regulatory *cis*-elements, including AREs, have been characterised in individual genes, a sample of which are summarised in Table 5.

Table 5 - Examples of genes producing differentially regulated UTR-APA isoforms

Gene	Gene function	Summary	References
<i>CCND2</i>	Cell cycle regulator	3'UTR shortening is seen in cancer cell lines relative to normal tissues, thereby avoiding regulation by miR-15/16. Preferential use of the proximal pA site has been shown to increase the number of cells present in S-phase.	(Mayr and Bartel, 2009)
<i>CDC6</i>	Cell cycle regulator	pA at the proximal sites avoids miRNA-mediated repression, resulting in increased CDC6 protein levels. This is triggered by the potent proliferation signal 17 β -estradiol (E2), and may, therefore, be a mechanism by which the cell promotes cell cycle progression in response to proliferation signals.	(Akman et al., 2012)
<i>ELAVL1</i> (HuR)	RNA-binding protein	The aUTR region of HuR mRNA region contains an ARE region where HuR and TTP competitively bind, resulting in mRNA stabilisation or destabilisation respectively. This, therefore, creates an autoregulatory loop, which may amplify the pathological role of HuR.	(Al-Ahmadi et al., 2009)
<i>MAPT</i>	Stabilises microtubules, specifically in neurons	UTR-APA isoforms are differentially regulated in neuroblastoma cell lines, with miR-34 family members targeting solely the distal APA isoforms. MAPT protein is one of the key components of protein aggregates formed during Alzheimer's disease.	(Dickson et al., 2013)
<i>MGMT</i>	DNA repair	Glioblastomas are shown to shift pA site usage to a distal site, resulting in the inclusion of target sites for miR-767-3p, miR-181d and miR-648, thus reducing the expression of <i>MGMT</i> .	(Kreth et al., 2013)
<i>PAX3</i>	Transcription factor which controls myogenesis	In quiescent muscle stem cells, APA results in the production of <i>PAX3</i> , a key myogenic regulator, with a shortened 3'UTR allowing escape from regulation via miR-206, that targets the aUTR.	(Boutet et al., 2012)
<i>PDCL</i>	Putative modulator of heterotrimeric G proteins	Several AREs are located between 2 pA sites in the 3'UTR of <i>PDCL</i> , which results in a significantly shorter mRNA half-life of the longer transcript.	(Lazarov et al., 1999)
<i>ZFR</i>	Neuron development	<i>ZFR</i> APA isoforms are differentially regulated by miR-579, which itself is co-transcribed with the <i>ZFR</i> gene. miR-579 also regulates CPSF2, creating a negative feedback loop wherein transcription of <i>ZFR</i> results in the production of miR-579, which targets CPSF2, which then favours the proximal APA, which is then resistant to regulation by miR-579.	(Hinske et al., 2015)

Although many of these individual events have been shown to be physiologically significant, the pervasiveness of UTR-APA effecting stabilisation on a global scale is much less evident. If the large-scale shifts in UTR-APA described in the numerous global studies performed were to have a significant impact on the RNA stability, it would be expected that the overall mRNA abundances of these genes would either increase in the case of 3'-UTR shortening or decrease on 3'-UTR lengthening. However, this expected pattern is not consistently seen. De Klerk and Venema found that ~45% of genes that showed general 3'-UTR shortening were accompanied by changes in transcript levels, but, of these, there was only a slight bias towards an overall increase in transcript levels (de Klerk and Venema, 2012). Similarly, in several other studies, only modest, if any, changes in transcript levels were observed in the genes that show a shift in pA site choice (Elkon et al., 2012; Fu et al., 2011; Morris et al., 2012).

Finally, a recent analysis that directly assesses the relative mRNA stabilities of APA isoforms globally draws similar conclusions. In NIH 3T3 mouse fibroblast cells, of the 3463 detected genes that produce several APA isoforms, only approximately 500 differed in their relative mRNA stability (Spies et al., 2013). Interestingly, in ~40% of these APA events where significant changes in stability are seen, the longer isoform is actually more stable than the shorter transcript, suggesting the presence of stabilising *cis*-elements in the aUTR. A similar association between longer UTRs and increased stability has also been described in human adipocyte stem cells (Spangenberg et al., 2013). Global studies in yeast, however, suggest that APA can have a profound impact on the stability of transcript isoforms, even in isoforms differing by a single nucleotide (Geisberg et al., 2014; Gupta et al., 2014).

1.5.2. Effects of UTR-APA on Translational Efficiency

As with stability, a number of studies have identified examples where APA events lead to alteration in translational efficiency (Akman et al., 2012; Mayr and Bartel, 2009; Pinto et al., 2011; Tranter et al., 2011; Yoon et al., 2013). With the help of reporter assays, protein output of constructs containing the short and long UTRs of *CCND2*, *IMP-1* and *DICER1* were compared and 3'-UTR shortening was found to be generally associated with increased translational efficiency, and thus increased expression (Mayr and Bartel, 2009). However, this view has become more complex in recent years as opposite effects have also been seen where the distally polyadenylated transcript is actually translated with a higher efficiency (Pinto et al., 2011; Spies et al., 2013; Yoon et al., 2013).

One of the most well-characterised examples of the latter is the anxiety-related gene *SERT* (serotonin transporter). Here, hnRNPK-binding in the aUTR of the distally polyadenylated *SERT* transcript prevents *miR-16* binding and promotes the translation of the

distal isoform relative to the proximal isoform (Yoon et al., 2013). A similar situation is also seen in the *Drosophila* cell cycle gene *polo*, where its longer APA isoform is translated at three times the rate of the shorter isoform (Pinto et al., 2011).

Global studies have found similar conflicting trends regarding correlations between APA mRNA isoforms and translational efficiency. Studies in HEK293T cells show a bias towards a greater polyribosomal association of the shorter UTR-APA mRNA isoforms (Sterne-Weiler and Martinez-Nunez, 2013). This is somewhat in contrast with a recent study using ribosomal profiling assays performed in mouse NIH 3T3 cells, where increases in ribosomal association of the distal APA isoform relative to their proximal counterpart were seen (Spies et al., 2013). However, these effects were minimal, with a median of less than 3% increase in ribosomal association of the distally polyadenylated transcript. This latter study concludes that the large majority of the APA events are unlikely to have a physiologically significant effect on either transcript stability or translational efficiency, at least in NIH 3T3 cells. Studies that couple quantitative mass spectrometry and 3'end sequencing also show that in proliferating T-cells 3'UTR shortening has very little effect on mRNA stability or protein abundance (Gruber et al., 2014).

1.5.3. Effects of UTR-APA on RNA localisation

There are only a small number of examples that demonstrate a role of UTR-APA in regulating the subcellular localisation of mRNA isoforms that have been presented to date. The inclusion of inverted *Alu* repeats in the distal APA isoform of *NICN1* has been shown to lead to its nuclear retention, whereas the proximal APA isoform, which lacks these sequences, is freely exported to the cytoplasm (Chen et al., 2008). The general role of 3'-UTRs in the localisation of mRNAs in highly polarised cells is well known (Andreassi and Riccio, 2009), and differential localisation of APA mRNA isoforms in neurons has been well

characterised. aUTRs play a critical role in regulating the localised translation of *BDNF* APA mRNA isoforms in neurons where the distal APA isoform localises to the dendrites and the proximal APA isoform is restricted to the soma (An et al., 2008). 3'-UTR lengthening in the brain is one of the most striking examples of tissue- specific APA (Miura et al., 2013), which may serve as a platform to control the localisation of these transcripts.

A further example of UTR-APA controlling localised translation has also been seen with the membrane protein *CD47*. The distal APA isoform of *CD47* allows binding of HuR and SET resulting in its localisation to the plasma membrane, where it is translated and the consequent protein inserts to the membrane. Conversely, the proximal APA isoform of *CD47* locates primarily to the endoplasmic reticulum (Berkovits and Mayr, 2015). Excluding work from the data presented here (Neve et al., 2016), only one other example of the effect of APA on subcellular localisation has been performed. This study was, again, in neurons (Taliaferro et al., 2016), comparing the RNA in the dendrites to that found in the soma. They isolated around 8% (~400) of expressed genes with APA isoforms displaying differential subcellular localisation between the dendrites and soma of neurons.

1.5.4. Coding Region APA in Impact on the Transcriptome and Proteome

In contrast to UTR-APA, CR-APA by definition, will affect the protein identity, as introducing a pA site upstream of a translational stop codon will result in a protein with a different C-terminus. This may be a simple truncation in the case of exonic CR-APA, or more commonly a different C-terminus in composite-intronic CR-APA or skipped-intronic CR-APA (Figure 1).

In the case of skipped-intronic APA, polyadenylation is coupled to an alternative splicing event of the alternative terminal intron and is therefore in dynamic competition with

this splicing (Evsyukova et al., 2013). Thousands of these intronic pA sites have been identified, but largely remain dormant through inhibition by factors such as U1 snRNP (C. Yao et al., 2012). However, these intronic pA sites can indeed be activated, as is seen during increased proliferation (Elkon et al., 2012).

The classic example of CR-APA is that of the switch between the membrane-bound form of IgM in B-cells to the secreted form in plasma cells (Alt et al., 1980). This switch is thought to be governed by changes in the CstF-64 levels, whereby an elevation of CstF-64 in plasma cells activates the weaker proximal pA site, triggering the switch to the upstream pA site, ultimately removing the membrane anchoring C-terminus (Takagaki et al., 1996).

One important aspect of CR-APA isoforms is that they do not result in the introduction of premature translational termination codons, so they are not subject to nonsense-mediated decay and therefore have the capability of producing alternative protein isoforms with potentially distinct physiological roles (Lejeune and Maquat, 2005). Furthermore, the polyadenylation event within the coding region of an exon can convert a tyrosine codon (TAT) into a stop codon (TAA) as was found with the truncated form of glutamyl-prolyl tRNA synthetase (EPRS) (P. Yao et al., 2012).

1.6. CONSERVATION OF APA

Interestingly, when looking at the evolution of the genome from *Caenorhabditis elegans* (*C. elegans*) to humans we see a large increase in the level of noncoding sequences, including that of 3'UTRs, while both maintain similar numbers of genes. Compared to *C. elegans*, humans have about twice the number of genes that contain multiple pA sites, and the average UTR length increases from ~140 nt in *C. elegans* to ~1200 nt in humans (Jan et al., 2011). This, therefore, suggests that more complex organisms may have higher levels of post-transcriptional regulation targeting these extended 3'UTR regions.

Comparative genomics has also been used in several studies to assess the conservation of APA between species. Global features, such as the general *cis*-elements that comprise the pA site, and the presence of multiple pA sites at a significant portion of genes are conserved between human and mouse (Tian et al., 2005). However, when the conservation of individual pA sites between human and mouse are considered, only ~500 genes contain conserved tandem pA sites. Consequently, it has been suggested that “in general, alternative polyA sites are species-specific and involve minor, non-conserved sites that are unlikely to play essential roles” (Ara et al., 2006).

Derti *et. al.* see a similar lack of conservation of pA sites (Derti et al., 2012). Of the 439,390 pA sites they identified in humans, only 2590 (0.6%) of them were found to have pA sites within the equivalent 30 base region in all four other mammals considered (rhesus, dog, mouse, rat) (Derti et al., 2012). Interestingly, it was found that of the conserved pA sites, the read numbers were more similar in equivalent tissues across species, rather than between tissues in the same species (Derti et al., 2012). However, since this approach is looking at absolute read numbers over individual pA sites and not looking at the relative usage of tandem pA sites, this may simply reflect a conservation of tissue-specific gene expression rather than the conservation of APA patterns themselves.

Of the conserved tandem pA sites, it is the distal pA sites that are conserved at a higher level than the proximal pA sites (Martin et al., 2012), and in general, it is the distal pA sites that have the highest processing efficiency (Beaudoing, 2000). This, therefore, suggests a mechanism by which the distal pA sites serve as the default pA site and functions to ensure efficient 3' end processing and transcription termination while the proximal pA site functions as a regulatory site that can potentially be activated to remove any downstream regulatory *cis*-elements (Shi, 2012).

Experiments in yeast have compared the conservation of APA between different yeast species including *Saccharomyces cerevisiae*, *Kluyveromyces lactis*, and *Debaryomyces*

hansenii (Moqtaderi et al., 2013). Here they find that, despite the similarities in the general sequence preferences surrounding the pA sites, the APA profiles show species specificity. Ultimately, they conclude that much of the heterogeneity seen in the pA sites is likely to be biological noise as defined by “biochemical events that occur *in vivo* as a result of a low-specificity process that has not been subjected to evolutionary optimisation.” (Moqtaderi et al., 2013).

Recently, single cell studies have started to expose the heterogeneity of pA site usage within a population of homogenous stem cells (Velten et al., 2015). Here they found that individual cells differ significantly in their choice of pA site choice. Interestingly, this study finds that the heterogeneity in pA site choice seen in these cells surpasses that of what would be expected if pA site choice was random, thereby concluding that this variability seen has the potential to contribute to the functional cell-to-cell heterogeneity (Velten et al., 2015).

1.7. ISSUES SURROUNDING THE GLOBAL ANALYSIS OF APA

APA is a well-studied field, but still has some issues surrounding the general approaches that are adopted in studying APA on a transcriptome scale. Firstly, it is generally considered that 3'UTR shortening acts to increase the stability of the transcripts through the exclusion of instability elements, including miRNA target sites and AREs (Sandberg et al., 2008). However, this poses a problem that is rarely addressed. As the proximal and distal APA isoforms potentially have varying levels of stability, the observed shortening of 3'UTRs could be the result of two separate mechanisms: 1) a change in pA site choice, resulting in an increased production of the proximal APA isoform; 2) an increase in post-transcriptional regulation targeted to the aUTR, which would result in the selected degradation of solely the distal APA isoform.

The studies described above showed large-scale shifts in the APA landscape in a variety of different situations, including cell transformation, differentiation and tissue-specific samples. However, each of these has been associated with significant changes in miRNA expression profiles (Calin and Croce, 2006; Letzen et al., 2010; Liang et al., 2007). When you consider that in mammalian systems $\geq 84\%$ of miRNA-mediated repression is through a reduction in mRNA levels (Guo et al., 2010), it is hard to see how these altered miRNA profiles would not have some effect on the observed changes in APA profiles. Indeed, in studies looking at how pA site usage can be used to improve miRNA-target predictions, it was found that cytoplasmic repression by miRNAs can shape the APA isoform landscape in the cell (Nam et al., 2014).

Furthermore, it is well known that different cancer subtypes have distinctive miRNA signatures (Calin and Croce, 2006), and one would expect these changes to be observed in the steady-state APA profile. Liaw *et al.* highlight this by cross-referencing miRNA profiles with previous APA data showing apparent 3'UTR shortening in the breast cancer cell line, MCF-7 (Fu et al., 2011; Liaw et al., 2013). Firstly, it was shown that genes with target sites for upregulated miRNAs in their aUTR showed a larger degree of 3'UTR shortening relative to those without target sites, the extent of which is influenced by the number of target sites. In addition, a search on the positioning of target sites for miRNAs upregulated in MCF-7 cells showed that they were enriched in the aUTRs of genes where 3'UTR shortening was seen, relative to down-regulated miRNAs. Combining this data, it was concluded that selective degradation of solely the longer transcript isoforms by upregulated miRNAs has a significant impact that resulted in the observed changes in APA profiles in MCF-7 cells (Liaw et al., 2013).

Additionally, Legendre et. al., using EST analysis looked at the ratio of different transcript isoforms where solely the distal APA isoform contained a specific miRNA target site. Here they saw that, in tissues where the corresponding miRNA was upregulated, the

distally polyadenylated transcript was less abundant (Legendre et al., 2006). Together, these results suggest that altering miRNA profiles may have a significant impact on the observed APA profile, even though the true pA site choice may not have altered. A similar conclusion was drawn when looking at APA profiles between cancer subtypes where shifts in APA profiles were shown to be a balance between shifts in pA site choice and changes in the differential regulation of APA isoforms (Singh et al., 2009).

Some studies have indeed observed that this is a problem, but it remains to be fully addressed. The most common method used to try and address this is to look at the expression levels of all transcript isoforms (Lembo et al., 2012; Morris et al., 2012; Sandberg et al., 2008). This approach argues that if observed shifts in APA patterns are indeed due to a change in pA site choice, the increased proportion of the shorter transcripts leads to a lower overall decay rate, as 3'UTRs have an overall repressive nature. The expression level observed would, therefore, increase in this situation. Conversely, if the shift in APA pattern is due to an increase in the decay rate of solely the longer transcript, a decrease in expression would be seen. This method has its limitations, as any number of mechanisms, aside from APA, can alter transcript levels of a gene. Furthermore, these tests are performed as a mass analysis. Therefore, if the majority of the shortening events were truly APA, this could mask a large number of other shortening events that are mediated by changes in differential degradation.

The only other previously proposed global approach to address this problem is the use of RNA Pol II occupancy data to compare the abundance of RNAPII in the aUTR compared to the constitutive UTR (Elkon et al., 2013). If the observed shift in APA is truly at the level of pA site choice, this should be reflected by a change in RNAPII occupancy in the aUTR. However, to the author's knowledge, this approach has never been put into practise.

1.8. TECHNIQUES TO STUDY SINGLE GENE APA

There is a wide range of traditional molecular biology techniques that can be used to analyse APA isoforms, each with the advantages and disadvantages. These include quantitative reverse-transcriptase PCR (RT-qPCR), northern blotting, RNase protection assays and 3'-rapid amplification of cDNA ends (3'RACE).

In RT-qPCR, using primers targeted at the aUTR and cUTR region of an individual APA event, the proximal and distal APA isoforms can be measured. The ratios of the aUTR to cUTR in different conditions can tell you any changes in the relative abundances of these isoforms. However, there are drawbacks since the amplification frequencies of distinct primer pairs means that an absolute ratio between the proximal and distal pA isoforms is hard to obtain, and the proximal pA isoform can be difficult to isolate individually as there are no sequences unique to it.

In northern blotting, the RNA from each sample is run on a polyacrylamide or agarose gel and transferred to a nylon membrane. A radioactive probe targeting the region of interest is then hybridised with their targets. Using this method APA isoforms can be separated by their difference in size, and probes targeted at either the cUTR or the aUTR can be used to distinguish the different APA isoforms. Difficulties in using northern blotting to analyse APA can arise when there is extensive alternative splicing in the gene of interest as this results in many bands and deciphering the exact identity of each band can be tricky.

RNase protection assays rely on the hybridisation of a radioactively labelled RNA probe and its target. This hybridisation protects it from degradation by RNase H, which cleaves all remaining single-stranded RNA. The digestion products can then be run on a polyacrylamide gel for analysis. Probes spanning pA sites can therefore quantitatively assess the usage of each pA site.

3'RACE uses a reverse transcriptase primer with an oligo(dT) region coupled to an additional adapter sequence. This preferentially primes polyadenylated transcripts for reverse

transcription. In the subsequent PCR reaction, a forward primer upstream of the suspected pA site and a reverse primer complementary to the adapter sequence on the RT primer can be used to map the exact location of the pA site. This approach can be combined with real-time PCR to quantitatively assess relative abundances of different pA isoforms. As with RT-qPCR, due to variations in primer efficiency, the absolute ratio between pA site usage cannot be obtained using this method. Furthermore, as this essentially relies on the specificity of solely the forward primer, nested PCR may be required to isolate just the gene of interest (Scotto-Lavino et al., 2006).

1.9. TECHNIQUES TO STUDY GLOBAL APA

1.9.1. *EST analysis from Sanger sequencing*

In the early years of large-scale APA analysis, the majority of work was done using expressed sequence tag (EST) databases (Beaudoing, 2000; Gautheret et al., 1998). ESTs are a short sub-segments of cDNA that have been cloned and subsequently sequenced. To analyse APA from ESTs, data is obtained from large online databases, such as GenBank, and these sequences can subsequently be aligned to their corresponding genome. Using unaligned adenosines at the 3'-ends of reads or thymidines at the 5'ends of reads are interpreted as deriving from the very 3'end of a polyadenylated transcript and can thus be used to map pA sites. The relative abundances of ESTs mapping to each pA site can be used to indicate the general levels of the extent to which each pA site is used, and, since the sample origin of each EST is generally known, any switches in pA site usage can be identified. This approach was used in the first large-scale analysis of differential pA site usage between tissues (Gautheret et al., 1998).

Limitations with this approach come from the fact that the EST databases, while large, lack the depth required for comprehensive APA analysis. This stems from the fact that

only a small percentage of ESTs will derive from the 3'-end of a gene. In addition, much like many of the methods discussed below, it suffers from the internal priming issue where A-rich genomic regions are falsely annotated as pA sites due to the binding of the oligo(dT) primer in the reverse transcription reaction. While key in the original identification of pA sites, these drawbacks mean that APA analysis from ESTs is no longer widely used.

1.9.2. *Microarrays*

Microarrays are generally used to assess the overall level of gene expression in a population of cells. This involves the binding of the microarray probes hybridising to their targets in the test DNA. This binding results in a signal that is proportional to the level of target in the test DNA. Therefore, using probes in the 3'UTRs of genes, targeting the cUTRs and aUTRs, the relative levels of different APA isoforms can be analysed. This approach was used in the PLATA (probe level alternative transcript analysis) method to analyse APA after the activation of T-cells (Sandberg et al., 2008). The PLATA approach relies on already knowing where the pA sites lie and then categorising the probes into cUTR versus aUTR probes. A subsequent study, however, used an approach which looked to identify "segmentation points" where the signal intensity between 2 subsequent probes changed, thereby indicating a pA site (Singh et al., 2009). However, due to the requirement for multiple probes per UTR, and the lack of ability for *de novo* identification of pA sites, the use of microarray analysis to study APA has dwindled.

1.9.3. *RNA-seq*

Whole transcriptome RNA sequencing can also be used to analyse APA. Similar to EST-based studies, reads mapping the very 3'-end of a transcript, and therefore containing non-genomically encoded adenosines, can be used to map pA site and quantify their usage

(Siegel et al., 2010; Wilhelm et al., 2008). However, since the proportion of reads overlapping pA sites is very low, the level of coverage required to comprehensively analyse APA is rarely obtained.

RNA-sequencing data can also be used to analyse APA by looking at all reads spanning the UTR and using the coverage across the UTR. Initial studies using this approach relied upon previously annotated pA sites to define the cUTR and aUTR (Katz et al., 2010; Wang et al., 2008; Wehrspaun et al., 2014). However, the recent DaPars (dynamic analysis of alternative polyadenylation by RNA-sequencing) approach uses a linear regression model to identify pA sites as switch-points in UTR coverage when comparing multiple RNA-seq samples (Masamha et al., 2014; Xia et al., 2014). This method is further discussed in Section 1.10.3.

Finally, RNA-seq data can also be useful for connecting far distal pA sites identified by targeted 3' end sequencing to specific genes as these can be tens of kilobases downstream of the annotated genes (Miura et al., 2013).

1.9.4. 3'-end targeted deep sequencing

The most popular method for studying APA on a transcriptome scale in recent years is 3'-end targeted deep sequencing. There have been a huge number of variations on this, but all essentially aim to quantify the usage of each pA site by sequencing the region directly upstream of the cleavage point. These methods fall into three categories - direct RNA sequencing, 3' end capture and deep sequencing and oligo(dT)-based priming approaches. Each category is described briefly below and an overview of each technique is outlined in Table 6.

Methods of 3'-end capture and deep sequencing utilise different techniques to isolate solely the 3' end of the mRNA, and this pool of 3' ends is then used as the input for library

preparation and deep sequencing. These methods include the ligation of biotinylated adapters to the polyA tail and subsequent purification using streptavidin (Jan et al., 2011), or the use of oligo(dT)-coupled beads (Hoque et al., 2014) to isolate the 3' ends. Depending on the exact method, these approaches suffer from varying levels of internal priming and biases can also be introduced during the adapter ligation steps in library construction (Hafner et al., 2011).

Oligo(dT) priming based approaches use a reverse transcriptase primer comprising an anchor sequence, which is later used for PCR amplification, followed by a string of thymidines. This, therefore, aims to bind specifically to the polyA tail. Second strand synthesis and PCR amplification follow, resulting in a sequencing library enriched for the very 3' ends of mRNAs. Drawbacks of this approach involve significant levels of internal priming necessitating bioinformatic pipelines that filter these out. This can be an issue as it can result in filtering out true A-rich pA sites.

Unlike oligo(dT) priming and 3' end capture approaches, direct RNA sequencing (DRS) does not rely on PCR amplification, thus avoiding potential biases that can be introduced at this step. DRS uses the Helicos BioSciences system which starts by using an oligo(dT) coated surface to which the polyA tail binds to. Reverse transcription using only dTTP is then used to fill in the entire polyA tail. Sequencing is then initiated from the most 3-prime non-A base (Ozsolak et al., 2010). Despite this resulting in very low internal priming, it can still occur when an internal A-rich region does bind to the oligo(dT) coated surface (Sherstnev et al., 2012). The major drawback of this approach is the error rate of the Helicos System, which is typically in the range of 4-5% (Ozsolak and Milos, 2011). Furthermore, Helicos BioSciences ceased trading in 2012 so this approach can no longer be widely used.

Table 6 - Methods of 3'end targeted deep sequencing

Category	Technique	Overview	References
3'end capture	3P-Seq	A biotinylated primer is added to the polyA tail and solely the polyA fragments can then be isolated using streptavidin. RT with just TTP is used to fill in polyA tail and RNase H is then used to cleave the polyA tail leaving just the very 3'end which is then used as input for sequencing library preparation.	(Jan et al., 2011)
	3'READS	Uses a unique chimeric CU ₅ T ₄₅ oligo isolation system, which completely eradicates internal priming and amplification of oligo adenylated transcripts.	(Hoque et al., 2013)
Direct RNA sequencing	DRS	As described in the text	(Lin et al., 2012; Ozsolak et al., 2010; C. Yao et al., 2012)
Oligo(dT)-based priming	3Seq	Standard oligo(dT)-based priming technique using an oligo(dT) ₂₅ containing RT primer, and sequencing the terminal 25 bp upstream of the cleavage site to map the pA site.	(Beck et al., 2010; Wang et al., 2013)
	3'end RNA Seq	Standard oligo(dT)-based priming technique using paired-end sequencing to obtain strand-specificity	(Yoon et al., 2012; Yoon and Brem, 2010)
	3'-end-seq	Also includes <i>in vitro</i> transcription step after reverse transcription to increase the quality of the input RNA and as an amplification step.	(Haenni et al., 2012)
	3'Seq	Adapted slightly from 3Seq	(Elkon et al., 2012; Jenal et al., 2012; Morris et al., 2012)
	3'T-fill	Directly prior to sequencing, the polyA tail is filled using TTP resulting in sequencing starting from the base directly upstream of the cleavage site.	(Wilkening et al., 2013)
	A-seq	Uses an RT primer consists of an anchor nucleotide, followed oligo(dT) sequence with a stem-loop containing the adaptor sequence for priming the subsequent PCR reaction, in the middle of the oligo(dT).	(Martin et al., 2012)
	MAPS	Standard oligo(dT)-based priming using a random primer for second-strand cDNA synthesis.	(Fox-Walsh et al., 2011)
	PAS-Seq	Variation to the standard oligo(dT)-based priming technique involves using the terminal transferase activity of MMLV reverse transcriptase, which allows generation of cDNAs with linkers in a single RT step, thus skipping several enzymatic steps.	(Shepard et al., 2011)
	PolyA-seq	Standard oligo(dT)-based priming using a random primer for second-strand cDNA synthesis.	(Derti et al., 2012)
	Quant-seq	Commercialised standard oligo(dT)-based priming using a random primer for second-strand cDNA synthesis.	(Moll et al., 2014)
	SAPAS	Oligo(dT)-based priming using template switching and optimised primer anchoring in the RT-reaction to avoid sequencing in the long polyA tail.	(Fu et al., 2011; Tian et al., 2014; You et al., 2014)
	SMPSS	Uses a single-molecule system based on the HeliScope single molecule sequencer. It is amplification and ligation-free, allowing very little bias in quantitation.	(de Klerk and Venema, 2012)

1.10. STRATEGY

1.10.1. Aims of the project

The main aim of this project was to investigate the impact of APA in the regulation of gene expression in eukaryotes. There are now extensive examples, as described above, of global APA studies that see general shifts in APA patterns between different cell types or cell states. The important thing now is to identify and characterise the differentially regulated APA isoforms as these will be the APA events that are most likely to have a physiological

impact on the final protein production. Once differentially regulated APA isoforms were identified, the question became what are the key post-transcriptional regulatory pathways that govern this differential regulation.

1.10.2. *The subcellular fraction approach versus previous approaches*

There have been two notable previous studies that aimed to identify, on a transcriptome scale, APA isoforms that differ in their stability (Gupta et al., 2014; Spies et al., 2013). Both of these attempted to quantify the decay rates of individual APA isoforms after blocking transcription either by actinomycin D in mouse cells (Spies et al., 2013) or by using temperature-sensitive strain of *Saccharomyces cerevisiae* containing a modified *rpb1-1* allele that arrests RNAPII transcription at 37°C (Gupta et al., 2014).

There are significant issues with both of these approaches in that transcriptional arrest can massively perturb the cell and induce major stress responses. These stress responses can majorly influence the RNA decay pathways and result in the stabilisation of specific transcripts (Bhattacharyya et al., 2006; Brennan and Steitz, 2001). The use of these approaches, therefore, ends up calculating the stabilities of these mRNAs under non-physiological conditions.

To combat this, this thesis develops a novel subcellular fractionation approach to compare the APA profiles in the cytoplasmic and nuclear fractions of the same cell population. The rationale behind our approach is that post-transcriptional regulatory mechanisms will primarily act on cytoplasmic mRNA (Stoecklin et al. 2006; Valencia-Sanchez et al. 2006). Thus, any APA isoforms that have different cytoplasmic decay rates, and/or nuclear-cytoplasmic export rates, will differ in their relative isoform abundance in the cytoplasmic and nuclear fractions. Therefore, using this approach we can identify APA isoforms that are differentially regulated, and thereby the APA events that are likely to be

contributing to the regulation of gene expression levels. The expected pattern of differentially regulated APA isoforms upon 3'-end sequencing and generic RNA-seq of subcellular fractions is shown in Figure 2.

The major advantage of this subcellular fractionation approach is that the RNA used for APA analysis is from cells that are completely unperturbed. Secondly, differentially regulated APA isoforms can be identified simply through the comparison of two samples, whereas transcriptional arrest experiments rely on multiple time points (Spies et al., 2013). Furthermore, this approach can be used on any cellular population whereby a clean subcellular fractionation can be obtained, meaning tissue samples can be analysed as well as cultured cells.

Despite these advantages, the subcellular fractionation approach does have limitations. Firstly, there are two different mechanisms that could result in the ratios of APA isoforms differing in the cytoplasm compared to the nucleus: through a difference in stability or a difference in nuclear export rates. These two mechanisms cannot be dissociated and individual gene studies are required to examine exactly how APA isoforms are differentially regulated.

Furthermore, with the transcriptional arrest approach a half-life of each APA isoform can be calculated (Spies et al., 2013). However, the ratios of APA isoforms in the nuclear and cytoplasmic fraction alone cannot give you this information; it can only highlight that there is a difference in the regulation of these APA isoforms.

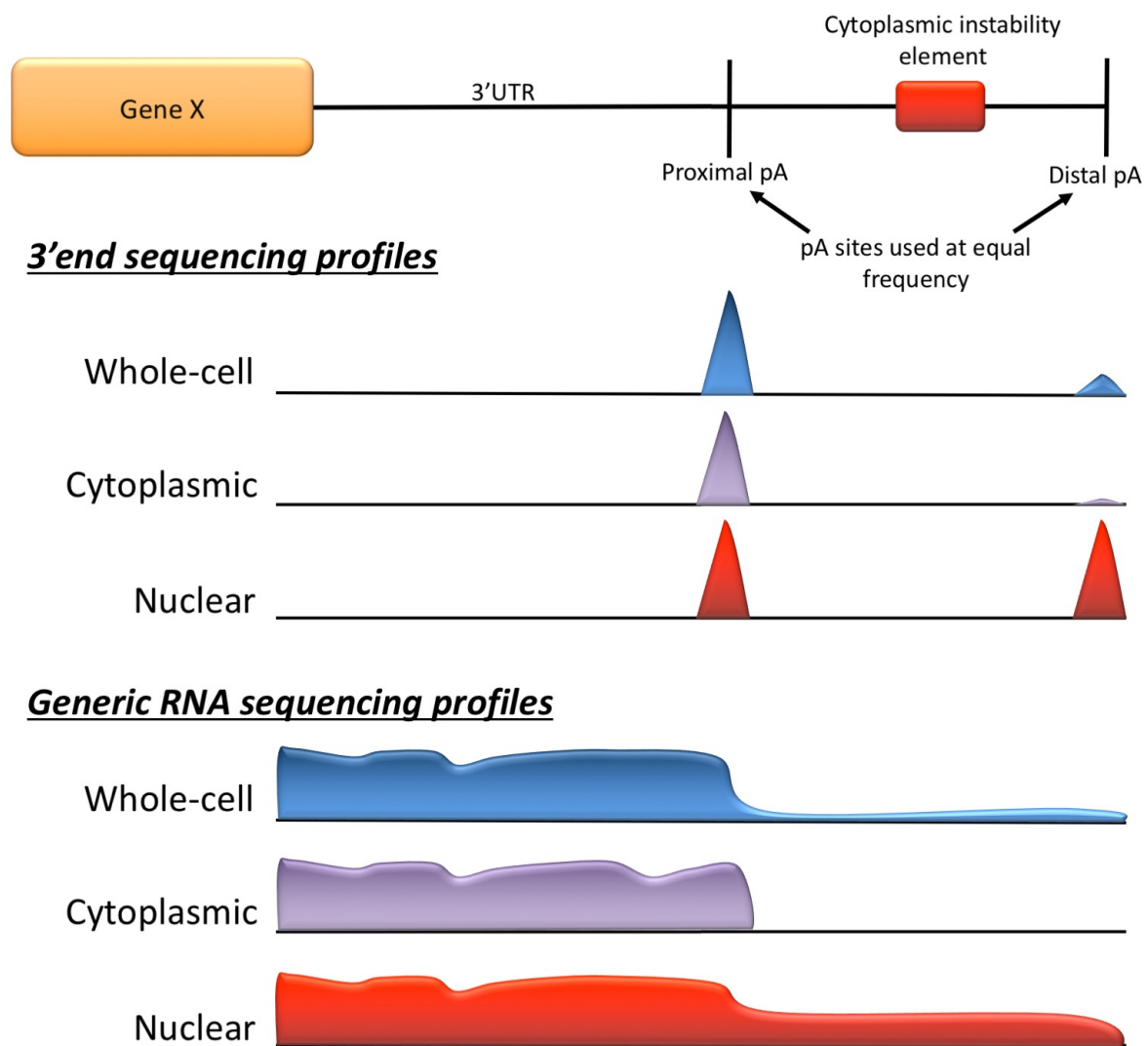


Figure 2: Sequencing profiles over a gene with differentially regulated APA isoforms

Upper panel shows a theoretical gene with two pA sites in the genes 3'UTR with an alternative UTR (aUTR) containing a cytoplasmic instability element.

Middle panel shows the expected 3'end-targeted sequencing profiles that would be expected if the two pA sites were used at equal frequencies. The nuclear profile shows equal levels of the proximal and distal APA isoforms while both the cytoplasmic and whole-cell fractions show an underrepresentation of the distal APA isoform due to the cytoplasmic instability element in the aUTR.

Lower panel shows the expected generic polyA+ RNA sequencing profiles that would be expected if the two pA sites were used at equal frequencies. The nuclear profile shows a small drop in coverage after the proximal pA site, but the whole-cell and cytoplasmic profiles show a much more significant drop off after the proximal pA site due to the cytoplasmic instability element in the aUTR.

1.10.3. Choosing pA mapping method

As previously discussed, there are a plethora of techniques for using deep sequencing to map pA sites (Table 6). One of the more recent methods is 3'READS, which is a method of isolating the ~50 nucleotides from the 3'terminus of all polyadenylated RNA transcripts in any given RNA sample (Hoque et al., 2013). An overview of the 3'READS library preparation is shown in Figure 3.

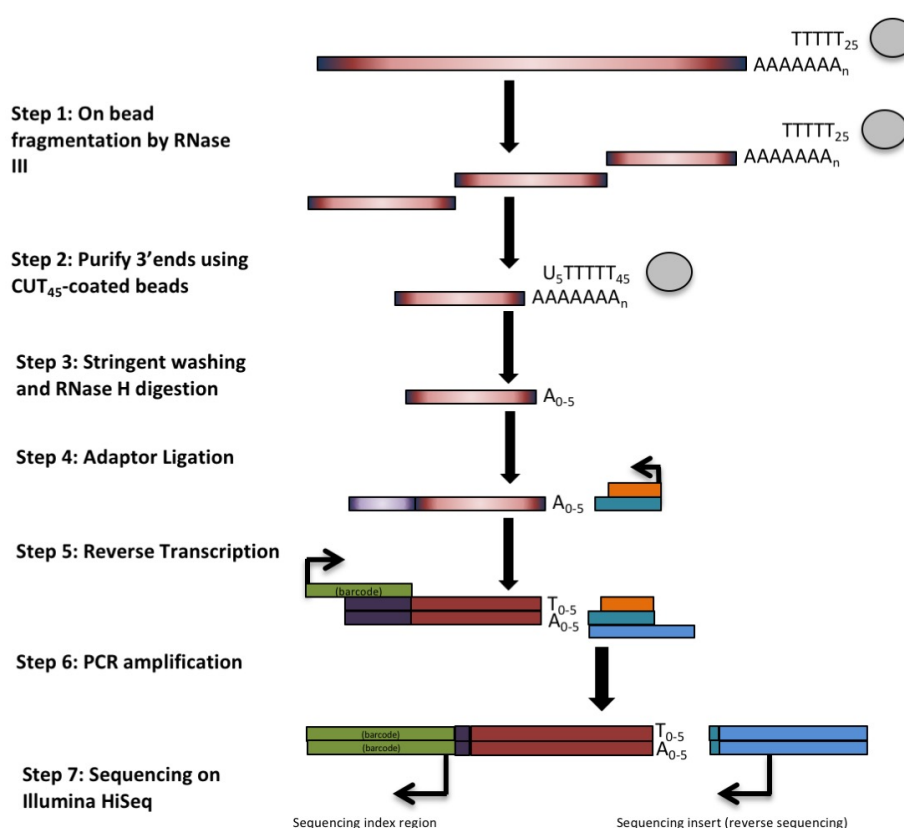


Figure 3: Schematic of 3'READS library preparation

Starting with a given sample of total RNA, polyA⁺ RNA is firstly isolated through oligo(dT) selection, then fragmented using RNase III. Subsequently, the polyadenylated fragments are rebound to magnetic beads using a chimeric oligo containing a stretch of 45 thymidines followed by five uridines. These thymidines create a DNA/RNA hybrid which is then cleaved by RNase H. The terminal uridines nucleotides create an RNA/RNA hybrid and protect the first few adenosines from cleavage. This is important as it leaves non-genomically encoded adenosines in the stretch of DNA sequenced, which are used to confirm that the reads have come from the 3'terminus of a polyadenylated transcript, and is not the result of internal priming. These fragments are then purified and subject to deep sequencing.

The key aspect of the technique is the use of a chimeric DNA-RNA oligo to capture the polyA tail. This oligo is composed of 45 thymidines (DNA) at its 5'end and 5 uridine (RNA) residues at its 3'end (CUT₄₅). The oligo is coupled to streptavidin-coated magnetic beads by a biotin tag on the 5'end of the oligo. The polyA tail of the fragmented RNA is hybridised to the CUT₄₅ oligo and, after several stringent washes, released by digestion with RNase H. Since RNase H only cleaves DNA/RNA hybrids, the five uridines essentially protect the first few adenosines of the polyA tail from cleavage. This, therefore, leaves up to five non-templated adenosines in the stretch of RNA released and is therefore subsequently present in the RNA fragment to be sequenced. These non-genomically adenosines can, therefore, be used in the data analysis to identify sequencing reads that are definitively from the 3'end of polyadenylated transcripts, and these are termed polyA site supporting (PASS) reads. This is a key advantage over other techniques as it is the only technique that can definitively eradicate sequence reads that originate from genomically encoded A-rich regions known as internal priming.

Furthermore, a high stringency wash step during the CUT₄₅ binding ensures that only fully polyadenylated transcripts, and not oligoadenylated transcripts, are isolated and sequenced (Hoque et al., 2013). This avoids complications that may arise from contamination of oligoadenylated transcripts that result from non-canonical polymerases such as the ones involved in exosome-mediated RNA decay (Schmidt and Norbury, 2010). Therefore, due to these significant advantages, 3'READS was one of the methods chosen for global pA site mapping for this project.

The QuantSeq 3' mRNA-Seq kit (Lexogen) was also used in conjunction with 3'READS. This uses oligo(dT) priming to prime from the polyA tail and isolate the 3'end of the transcripts for sequencing. The advantage of this approach is much lower input material is required (500 ng versus 25 µg used in 3'READS). The internal priming issue encountered

with this method can easily be avoided by only considering sequencing reads that span previously known pA sites that are mapped using 3'READS.

In addition to 3'end sequencing methods, DaPars was also used throughout this project (Masamha et al., 2014). DaPars calculates APA shifts by examining standard polyA+ RNA-seq coverage over annotated 3'UTRs so has the advantage of being able to analyse the vast array publicly available RNA-seq datasets in the context of APA. The approach uses a linear regression model to identify switch points in sequencing coverage over annotated 3'UTRs when comparing two samples. The basis of this approach is that if there is a switch between pA sites within the annotated UTR between samples, this will result in a change in the sequencing reads over this region, and the point at which this coverage changes will correspond to the proximal pA site. Subsequently, analysing the coverage in the aUTR region can be used to calculate the percentage distal usage index (PDUI) for each sample, and ultimately the change of the PDUI (Δ PDUI) between the two samples.

There are however a couple of drawbacks associated with using DaPars analysis. Firstly, for the algorithm to calculate the coverage over 3'UTRs, it needs to be provided with the coordinates of these regions. This is done by using standard genome annotations, but this relies on the end of the annotated 3'UTR being the distal pA site, and the proximal pA site lying within the annotated UTR. Therefore, this approach cannot identify CR-APA events and can potentially miss events where the distal pA site lies downstream of the annotated UTR. For example, this would be a major issue when looking at APA profiles in the brain as 3'UTR-extension beyond the annotated 3'UTRs is widespread (Miura et al., 2013). Finally, DaPars also indirectly predicts the proximal pA sites based on changes in the UTR coverage. However, through coupling this approach to our known pA site coordinates this issue is avoided. This approach is therefore used to identify differentially regulated UTR-APA isoforms from publicly available RNA-seq data from a variety of cell types.

1.10.4. Identifying mechanisms responsible differential APA isoform regulation

After identifying differentially regulated APA isoforms, the question then became what are the main regulatory pathways that lead to this differential regulation? From previous studies, the most likely candidate in regards to differential stability was miRNA-mediated degradation (Di Giammartino et al., 2011; Sandberg et al., 2008). To identify the subset of genes which are subject to this pathway, an inducible DICER1 knockdown HEK293 cell line was used (Schmitter et al., 2006) to compromise the miRNA biogenesis machinery. Combining this approach with 3'READS, allowed the monitoring of changes in APA profiles in the subcellular fractions upon DICER1 knockdown.

1.10.5. Individual gene studies

Once these approaches identified genes of interest where the APA isoforms appeared to be subject to differential regulation, individual gene studies were used to try and identify the exact mechanism by which this differential regulation occurred. This involved using a range of methods as described in Section 1.8, including northern blotting, standard RT-PCR and RT-qPCR.

2. CHAPTER 2 – MATERIALS AND METHODS

2.1. INTRODUCTION

The work described in this thesis can be split into wet lab work and bioinformatics analysis. The wet lab work includes subcellular fractionation, deep sequencing library preparation as well as individual gene analysis involving northern blotting, RT-PCR, RT-qPCR, chromatin immunoprecipitation (ChIP) analysis. The bioinformatics analysis includes the analysis of 3' end sequencing data and full RNA-sequencing as well as motif analysis, and gene ontology analysis. Materials and methods for both parts are described below, considering the wet lab techniques first.

2.2. CHEMICALS AND KITS

All chemicals and kits that were used in this study are listed below.

Table 7 - Chemicals and kits used in this study

Chemical	Manufacturer and Catalogue Number
10 x Taq polymerase buffer with MgCl ₂	NEB M0267
Acid phenol	Sigma P4682
Actinomycin D	Sigma A9415
Agarose	Sigma A9539
Alternative NP-40	Merck Millipore 492016
BIX 01294 trihydrochloride hydrate	Sigma B9311
Boric Acid	Sigma B0394
Bromophenol Blue	Sigma B0126
BSA (Bovine serum albumin)	Sigma A2153
Chloroform	Fisher C/4920/17
Deoxynucleoside triphosphate set (dNTPs)	ThermoFisher R0181
DMEM (Dulbecco's Modified Eagle's Medium)	Sigma D6546
DNase I	Roche 04716728001
Doxycycline hyclate	Sigma D9891
DTT (1,4-dithiothreitol)	Sigma 000000010197777001
ECL (Enhanced chemiluminescence) kit	Perkin Elmer NEL104001EA
EDTA	Sigma ED
Ethanol	VWR 20831.330

Ethidium bromide	Sigma E1510
FCS (Foetal calf serum)	Sigma F7524
Formaldehyde (37%)	Sigma F8775
Formamide	Sigma F9037
Glycerol	Fisher Scientific 10579570
Glycine	Sigma G8898
Hybond-N ⁺ Nylon Membrane	GE Healthcare RPN203B
Hydrochloric acid	Sigma 320331
L-Glutamine	Sigma G7513
Leupeptin	Sigma L2884
Lithium chloride	Sigma 62476
Magnesium chloride	Sigma M8266
MOPS (4-Morpholinepropanesulfonic acid)	Sigma M1254
MyOne Streptavidin C1 Beads	ThermoFisher 65001
NEBNext® PolyA mRNA Magnetic Isolation Module	NEB E7490
PBS (Phosphate-buffered saline)	Sigma D8537
Penicillin-streptomycin	Sigma P4333
Pepstatin A	Sigma P5318
Phosphatase inhibitor cocktail	Sigma P2850
Phusion high-fidelity polymerase	NEB
PIPES (1,4-Piperazinediethanesulfonic acid)	Sigma 80635
PMSF (Phenylmethanesulfonyl fluoride)	ThermoFisher 36978
Poly-A Tract Kit	Promega Z5300
PolyAPurist™ MAG kit	Ambion AM1922
Potassium Chloride	Sigma P9541
Protease inhibitor cocktail	Sigma P8340
Protein A agarose beads pre-blocked with salmon sperm DNA	Millipore 16-157
Proteinase K	Sigma P8044
PVDF membrane	ThermoFisher 88518
QuantSeq 3'mRNA-seq library kit	Lexogen 012.24
RevertAid Reverse Transcriptase	ThermoFisher EP0441
RNA fragmentation kit	Ambion AM8740
RNase A	ThermoFisher EN0531
RNase H	NEB M0297
RNase OUT	ThermoFisher 10777019
RNeasy kit	Qiagen 74104
SDS (sodium dodecyl sulphate)	Sigma L4390
SensiMix SYBR No-ROX	Bioline QT650
Sodium acetate	Sigma S2889
Sodium bicarbonate	Sigma S6014
Sodium chloride	Sigma S9625
Sodium citrate	MerckMillipore 567446
Sodium deoxycholate	VWR 430353P
Sodium phosphate dibasic	Sigma S9763
Superscript III	ThermoFisher 18080044
T4 kinase	NEB M0201
T4 RNA ligase I	NEB M0437
Taq polymerase	NEB M0267

Tris (Tris(hydroxymethyl)aminomethane)	Sigma 252859
Triton X-100	Sigma X100
Trizol	ThermoFisher 15596-026
Truncated T4 RNA ligase II	Bioo Scientific 512105
Trypsin	Sigma T3924
Tween-40	Sigma P1504
Xylene cyanol	Sigma X4126
Hybond-P	GE Healthcare RPN303F

2.3. ANTIBODIES

All primary antibodies used in this study are listed below.

Table 8 - Primary antibodies used in this study

Antibody Target and Name	Manufacturer and Catalogue Number
DICER1 - 13d6	Abcam ab14601
DICER1 – A-2	Santa Cruz sc-136981
EHMT2/G9a	Abcam ab40542
H3K9me2	Abcam ab1220
Protein A - polyclonal	Abcam ab7243
RNAPII - N-20 X	Santa Cruz, sc-899 X
TAP - IgG Sepharose 6 Fast Flow beads	GE Healthcare 17-0969-01
α -Tubulin - YL1/2	Abcam ab6160

2.4. BUFFERS AND COMPOSITIONS

All buffers and solutions were prepared in MilliQ water and pH was adjusted with 5M NaOH or 40% HCl.

Table 9 - Composition of buffers used in this study

Buffer Name	Composition
Binding buffer	10 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA
Cell lysis buffer	5 mM PIPES, pH 8.0, 85 mM KCl, 0.5% (v/v) NP-40, 1 mM PMSF, 1 µg/ml pepstatin A, 1 µg/ml leupeptin and 1 x protease inhibitor cocktail
Church Buffer	1% BSA, 1 mM EDTA, 7% (w/v) SDS, 0.5 M NaHPO ₄
Crosslink reversal buffer	0.3 M NaCl, 3 µg/ml RNase A, 10 µl of 0.5M EDTA, 20 µl of 1 M Tris-HCl, pH 6.5
Detergent Stock Solution	3.3% (w/v) sodium deoxycholate, 6.6% (v/v) Tween 40
DNA loading dye (5 x)	0.25% (w/v) bromophenol blue, 0.25% (w/v) xylene cyanol, 0.1 M EDTA, 30% (v/v) glycerol
High-salt wash buffer	0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl, pH 8.0, and 500 mM NaCl
Immunoprecipitation dilution buffer	0.01% (w/v) SDS, 1.1% (v/v) Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl, pH 8.1, 167 mM NaCl, 0.5 mM PMSF, 0.8 µg/ml pepstatin A, 1 µg/ml leupeptin and 1 x protease inhibitor cocktail
Immunoprecipitation elution buffer	1% SDS (w/v) and 0.1 M NaHCO ₃
LiCl wash buffer	0.25 M LiCl, 1% (v/v) NP-40, 1% (w/v) sodium deoxycholate, 1 mM EDTA and 10 mM Tris-HCl, pH 8.0
Low salt buffer	10 mM Tris-HCl pH7.5, 1 mM NaCl, 1 mM EDTA, 10% (v/v) formamide
Low-salt wash buffer	0.1% (w/v) SDS, 1% (v/v) Triton X-100, 2 mM EDTA, 20 mM Tris-HCl, pH 8.0, and 150 mM NaCl
Lysis Buffer B	10 mM Tris-HCl (pH 8-8.4) 0.14 M NaCl, 1.5 mM MgCl ₂ , 0.5% (v/v) NP-40
MOPS running buffer (1 x)	40mM Sodium acetate, 5 mM EDTA pH 8, 0.1 M MOPS pH 7
Nuclear Lysis Buffer	1% (w/v) SDS, 10 mM EDTA, 50 mM Tris-HCl, pH 8.0, 0.5 mM PMSF, 0.8 µg/ml pepstatin A, 1 µg/ml leupeptin and 1 x protease inhibitor cocktail
SDS Loading Buffer (4 x)	200 mM Tris-Cl (pH 6.8), 400 mM DTT, 8% (w/v) SDS, 0.4% bromophenol blue (w/v), 40% (v/v) glycerol
SSC Solution (20 x)	3 M NaCl, 0.3 M Sodium Citrate
TBE Buffer	89 mM Tris, 89 mM Boric acid, 2 mM EDTA
TE Buffer	10 mM Tris-Cl (pH 7.5), 1 mM EDTA
Towbin Buffer (1 x)	25 mM Tris, 192 mM glycine, 15% (v/v) methanol (pH 8.3)

2.5. CELL CULTURE

Inducible DICER1-knockdown HEK293 (2B cells) (Schmitter et al., 2006), HBL (Sánchez-Laorden et al., 2009), SW480 and SW620 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) containing 10% (v/v) foetal calf serum (FCS), 2 mM glutamine and ampicillin (100 mg/ml). Cells were grown to 60-80% confluence before harvesting or passaging unless specified otherwise. 1CT cells were cultured as described previously (Roig et al., 2010).

2.6. SUBCELLULAR FRACTIONATION

The subcellular fractionation protocol used was adapted from previous protocols (Topisirovic et al. 2003). Around 5×10^7 cells were trypsinized, washed two times with ice-cold PBS and the cell pellet was then resuspended by slow pipetting in 1 ml of Lysis Buffer B (Table 9). If required, 100 μ l of lysate was removed as the whole-cell RNA fraction. The nuclei were then spun down at 1000 g for 4 minutes. The supernatant was removed, which comprised the cytoplasmic RNA fraction and spun at 11,000 g for 1 minute to remove any remaining nuclei. The nuclei pellet was resuspended in 1 ml of Lysis Buffer B, and 100 μ l of the Detergent Stock Solution (Table 9) was then added under slow vortexing. The stripped nuclei were then spun down at 1000 g for 4 minutes at 4°C. The nuclei pellet was then washed once more in 1 ml Lysis Buffer B and spun down at 1000g for 4 minutes at 4°C. The nuclei pellet was then resuspended in 1 ml of Trizol by passing it through a 21-gauge syringe. Trizol was also used to extract RNA from the cytoplasmic and whole-cell fractions as per manufacturer's instructions. The RNA pellets were then DNase I-treated as per manufacturer's instructions. RNA was then phenol-chloroform extracted and precipitated with ethanol and the pellet was resuspended in RNase-free water. Concentrations were subsequently determined using a NanoDrop ND1000. For experiments where solely whole-

cell RNA was being extracted, the hot phenol method was used, as has been previously described (Furger et al., 2001). Extracted RNA was DNase I-treated and processed as in the subcellular fractionation approach.

2.7. REVERSE-TRANSCRIPTION PCR

One microgram of DNase-treated RNA was reverse transcribed into cDNA using RevertAid reverse transcriptase as per manufacturer's instructions with an oligo(dT) primer unless otherwise stated. The components of the RT-PCR reactions are shown Table 10. For each RNA sample, a reverse transcriptase negative control was performed which included all components of the reverse transcription reaction, except the reverse transcriptase itself. This was to ensure products were not being amplified from any contaminating DNA. Primers were designed using Primer-Blast (<http://www.ncbi.nlm.nih.gov/tools/primer-blast/>) to ensure specificity for the intended amplicon. Primers were synthesised by MWG Eurofins. All oligos used in the study can be found in the Appendix Table 1 and 2. 20 µl of RT-PCR product was mixed with 5 µl of 5 x gel loading dye and run on a 1-2% agarose/TBE gel containing 0.5 µg/ml ethidium bromide (EtBr) alongside a DNA ladder. The agarose gels were run at 120 V in 1 X TBE buffer for ~ 60-90 minutes.

Table 10 - Components of RT-PCR reaction

Component	Volume (µl)
Taq Polymerase (NEB)	0.2
Forward primer (2.5 µM)	1.0
Reverse primer (2.5 µM)	1.0
dNTPs (2 mM for each of dATP, dCTP, dGTP, dTTP)	4.0
10 x PCR Buffer with MgCl ₂ (NEB)	2.5
RNase-free H ₂ O	15.3
cDNA from DNase-treated RNA (RT- in RT- control)	1

2.8. DENATURING AGAROSE RNA GEL ELECTROPHORESIS AND NORTHERN BLOTTING

For northern blotting, purified RNA was resolved on 1% (w/v) formaldehyde-agarose gels (Hazzalin et al., 1997). The gel was run at 65 V for 2 hours in 1 x formaldehyde gel running buffer (Table 9) and then blotted onto nylon membranes by capillary transfer overnight in 20 x SSC solution. ³²P-labelled probes were synthesised using asymmetric PCR from gel purified RT-PCR products and hybridised with the membranes overnight at 65°C in Church buffer (Table 9). Membranes were then washed twice in 1 x SSC, once in 0.2 x SSC and finally once in 0.1 x SSC, with each wash being performed at 65°C for 20 minutes. Quantitation was subsequently performed on a phosphoimager. polyA⁺ purification was performed using the Poly-A Tract Kit according to manufacturer's instructions.

2.9. CHROMATIN IMMUNOPRECIPITATION AND QUANTITATIVE RT-PCR

Non-induced and doxycycline-induced (3 µg/ml, 7 days) HEK293 cells (2B cells) were collected from 175 cm² flasks. Treatment with BIX 01294 trihydrochloride hydrate was 5 µM for 72 hours prior to ChIP. 37% formaldehyde was added directly to the tissue culture medium at a final concentration of 1% (v/v) and then incubated for 10 min at 37°C. Formaldehyde was inactivated by adding glycine to a final concentration of 0.125 M. Cells were washed twice with 5 ml ice-cold PBS containing protease and phosphatase inhibitors and then scraped into 15 ml tubes. Samples were centrifuged for 5 min at 1600 rpm and 4°C. Cells were gently resuspended in 500 µl of cell lysis buffer (Table 9) and incubated on ice for 10 min. Nuclei were collected by centrifugation for 5 min at 3000 rpm at 4°C and resuspended in 400 µl ice-cold nuclear lysis buffer (Table 9) and then incubated on ice for 10 min. Samples were sonicated to an average length of 300–500 bp, kept on ice (15 s sonication and 15 s rest) and spun for 10 min at 13,000 rpm and 4°C to remove cell debris. The supernatant was diluted by adding immunoprecipitation dilution buffer and aliquoted into

various immunoprecipitation samples. Specific antibodies (5 µg/100 µg chromatin) were added to samples and incubated overnight at 4°C on a rotating wheel. Immune complexes were pulled down at 4°C with 40 µl of protein A agarose beads pre-blocked with salmon sperm DNA, or IgG Sepharose 6 fast flow beads, and then washed once with low-salt wash buffer, and twice with each of high-salt wash buffer, LiCl wash buffer and TE buffer. Immune complexes were eluted with 500 µl immunoprecipitation elution buffer for 30 min on a rotating wheel. Reversal of cross-links was performed by adding crosslink reversal buffer and 2 µl of 10 mg/ml proteinase K and incubating at 65°C overnight. DNA was purified by phenol/chloroform extraction and recovered in 60 µl of H₂O. Samples were then analysed using the SensiMix SYBR No-ROX on a Rotor-Gene RG 3000 (Corbett Research). Oligos used to target the specific amplicons are shown in Appendix Table 2. RT-qPCR signals represent the average of at least three biological repeats expressed as a percentage of the input signal.

2.10. REVERSE-TRANSCRIPTION QUANTITATIVE PCR (RT-qPCR)

Non-induced and doxycycline-induced HEK293 cells (2B cells) were collected from 75 cm² flasks. Whole-cell RNA was isolated using TRIzol according to manufacturer's instructions. Whole-cell RNA was treated with DNase I for 90 minutes at 37°C. DNase I was inactivated for 15 minutes at 75°C. RNA quality was assessed by using denaturing 1.25% (w/v) formaldehyde agarose gel electrophoresis and detection of 18S ribosomal RNA with ethidium bromide under ultraviolet light. For detection of gene-specific transcripts, whole-cell RNA was reverse transcribed with SuperScript III reverse transcriptase according to manufacturer's instructions using gene-specific primers (Appendix Table 1). To measure gene-specific antisense transcripts by RT-qPCR cDNA synthesis was performed on cDNA generated using sequence-specific forward primers complementary to the antisense RNA.

Gene-specific cDNA levels were then analysed using RT-qPCR and transcript levels were normalised to cDNA levels of the housekeeping genes *HPRT1* or *GAPDH*. RT-qPCR was performed using the SensiMix SYBR No-ROX on a Rotor-Gene RG 3000 (Corbett Research).

2.11. WESTERN BLOTTING

Non-induced and doxycycline-induced HEK293 (2B) cells were collected from 75 cm² flasks and lysed in 4 x SDS loading buffer. Lysates were sonicated, boiled, separated on pre-cast SDS polyacrylamide gel. Protein was then transferred to PVDF membrane in Towbin buffer overnight at a constant current of 200 mA at 4°C. The membrane was blocked with 5% milk and incubated with specific primary antibodies at 4°C overnight. After washing, membranes were incubated with horseradish peroxidase (HRP)-conjugated secondary antibodies for 1.5 hours at RT, washed and signals were detected using ECL kit.

2.12. 3'REGION EXTRACTION AND DEEP SEQUENCING (3'READS)

The 3'READS protocol was originally described in (Hoque et al. 2013) and is outlined in Figure 3. Briefly, 25-30 µg of RNA was subjected to 1 round of polyA selection using the PolyAPuristTM MAG kit according to the manufacturer's protocol, followed by fragmentation using Ambion's RNA fragmentation kit at 70°C for 5 min. PolyA-containing RNA fragments were isolated using the CU₅T₄₅ oligo (a chimeric oligo containing five uridines and 45 thymidines) which were bound to the MyOne streptavidin C1 beads through biotin at its 5' end. Binding of RNA with CU₅T₄₅ oligo-coated beads was carried out at room temperature for 1 hr in 1x binding buffer, followed by washing with low salt buffer. RNA bound to the CU₅T₄₅ oligo was digested with RNase H (5 U in 50 µl reaction volume) at 37°C for 1 hr, which also eluted RNA from the beads. Eluted

RNA fragments were purified by phenol:chloroform extraction and ethanol precipitation, followed by phosphorylation of the 5' end with T4 kinase. Phosphorylated RNA was then purified by the RNeasy kit and was sequentially ligated to a 5'-adenylated 3'-adapter (5'-rApp/NNNNGATCGTCGGACTGTAGAACTCTGAAC/3ddC) with the truncated T4 RNA ligase II and to a 5' adapter (5'-GUUCAGAGUUCUACAGUCCGACGAUC) with T4 RNA ligase I. The resultant RNA was reverse-transcribed to cDNA with Superscript III, and the cDNA was amplified by 12 cycles of PCR with Phusion high fidelity polymerase. cDNA libraries were sequenced on an Illumina HiSeq 2500.

2.13. BIOINFORMATIC ANALYSIS OF 3'READS DATA

An overview of the bioinformatics pipeline used to analyse the 3'READS data is shown in Figure 4, and was developed by the Tian lab (Hoque et al., 2013). Raw reads were mapped to the human genome (hg19) with bowtie2 using the option “-M 8 --local”. This mode allows the ends of the sequence not to be part of the alignment (soft clipped). This is useful for reads containing a sequence corresponding to the polyA tail which is not present in the genomic sequence. Reads that were shorter than 15 nt, were non-uniquely mapped to the genome (MAPQ<10), or contained more than 2 mismatches in the alignment were discarded. If reads contained ≥ 2 non-genomically encoded thymidines at the 5' end after the last aligned nucleotide, they were selected as polyA site supporting (PASS) reads, as these non-genomically encoded thymidines in all likelihood derive from the polyA tail. An individual pA site was assigned if it contained more than 2 reads in any one sample, or one read in any two samples and constituted more than 5% of all reads mapped to that particular gene.

As the exact point of cleavage for a single pA site can vary within a small window (Tian et al., 2005), cleavage sites located within 24 nt as mapped by PASS reads from all

samples were clustered. The site with the most PASS reads was selected to represent the exact coordinate for the pA site cluster. All the reads in a cluster were used to calculate isoform abundance. pA sites were mapped to RefSeq transcripts. pA sites located downstream of RefSeq transcripts were linked to the transcript by cDNA, EST and directional paired-end RNA-seq data from the ENCODE project. An extended region must be covered by cDNA/EST sequences or RNA-seq reads without a gap greater than 40 nt. Furthermore, the 3'UTR extension must not exceed the transcriptional start site of any other gene on the same strand. The relative abundance of an APA isoform in a sample was defined as the proportion of PASS reads supporting that APA isoform compared to all the PASS reads mapped to the gene. To remove data noise, pA sites with a relative abundance of less than 5% in all of the samples were discarded. The combined reads for all pA sites mapping to a single gene were used to calculate the gene expression values, which were then normalised to reads per million (rpm) PASS reads in each sample.

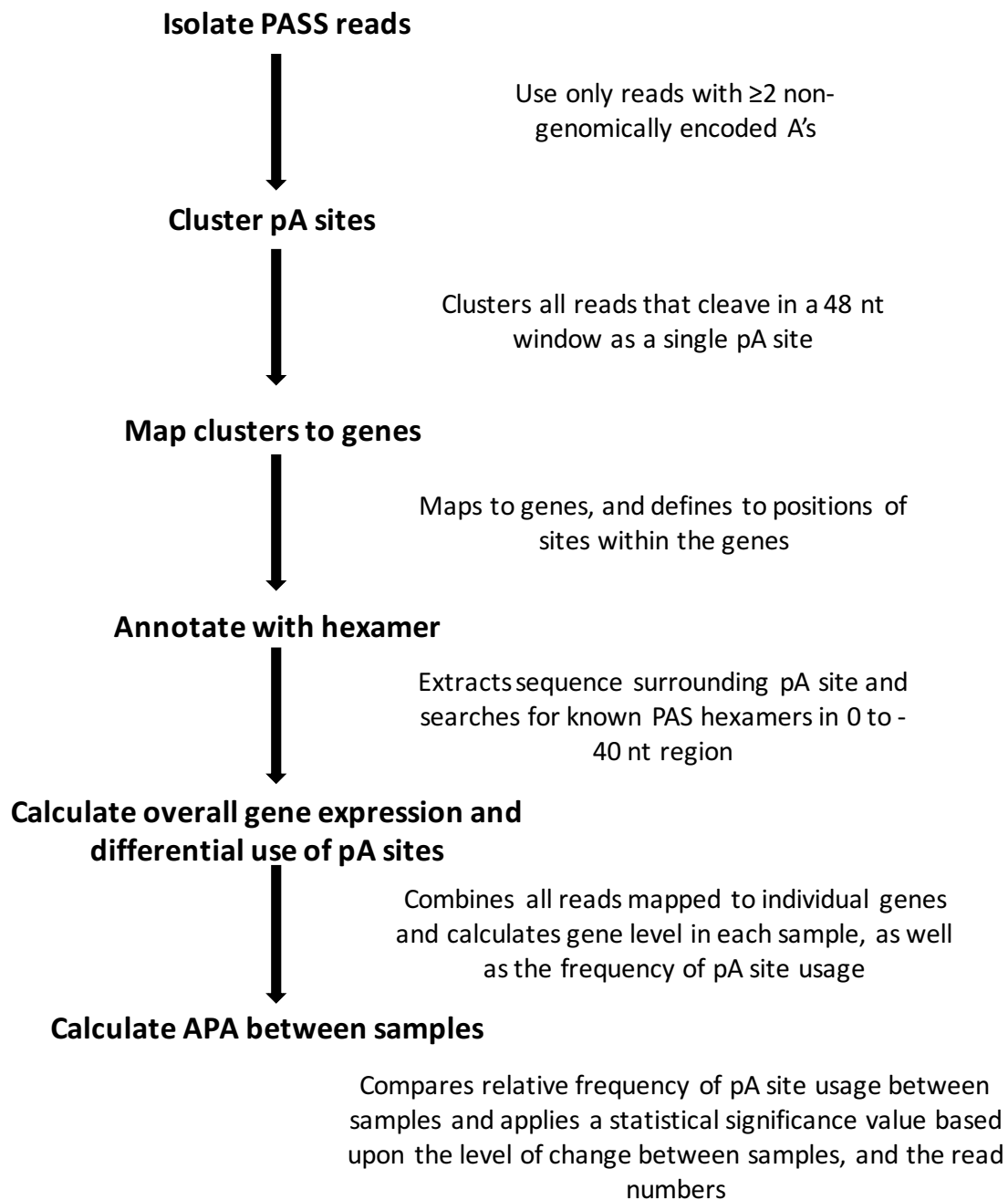


Figure 4: Schematic of 3'READS bioinformatics analysis pipeline

2.14. SIGNIFICANCE ANALYSIS OF ALTERNATIVE POLYADENYLATION

To study the difference of cytoplasmic to nuclear ratio between pA isoforms in genes with UTR-APA, significance analysis of alternative polyadenylation (SAAP) was used as previously described (Li et al., 2015). Briefly, total PASS read numbers per sample were

bootstrapped to 1 million. This allowed control for the impact of sequencing depth. For each gene, the two most expressed UTR-APA isoforms were picked for study. The observed change in cytoplasmic to nuclear ratios ($\Delta C/N$ ratio) was calculated between the two pA sites. To get the expected $\Delta C/N$ ratio, we bootstrapped reads for the two pA sites in cytoplasmic and nuclear fractions by assuming that the rpm ratio between cytoplasmic and nuclear fractions follows the averaged ratio for the two pA sites. We bootstrapped the reads and calculated expected $\Delta C/N$ ratio 100 times for each gene. Based on the 100 expected values, we converted all observed and expected $\Delta C/N$ ratio to Z scores. For the whole data set, we selected genes with significant C/N ratio difference between pAs by choosing observed absolute Z scores greater than a cut-off (Z_{cut}). The False Discovery Rate (FDR) was estimated by calculating the fraction of the average number of genes with $|\text{expected Z scores}| > Z_{cut}$ among the number of genes with $|\text{observed Z scores}| > Z_{cut}$. The FDR were controlled at less than 10% for this study. The whole SAAP method was performed 20 times for each comparison.

2.15. QUANT-SEQ APA ANALYSIS

The Quant-Seq 3'mRNA-Seq kit was also used in this project for deep sequencing of solely the 3'end. The manufacturer's protocol was followed, using 500 ng of input RNA for each sample and 13 cycles of PCR. An overview of the library preparation protocol is summarised in Figure 5. The resulting library was then loaded onto the Ion Chef system for template preparation and then the prepared chip was then sequenced Ion Proton Sequencing platform as per manufacturer's instructions.

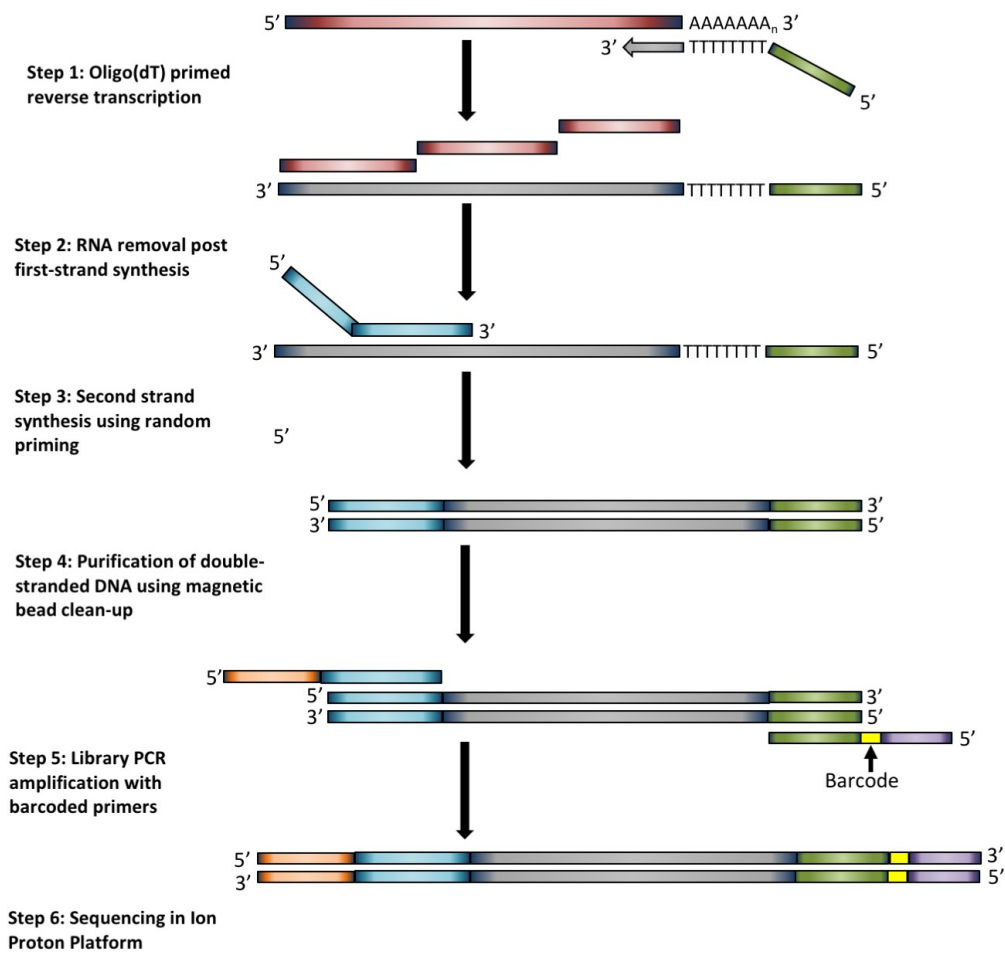


Figure 5: QuantSeq 3'mRNA sequencing library preparation

Sequences were aligned using the Ion Torrent Server TMAP aligner. Reads mapping to true pA sites were then extracted by using the Bedtools Intersect function to select only reads that fell within 100 nucleotides upstream of a previously confirmed pA site. These pA sites were those identified using 3'READS in HEK293 or HBL cells, or those identified using PolyA-Seq in several different human tissues under accession number GSE30198 (Derti et al., 2012). The resulting selected reads were then used as the input into the same scripts used for 3'READS without the requirement for non-genomically encoded adenosines to be classified as polyA supporting reads.

2.16. DYNAMIC ANALYSIS OF APA FROM RNA-SEQ (DAPARS)

DaPars analysis was performed as described in the original paper (Masamha et al. 2014) using the code available at <https://code.google.com/p/dapars> with default settings unless stated otherwise. For APA isoforms to be considered to be differentially represented between samples, the Δ PDUI used was 0.25. This means that the ratio of proximal to distal APA isoform had to differ by a minimum of 25%. Minor changes included the addition of a filter to only consider events whereby the predicted proximal pA site was located within 250 bp of an experimentally confirmed pA site. The pA site coordinates were obtained from APASdb (<http://mosas.sysu.edu.cn/utr/>) which involved 3'end mapping in 22 separate human tissues.

2.17. MOTIF ANALYSIS

Alternative UTR sequences were extracted, and the Discriminative Regular Expression Motif Elicitation (DREME) motif analysis tool (Bailey, 2011) was used with default parameters to identify overrepresented motifs. The reference sequences used were aUTRs from examples which showed either no difference between the cytoplasmic and nuclear subcellular fractions, or the opposite trend to those being analysed. Subsequently, the motifs identified were submitted to the TOMTOM motif comparison tool (Gupta et al., 2007) to identify any RNA-binding proteins that may be binding these motifs. The motif database used for known RNA-binding proteins was the those gathered using the RNAcompete method (Ray et al., 2013).

2.18. GENE ONTOLOGY ANALYSIS

Gene ontology analysis was performed using the **P**rotein **A**nalysis **T**hrough **E**volutionary **R**elationships (PANTHER) Classification System (Mi et al., 2016). Gene lists were submitted and analysed using the statistical overrepresentation test with default settings looking for enrichment in “Biological Process” categories. Only categories which show significant enrichment ($p < 0.05$), and have more than five genes, are shown.

3. CHAPTER 3 – GLOBAL MAPPING OF **POLYADENYLATION SITES**

3.1. INTRODUCTION

The aim of this chapter is to take a step back and gather an overall perspective of the basic statistics surrounding cleavage and polyadenylation in human cells. This is with the aim of generally considering the APA landscape in human cells, and what evolutionary pathways may have resulted in the observed profiles.

Firstly, an efficient subcellular fractionation technique using HEK293 cells is developed and quality control steps are introduced to ensure clean subcellular fractions are obtained. 3'READS is then used to map the 3'ends of the polyadenylated transcripts from these subcellular RNA fractions. The 3'READS data generated is then analysed, looking at the frequency and positioning of pA sites and the *cis*-elements that form them.

3.2. RESULTS

3.2.1. *Establishing a clean subcellular fractionation protocol*

As this project aims to compare cytoplasmic and nuclear APA profiles, it is imperative that the subcellular RNA fractions be highly pure and there is as little cross contamination between fractions as possible. To this end, a subcellular RNA fractionation protocol is established, that is both highly efficient and reproducible. Furthermore, suitable control markers are identified to measure the purity of the individual subcellular fractions before sequencing. As this study is looking at fractionated RNA, it is essential that the markers be RNA themselves, as opposed to protein. Two separate markers were employed to confirm the purity of the individual fractions, precursor ribosomal RNAs (pre-rRNAs) and the nuclear-retained long non-coding RNA (lncRNA), *MALAT1* (Wilusz et al., 2008).

In humans, the 45S (13.7 kb) precursor ribosomal RNA (pre-rRNA) is processed, via the 32S (5.5 kb) pre-rRNA, into the 28S (5.0 kb), 18S (1.9 kb) and 5.8S (160 bases) rRNAs in the nucleus prior to export into the cytoplasm (Lafontaine, 2015). To observe the localisation of these pre-rRNA bands, 2 µg of RNA from each fraction was run on a denaturing 1% agarose gel (Figure 6A). The cytoplasmic fraction is completely clear of pre-rRNAs confirming a clean cytoplasmic fraction with minimal contamination with nuclear RNA. The large abundance of the pre-rRNAs in the nuclear fraction, relative to the whole-cell fraction, shows a vast enrichment of nuclear RNA. The whole-cell fraction shows a small amount of pre-rRNAs, but with closer overall similarity to that of the cytoplasmic fraction. This is expected as, from cytoplasmic and nuclear RNA yields obtained during the harvests, it is estimated that in HEK293 cells the ratio of cytoplasmic to nuclear RNA is around 15:1.

In addition, reverse transcription PCR (RT-PCR) of *MALAT1* was used as a more sensitive method for verifying subcellular fractionation (Figure 6B). *MALAT1* is a highly expressed lncRNA, which localises almost entirely to the nuclear speckles (Wilusz et al., 2008). *GAPDH* was used as a reference gene as data from the ENCODE project shows that it is present in both the nucleus and the cytoplasm, but enriched in the cytoplasm relative to the average gene. This, therefore, allows a comparison of the *MALAT1:GAPDH* ratios in each subcellular fraction to estimate fractionation purity. The doublet seen in the *GAPDH* RT-PCR represents different splice isoforms of *GAPDH*.

Finally, to ensure that the RNA is of high quality, the RNA was run on an Agilent Bioanalyser 2100. Samples were only sent for sequencing if they had an RNA integrity number (RIN) of above 8.0. As sequencing is quantitative, accurate fractionation purity for each set of samples can be obtained once sequenced. This is shown in Figure 6C, where 3'READS maps the pA site of the nuclear-retained lncRNA *XIST* (Brown et al., 1992) with a level of ~5000 rpm in the nuclear fraction, whereas the cytoplasmic levels are ~3 rpm and

whole-cell ~30 rpm over this same region. Combined, these data show that this is a highly efficient fractionation, yielding RNA of high purity and quality.

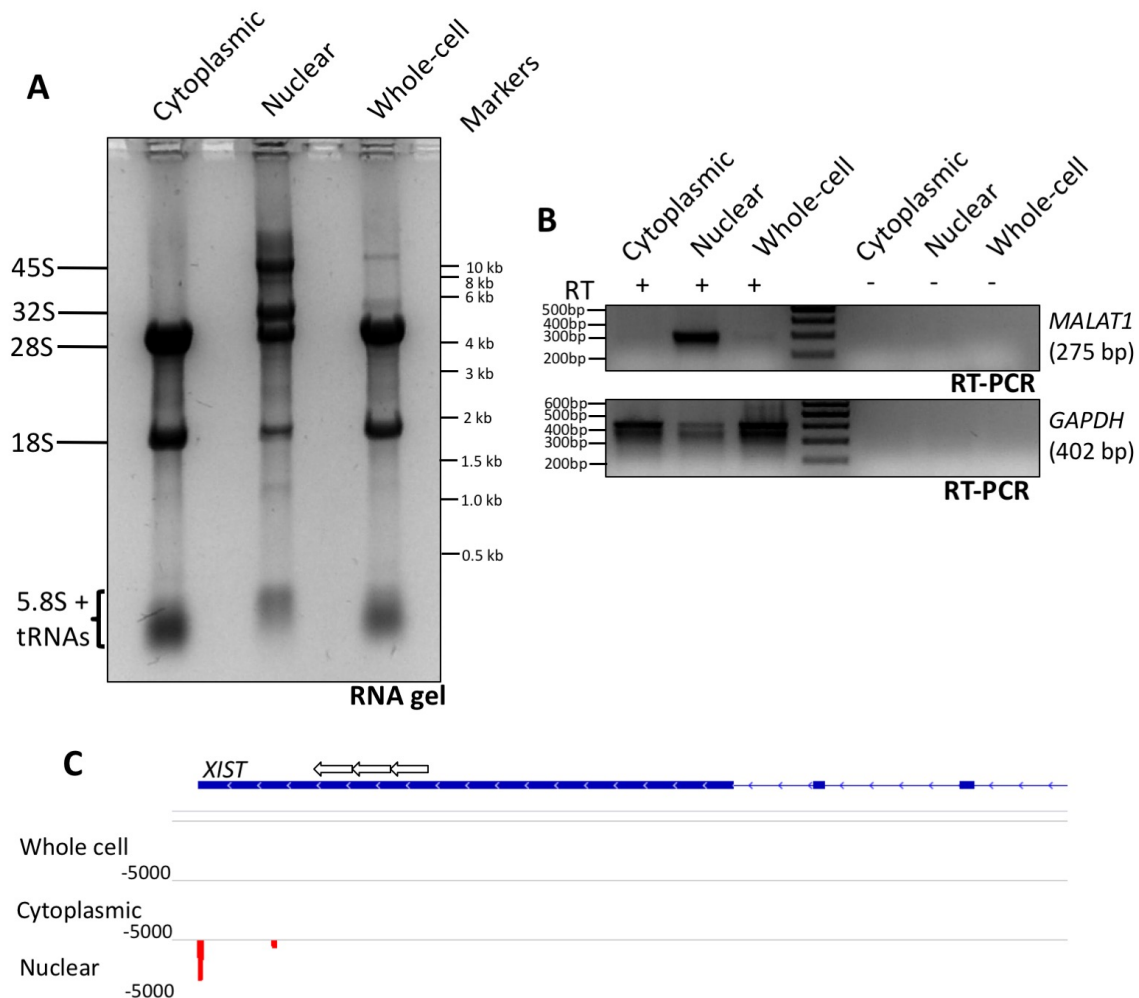


Figure 6: Subcellular Fractionation Controls for HEK293 cells

Panel A – Denaturing RNA gel with 2 ug of RNA from each subcellular RNA fraction. The identity of each ribosomal RNA band is shown on the left-hand side of the gel with size markers on the right-hand side of the gel.

Panel B – Reverse transcription PCR (RT-PCR) with primers targeting the nuclear-retained *MALAT1* (upper panel) and *GAPDH* in the lower panel. These were performed in each of the subcellular fractions using 1 ug of RNA from each subcellular fraction. Reverse transcriptase negative controls are on the right-hand side of each panel with the DNA ladder in the middle. The size of each band of the ladder is shown on the left-hand side, and the expected size of the targeted amplicon is shown on the right of each panel.

Panel C – Representative 3'-end targeted RNA sequencing density plot over the nuclear-retained *XIST* gene for the subcellular fractions of HEK293 cells. The reads levels are all normalised to reads per million. The direction of the *XIST* gene is shown by the arrows above the gene schematic.

3.2.2. *Reproducibility of 3'READS Results*

In the initial studies describing the development of the 3'READS technique, it was demonstrated that the method is highly reproducible (Hoque et al., 2013). The comparison of gene expression levels calculated by 3'READS across biological replicates resulted in Pearson correlations ranging from 0.95 to 0.98. A Pearson correlation of 0.89 was found when the gene expression levels calculated from 3'READS were then compared to those calculated from standard RNA-seq. Together this indicates that 3'READS is highly reproducible and can be employed to accurately quantitate gene expression levels (Hoque et al., 2013).

In collaboration with Prof. Bin Tian at Rutgers University, the 3'READS pipeline was used to assess the APA profiles of the HEK293 subcellular fractions from Section 3.2.1. Two biological replicates were performed, which were harvested several months apart. To measure the correlation between the different subcellular fractions the Pearson correlations of the gene expression values from each biological replicate were calculated. These ranged from 0.68 when comparing the nuclear fractions to 0.79 for the cytoplasmic fractions (Figure 7). These correlations are indeed lower than that found in the original paper (Hoque et al., 2014), despite the samples being processed through the same pipeline, and by the same people. This indicates that the lower correlations seen between these biological replicates are likely due to biological variation between the cell populations, possibly due to different passage numbers or fluctuations in culture conditions, most probably as a consequence of them being harvested several months apart. Despite this, combined with the previous studies (Hoque et al., 2014), it shows that 3'READS is suitable to accurately and quantitatively map the pA site usage within a given RNA sample.

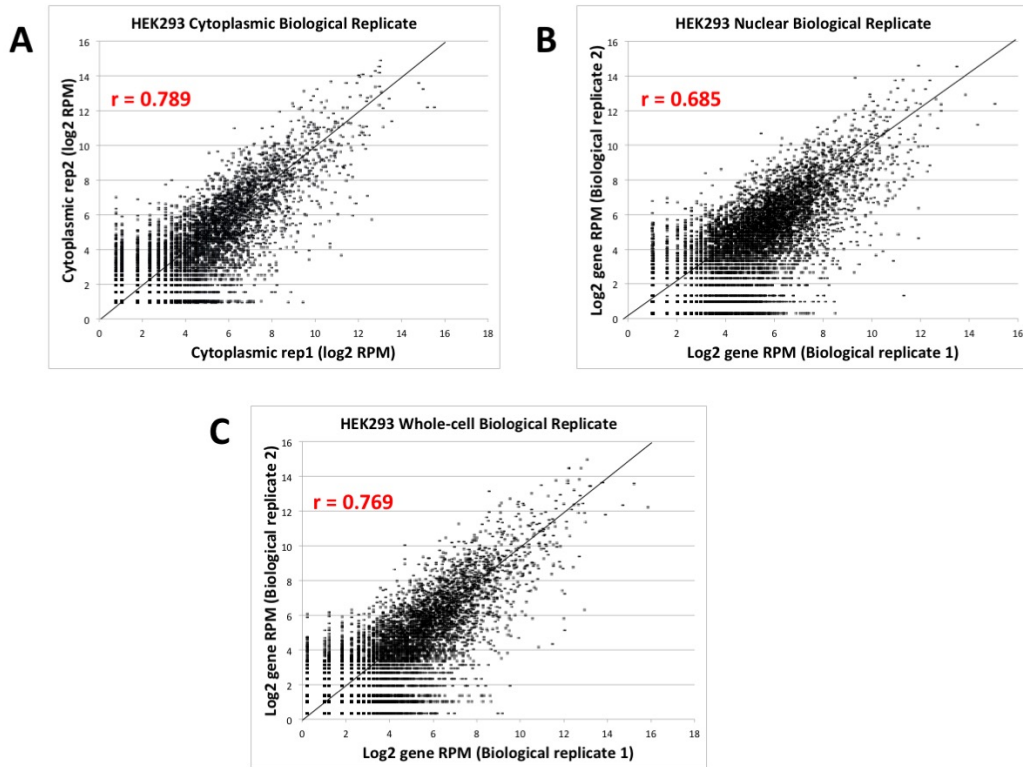


Figure 7: Correlation plots of 3'READS data on HEK293 subcellular fractions

The reads per million (RPM) value of each gene was calculated from the 3'READS data for each subcellular fraction and the $\log_2(\text{RPM})$ for each gene were plotted for each biological replicate. Panel A shows this for the cytoplasmic fraction biological replicates where Panel B shows this for the nuclear fractions and Panel C for the whole-cell fractions. The Pearson correlation for each comparison is shown in each plot.

3.2.3. *Classification of a polyadenylation site*

As described in Section 1.4.2, the *cis*-acting sequence elements that define a pA site can deviate significantly from the canonical pA site and still direct cleavage and polyadenylation. As a result of this, there are many sites throughout the genome that are termed “cryptic” pA sites and are used at very low levels, many of which may not have any physiological significance. Therefore, to minimise this background noise, these pA sites need to be filtered out bioinformatically.

In the 3'READS pipeline, a pA site is defined by a site that contains more than 2 PASS (Section 2.13) reads in a single sample or 1 PASS read in 2 or more samples.

Additionally, the reads associated with that pA site comprising more than 5% of all reads mapped to that specific gene in at least 1 of the samples being analysed. Using these parameters, 3'READS mapped 43,425 utilised pA sites corresponding to 14,632 individual genes in the two biological replicates of the HEK293 subcellular RNA fractions. All subsequent analysis is performed on the first biological replicate of HEK293 subcellular fractions unless stated otherwise.

3.2.4. *Location of pA sites throughout HEK293 cells*

The initial analysis looked at the general frequency and location of identified pA sites throughout the genome. As described in the Materials & Methods Section, each pA site was assigned to a gene. The specific location of each pA site within the gene was determined and classifications were made as described in Section 2.13 and this distribution is shown in Figure 8.

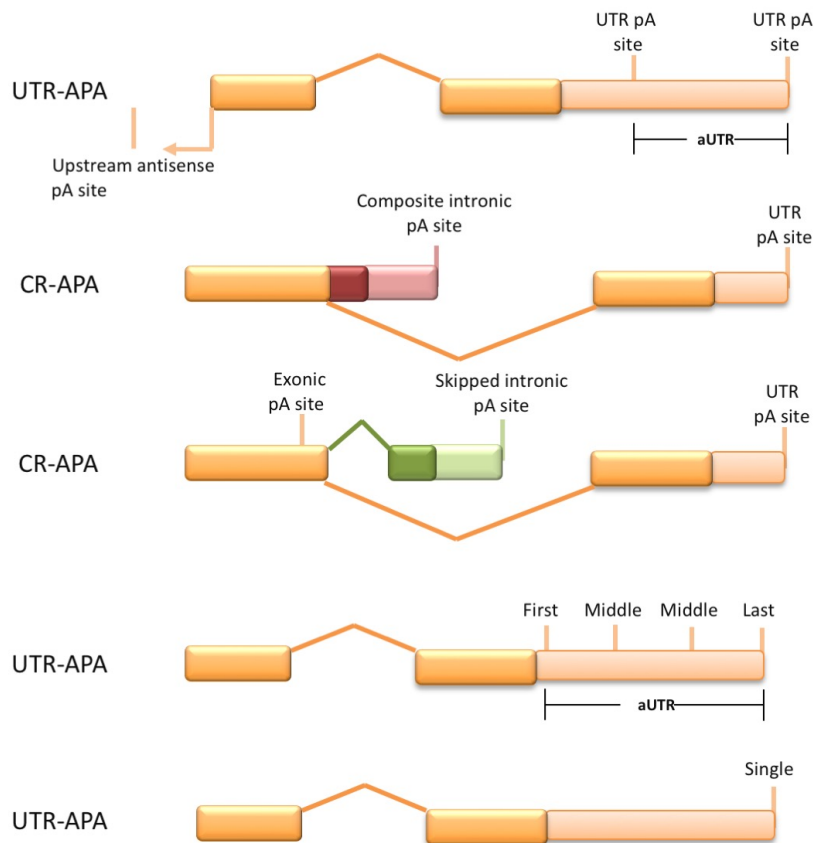


Figure 8: Classification of pA site positioning

The upper panel shows untranslated region APA (UTR-APA) where multiple pA sites occur within a single UTR with the alternative UTR (aUTR) between these pA sites. It also shows the classification of upstream antisense pA, which arises from the cleavage and polyadenylation of antisense transcripts upstream of the annotated gene promoter.

The second and third panels show coding region APA (CR-APA) where pA sites are utilised after a 5' splice site is missed (composite-intronic pA) or a pA site contained within an alternative exon (skipped-intronic pA).

The bottom two panels show the classification of UTR-APA sites depending on their relative positions to other UTR-APA sites. This includes whether it is the first, middle or last pA site in the UTR, or the only (single) pA site for that annotated UTR.

When looking at frequency alone, around 81% of all mapped pA sites occur within the annotated 3'UTRs of genes, with a significant number of pA sites also located in the intronic regions (Figure 9A). To take into account the number of transcripts that terminate at these different pA sites, each of these figures was adjusted to the read count at each pA site in the whole-cell HEK293 sample (Figure 9B). The same was repeated for nuclear and

cytoplasmic fractions and there was no significant difference in the whole-cell sample so these results are not shown. This analysis demonstrates that of all polyadenylated transcripts isolated from these HEK293 cells, over 95% of these were using pA sites located in the 3'UTR of the gene (Figure 9B). Therefore, despite the fact that a significant number of pA sites are located in the intronic regions, these are either lowly used pA sites or result in the production of unstable transcripts.

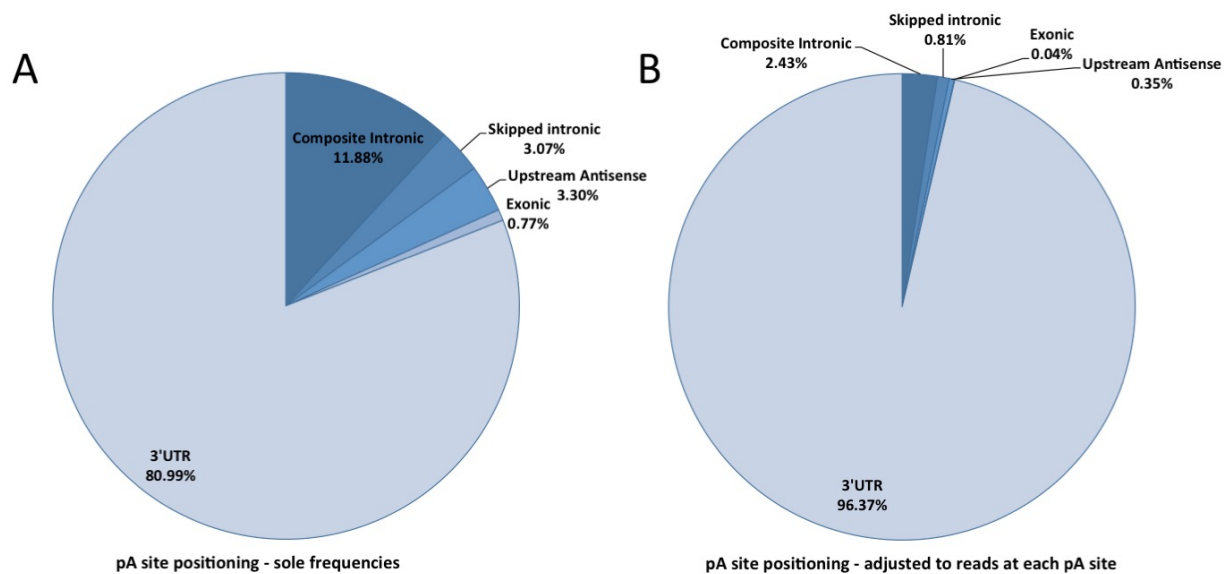


Figure 9: Positioning of pA sites in HEK293 cells

Panel A shows the positioning of pA sites relative to different annotated gene features as mapped by 3'READS in the whole-cell fraction of HEK293 cells. This is based on the sole frequency of pA sites mapping to each gene feature. Categorisation of pA site positioning is defined as in Figure 8.

Panel B shows the same as Panel A but with each pA site weighted to the number of reads mapping to that pA site in the whole-cell fraction of HEK293 cells.

3.2.5. *Number of pA sites per gene*

Previous large-scale studies using 3'end mapping techniques have found that ~70% of human genes contain multiple pA sites, with ~50% of genes having 3 or more (Derti et al., 2012). Here, mapping of pA sites in HEK293s using 3'READS yields similar numbers, with

~60% of mapped genes having multiple pA sites (Figure 10). The slightly higher figure obtained previously, likely reflects the broader range of samples analysed.

To determine the relative usage of these multiple pA sites in the gene they were classified as either the dominant or alternative site based on their abundance in the whole-cell fraction relative to other APA isoforms. When considering only genes with multiple pA sites, it was found that in the whole-cell fractions 82% of overall reads came from the dominant pA site indicating that the alternative pA sites are generally lowly used but potentially provide a site to switch to upon a change in growth conditions or external stimuli.

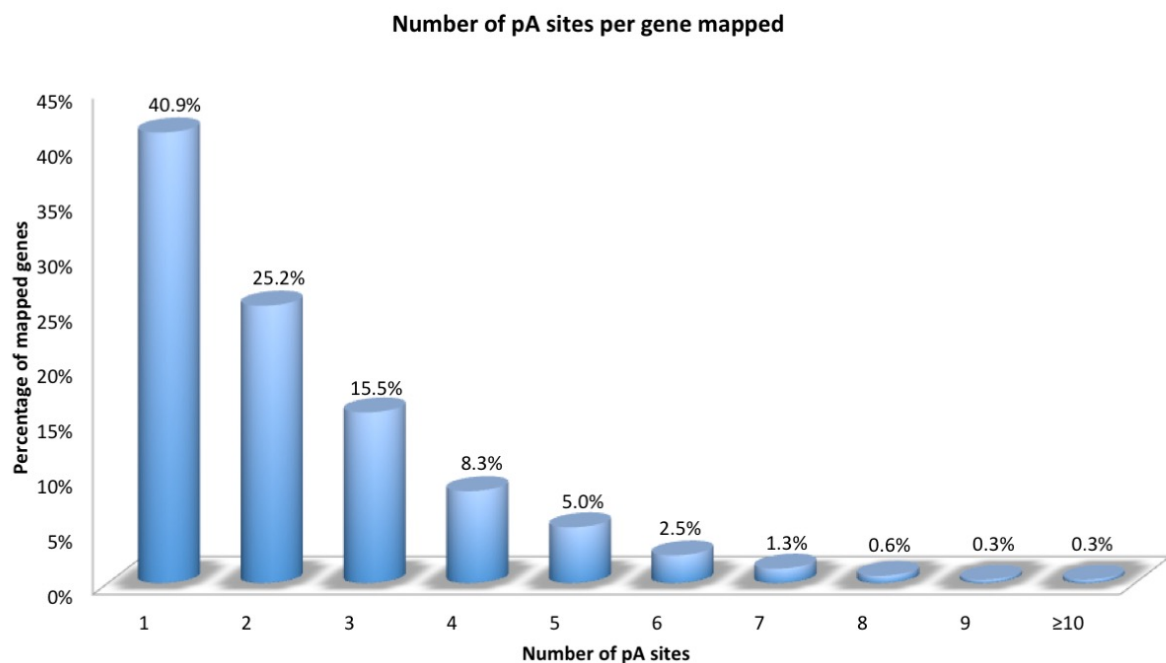


Figure 10: Number of pA sites per gene

The genes mapped using 3'READS on RNA from HEK293 cell are plotted with the number of pA sites per gene. pA sites were from all nuclear, cytoplasmic and whole-cell fractions combined.

3.2.6. *Polyadenylation Signal Hexamer usage*

One of the key factors that govern the strength of a pA site is the identity of the upstream PAS (Sheets et al., 1990). Previous studies have looked at the relative frequencies of the different hexamers at pA sites throughout the human genome (Beaudoing, 2000; Wang et al., 2013). Both of these studies found that around 58% of pA sites throughout the human genome contain the canonical AAUAAA sequence and around 15-20% AUUAAA hexamer.

Analysis of the hexamer frequency was also performed on the 3'READS data here from the HEK293 cells. Initially, the 0 nt to -40 nt regions directly upstream of the mapped pA site were extracted. The hexamer variants AAUAAA, AUUAAA, AGUAAA, UUUAAA, CAUAAA, UAUAAA, AAUACA, GAUAAA, ACUAAA, AAUAUA, AAUGAA and AAUAGA were then searched for in a hierarchical fashion in this order. Namely, if both the AAUAAA and CAUAAA hexamers were both present in the 0 to -40 nt region of a single pA site, that site would be annotated with the AAUAAA hexamer.

The order was determined by the relative efficiency that each of these hexamers can direct cleavage and polyadenylation from previous studies (Sheets et al., 1990). This was based on the premise that if multiple hexamers were present, the one that is known to direct cleavage and polyadenylation more efficiently is most likely to play the major role in recruiting CPSF to the pA site. A further classification of A-rich was added for if none of the aforementioned hexamers was found in the 40 nucleotides directly upstream of the pA site. To classify as an A-rich sequence there must be a string of five adenosines in the 40 nucleotides upstream of the pA site with a maximum of one other base separating this string of adenosines i.e. any of the following: AAAAA, A(C/U/G)AAAA, AA(C/U/G)AAA, AAA(C/U/G)AA or AAAA(C/U/G)A. This classification of pA site is based on previous studies that have shown that, when combined with a potent DSE, an A-rich stretch is sufficient to direct efficient cleavage and polyadenylation (Nunes et al., 2010). Any pA sites

which did not contain any of the described hexamers or an A-rich region were classified as “No PAS” pA sites.

Hexamer frequency observed from the 3'READS data here shows a similar pattern to the previous studies with 53% of sites using the canonical AAUAAA and 17% of sites using the AUUAAA hexamer. Around 70% of pA sites used in HEK293 cells contain either of the canonical hexamers (A(A/U)UAAA) leaving 30% that do not contain either of the canonical hexamers (Figure 11A). This is a significant number of pA sites (~13,000 sites), notably of which ~2700 fall under the “No PAS” classification.

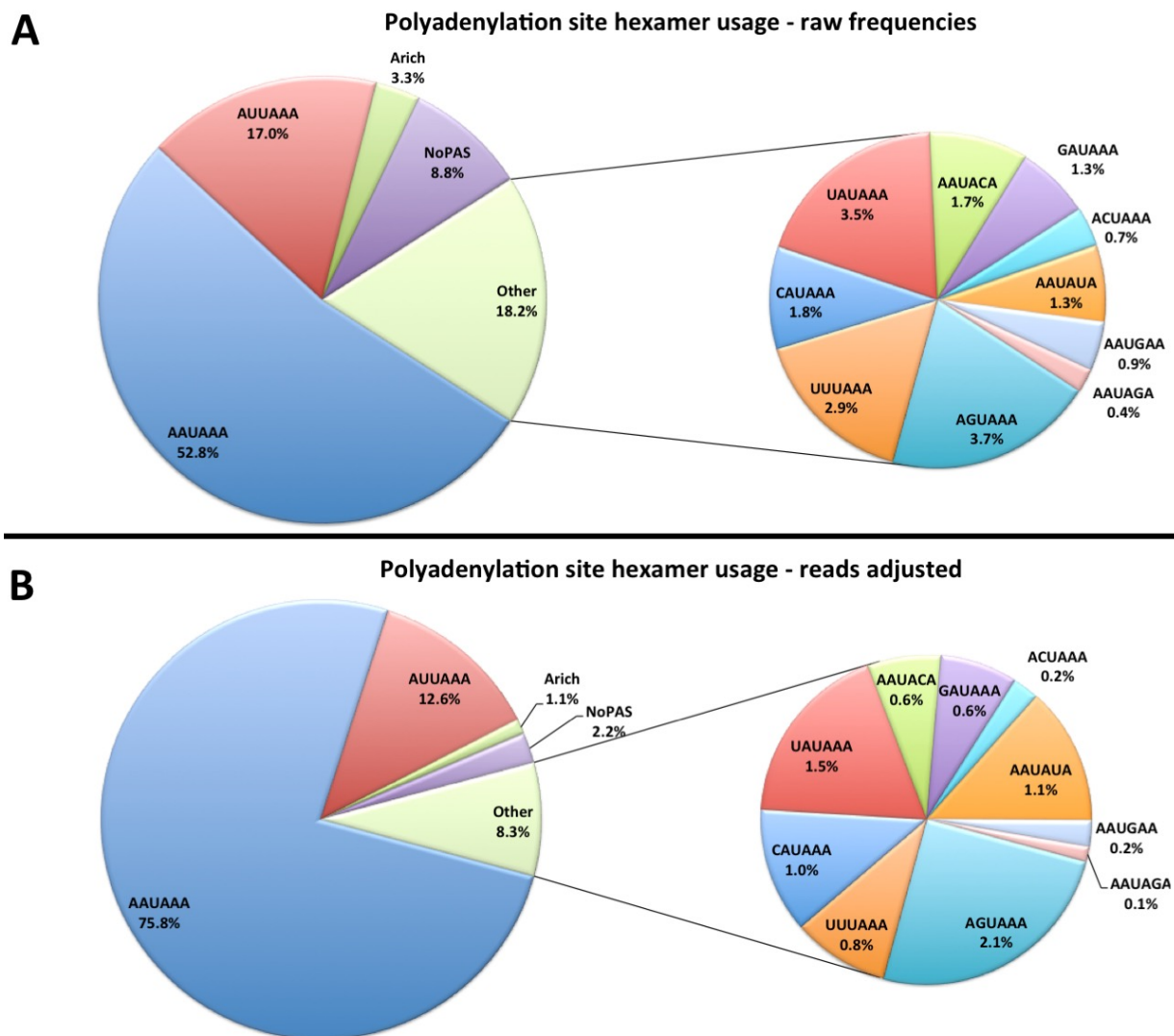


Figure 11: Frequency of polyadenylation signal usage

Panel A – Polyadenylation signal (PAS) frequencies in the 40-nt region upstream of all mapped pA sites in HEK293 cells. Here, only frequencies are taken into account, with all pA sites considered equal, regardless of read number.

Panel B – As in A, but with each PAS weighted to the number of reads mapping to the corresponding pA sites in the whole-cell fraction of HEK293 cells.

To quantify the actual number of transcripts coming from these weaker pA sites, each pA site was weighted based upon the number of reads at the specific pA site. The 3'READS data from RNA isolated from the whole-cell was used for this analysis. As can be seen in Figure 11B, when read number is taken into account the contribution of the canonical hexamers increases significantly and the weaker PASs reduce. This shows that in HEK293 cells, nearly 90% of all transcripts cleave and polyadenylate at pA sites that have an upstream canonical hexamer. Therefore, it appears that despite there being a large number of “No PAS” and alternative PAS pA sites these sites appear to be lowly used sites which may reflect their reduced efficiency at directing efficient cleavage and polyadenylation.

This is confirmed when looking at the average number of reads that are mapped to pA sites containing different PAS types (Figure 12). Here we see that pA sites using canonical hexamer AAUAAA have, on average, almost twice as many reads mapping to these sites compared to pA sites using other hexamers. The other PAS types generally decrease in the average reads mapping to these pA sites with the relative efficiency in which they can direct efficient cleavage and polyadenylation (Sheets et al., 1990).

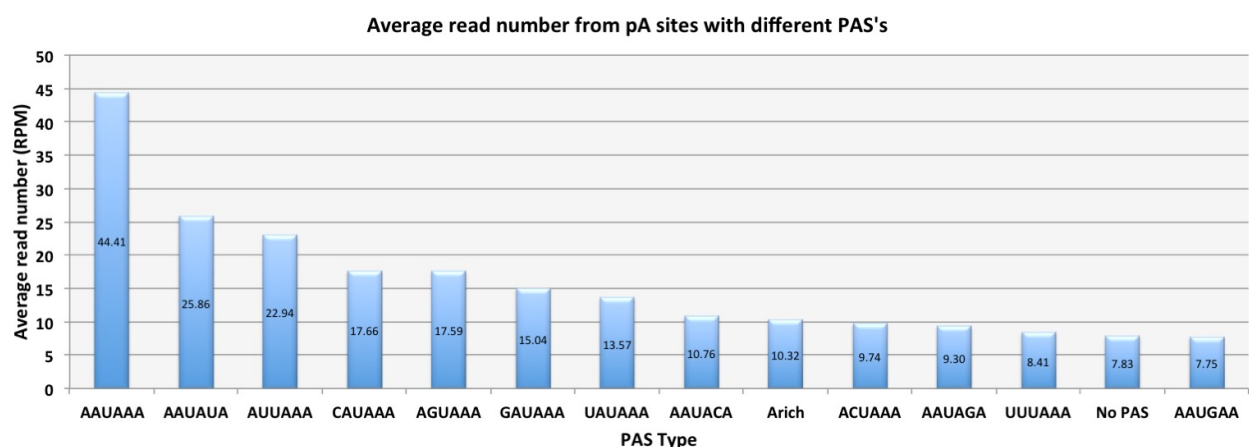


Figure 12: Correlation between read number and PAS usage

Bar chart showing the average number of reads mapping to pA sites containing each of the polyadenylation signals (PAS)

To see if this same pattern was conserved throughout different species, a similar analysis was done using publically available 3'end mapping data from mouse cells (specifically in cells involved in mouse thymopoiesis) (You et al., 2014). This data is derived from a slightly different mapping technique called SAPAS (Table 6). Here RNA from various stages of mouse thymopoiesis was isolated and subject to SAPAS and reads were subsequently mapped to mm9 and filtered for internal priming (You et al., 2014). If the number of reads throughout all samples was less than two that pA site was discarded. The pA sites were then annotated with an upstream PAS and, as before, the PAS frequency was analysed by sole frequency (Figure 13A), and also adjusted to read number at each pA site (Figure 13B).

Here, the canonical hexamers account for ~50% of PASs, but, as with in HEK293 cells, this increases greatly when adjusted for read number at the corresponding pA sites. This, therefore, shows that the higher read numbers at pA sites with more efficient PAS, are conserved between human and mouse.

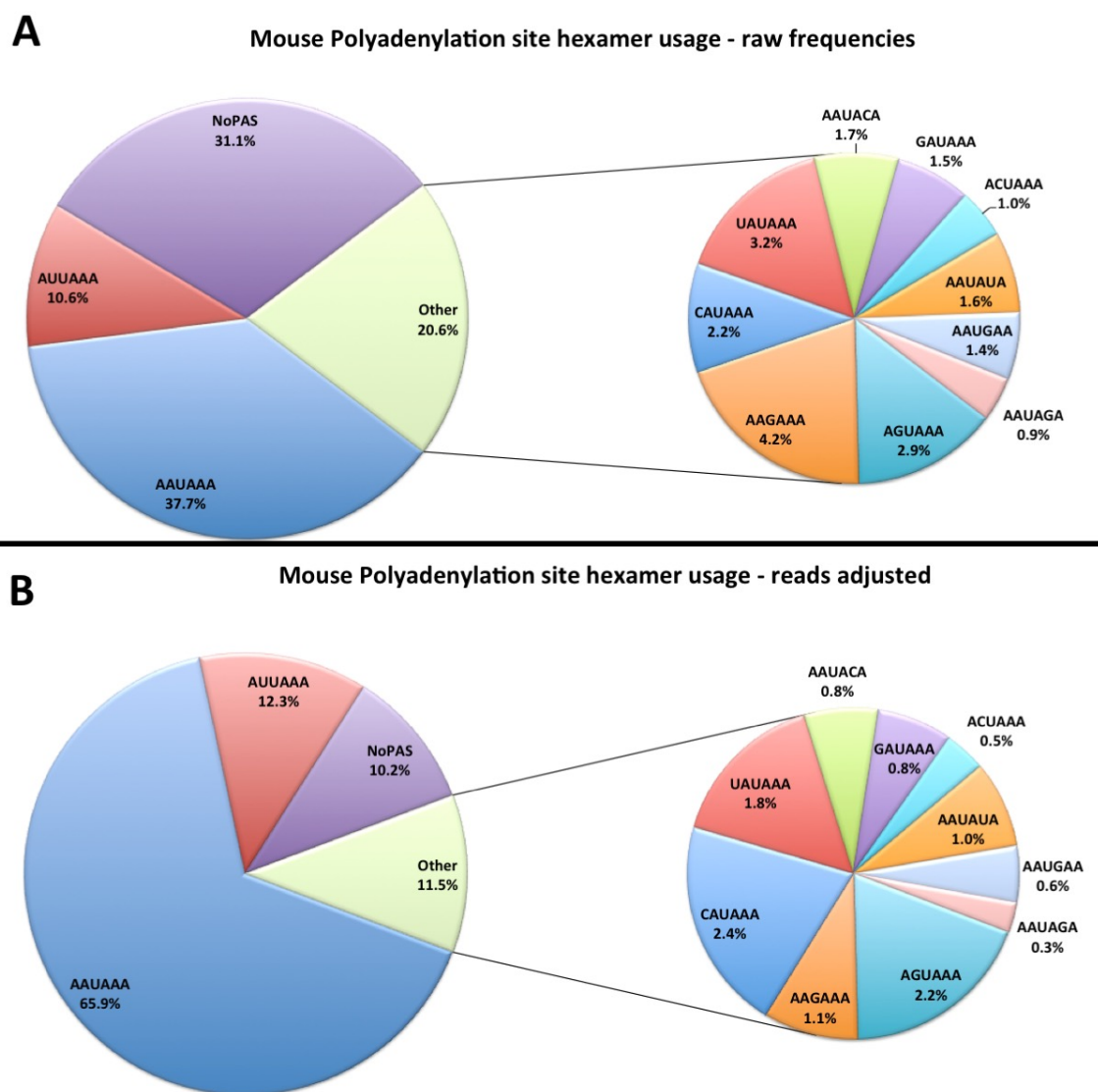


Figure 13: Frequency of polyadenylation signal usage in mouse

Panel A – Polyadenylation signal (PAS) frequencies in mouse cells as determined in You et al., 2014. Here, only frequencies are taken into account, with all pA sites considered equal, regardless of read number.

Panel B – As in A, but with each PAS weighted to the number of reads mapping to the corresponding pA sites in the mouse cells being analysed.

3.2.7. *Positional effect on hexamer usage*

The pA sites using different PASs appear to differ significantly in the reads associated with that pA site. To further investigate this the positional effect of the different pA sites was analysed. pA sites were classified depending on their positioning as in Figure 8, and the frequency of the different PAS types was analysed at each of the classes of pA site positioning (Figure 14). All hexamer variants were combined as “Other PAS”, as seen in the subset of the pie chart in Figure 11. As can be seen in Figure 14, in ~85% of cases when there is only a single pA site in the UTR of a gene (single), there is one of the two canonical (A(A/U)UAAA) hexamers. Similar frequencies are seen when looking at the last pA site when there are multiple pA sites present, with nearly 80% of these pA sites using the canonical hexamers. This is significantly higher than the frequencies seen for the canonical hexamers when considering pA sites that have an additional pA site downstream of them (first and middle) where the frequencies are ~58-61% (Figure 14). Exonic pA sites have an even lower incidence of canonical hexamers at ~55% while intronic pA sites lie around mid range at ~68%-77% canonical hexamers.

One quite striking example of the infrequency of using weaker pA sites as the sole pA site in a gene can be seen when only the genes that contain a No PAS pA sites are taken and then the number of pA site per gene is replotted (Figure 15). This shows that there are very few (316) genes that have a single No PAS pA site.

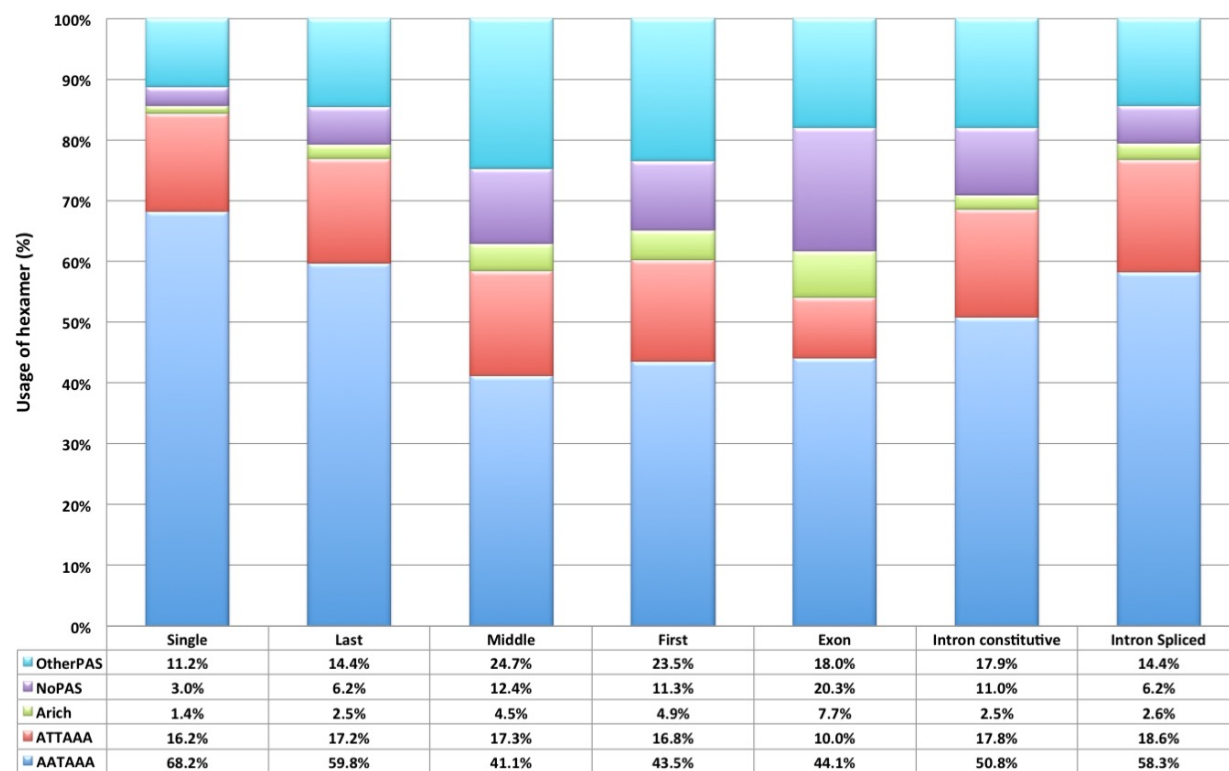


Figure 14: Positional dependency of polyadenylation signal frequencies

Stacked column bar chart which shows polyadenylation signal (PAS) frequencies for each different positional classification of pA site (shown in Figure 8) as mapped by 3'READS in HEK293s. The exact frequencies for each PAS are shown under the bar chart.

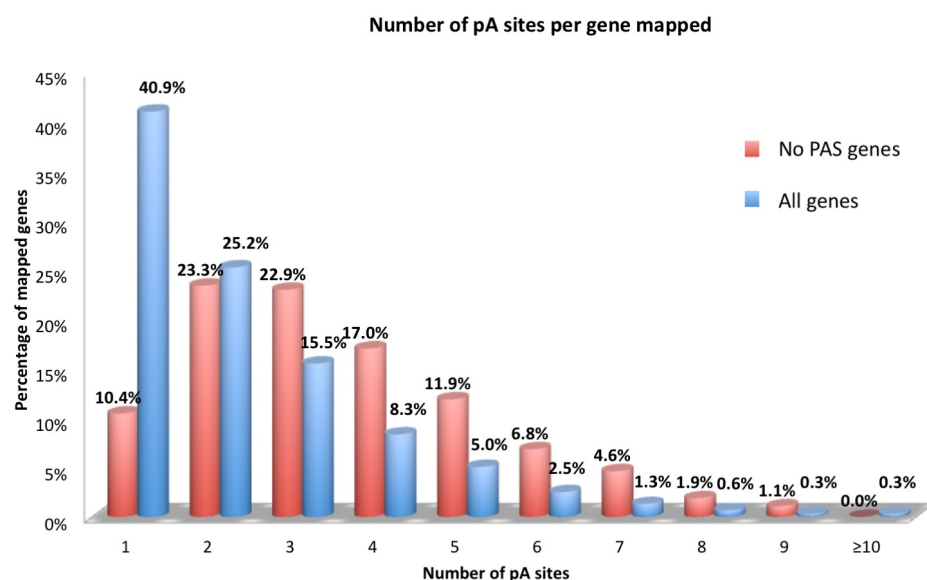


Figure 15: Distribution of No PAS pA sites

Bar chart showing the number of pA sites per gene in all genes (blue) and only genes containing a No PAS pA site (red) in HEK293 cells.

3.3. DISCUSSION

3.3.1. *Advantages and disadvantages of 3'READS technique*

To date, there have been a huge number of different methods developed to assess the polyadenylation landscape in a given RNA sample. To globally identify all pA sites used in a specific sample 3'READS has been demonstrated to be one of the most quantitative and most accurate, due to its ability to eradicate internal priming (Hoque et al., 2013). 3'READS was therefore chosen to analyse the pA landscape in the different subcellular fractions of HEK293 cells.

Despite these advantages, 3'READS does suffer from drawbacks that are seen throughout 3'end mapping techniques. Firstly, due to the stringent wash steps, the input level of RNA is relatively high, requiring more than 25 ug per sample. Since nuclear RNA is being extracted, a large cell population (~50 million cells) is required to achieve this quantity, thus restricting the approach to cells which can easily be grown on a large scale.

Furthermore, it also suffers from a drawback found with all 3'end sequencing approaches: There is no additional information about the identity of that transcript upstream of the 3' tag. Consequently, features including splice variants, overlapping genes and small RNAs that lie within genes can all create complications in the analysis of 3'READS data. Some studies have tried to address these issues by sequencing the full-length transcript (Pelechano et al., 2013; Sharon et al., 2013), but techniques that use this approach are generally found to lose their quantitative ability. However, as will be seen in later sections, our approach tackles this issue by using other methods, including standard RNA-seq and northern blotting, alongside 3'READS to characterise transcripts of interest.

3.3.2. *Potential reasons for having multiple pA sites per gene*

In line with previous studies (Derti et al., 2012) here it is found that the majority (~60%) of genes in HEK293s have multiple pA sites (Figure 10). This raises the question of why the genome has evolved such an excess of pA sites. There are a number of potential scenarios that would result in several pA sites at a single gene.

Firstly, if an additional pA site is located in the body of a gene (i.e. in an intron or the coding region), it is guaranteed to have an effect on the protein as it changes the actual open reading frame by either truncation of the protein or by including or excluding an alternative C-terminus. Here, the presence of multiple pA sites can, therefore, function to provide a mechanism to produce multiple protein isoforms, potentially with different functions, from the same gene.

Secondly, multiple pA sites in the 3'UTR can potentially have a regulatory impact. Depending on the identity of the sequence between the pA sites, in the aUTR, the resulting APA isoforms could potentially differ in their stability, localisation or translational efficiency. If this was the case it is easy to see how this could have a physiological role: you have multiple pA sites that produce transcripts that are differentially regulated and, in response to environmental changes or extracellular stimuli, you switch between these pA sites to ultimately change the level of corresponding protein produced to suit that environment. This has generally been accepted as the function of having multiple pA sites in the 3'UTR and has indeed been shown to be true for a significant number of genes (as discussed in Section 1.5). However, the pervasiveness of this mechanism has recently been questioned (Gruber et al., 2014; Neve et al., 2016; Neve and Furger, 2014; Spies et al., 2013).

There are also other biological pressures that may favour the presence of multiple pA sites in their UTRs. In yeast, to avoid transcriptional interference at downstream genes, it has previously been proposed that multiple pA sites may function as a fail-safe mechanism to ensure the transcript still undergoes efficient cleavage and polyadenylation, even if the first

pA site is missed (Yu and Volkert, 2013). It is easy to imagine how this fail-safe mechanism could apply to mammals as several chromosomal regions feature a high density of genes. Also, multiple successive pA sites may confer a slight evolutionary and energetic advantage as it would maximise the transcriptional output of the gene by reducing the likelihood of the production of unstable read-through transcripts. In this case, the identity of the aUTR is somewhat irrelevant, as the positive selection for multiple pA sites is not linked to the ability to differentially regulate the APA isoforms, but merely to the ability to ensure efficient transcriptional termination.

Since the sequence requirements to form a pA site are relatively relaxed, the probability of a pA site to arise through random evolutionary mutation is likely to be a fairly common event. Furthermore, because the UTR is non-coding, it is not under the same selective pressure as the coding region meaning that mutations are more likely to propagate and increase the likelihood of new pA sites arising. However, these pA sites are likely to be neutral regarding evolutionary advantage so there would be minimal conservation of these pA sites. An equilibrium would exist where existing pA sites would disappear and new ones arise, solely based on the probability of any random sequence being capable of directing cleavage and polyadenylation. In evolutionary terms, this would lead to a highly dynamic APA landscape, with little conservation seen between species, due to the constant introduction and removal of pA sites. This lack of conservation is seen to a large extent (Ara et al., 2006), as discussed in Section 1.6.

3.3.3. *Distribution of hexamer usage*

Analysis of hexamer usage from our 3'READS data yielded similar results to previous studies (Beaudoing, 2000; Wang et al., 2013). However, previous studies simply analysed frequencies of PASs with all pA sites being treated as equal. This approach, though

informative, does mask much information, as giving the primary pA site in *GAPDH* the same weighting as that of a pA site that only has two reads mapping to it is misrepresentative. Weighting each pA site to the number of reads mapped to it (Figure 11B) essentially treats all **transcripts** individually, regardless of whether they use the same pA site.

From this analysis, it is found that over 80% of transcripts in the mouse and human cell lines analysed contain one of the two canonical hexamers in the 40 nucleotides immediately upstream of the utilised pA site. This observation highlights that, even though PAS variants can and do direct cleavage and polyadenylation at many sites throughout the genome, these tend to be either at lowly expressed genes or lowly used pA sites. This is reinforced when looking at the average read number stemming from pA sites containing the different PASs (Figure 12).

When looking at the positioning of pA sites containing specific PASs it is found that the proximal pA sites (first and middle classification) tend to deviate more from the canonical hexamers (Figure 14). This has been observed before in a range of metazoans (Beaudoing, 2000; Jan et al., 2011; Martin et al., 2012; Shepard et al., 2011; Smibert et al., 2012; Ulitsky et al., 2012). This arrangement is generally described as an efficient distal pA site serves as the default site, with the weaker proximal pA site being regulatory and as a way to remove the downstream regulatory elements that lie in the aUTR (Shi, 2012).

An alternative cause for this observed pattern of weaker proximal pA sites may be simply because the introduction of an efficient pA site downstream would reduce the selection pressure on the proximal pA site to direct highly efficient cleavage and polyadenylation, thereby allowing the degeneration of the proximal PAS without any deleterious effect. This brings us back to the fail-safe mechanism as described in Section 3.3.2. Furthermore, the fact that the distal pA site is stronger may itself define the fact that it is the distal site, i.e. there may be other sites downstream of the designated distal pA site

capable of directing cleavage and polyadenylation, but the fact that the designated distal pA site is so strong means that these downstream sites are never used or observed.

3.3.4. *Conclusions*

In this chapter, the general polyadenylation landscape in HEK293s was analysed and the potential reasons for such an excess of pA sites are considered. Depending on which of these scenarios are true, APA may or may not play a significant physiological role in the regulation of gene expression. In reality, all of the proposed reasons are likely to be true to some extent. From previous studies, as described in Section 1.5, it is clear that in the case of many individual genes, the specific pA site used does indeed influence the ultimate protein level. However, the pervasiveness of this is still unclear, and the following chapters will attempt to address this issue and focus on the isolation of genes in which APA is likely to play a major physiological role.

4. CHAPTER 4 – DIFFERENTIAL REPRESENTATION OF APA ISOFORMS IN SUBCELLULAR FRACTIONS

4.1. INTRODUCTION

In the previous chapter, a clean subcellular fractionation protocol was established using HEK293 cells and 3'READS was used to quantitatively map the pA sites in these subcellular fractions. This data was then used to give an overall view of the polyadenylation landscape in HEK293 cells and suggested a few evolutionary pressures that may shape this APA landscape. The most accepted of these is that multiple pA sites are present within a single gene to provide a mechanism for gene regulation by switching between these pA sites, to include or omit potent regulatory elements in the mRNA, and thereby alter its final protein output. It is these APA events, where the APA isoforms are differentially regulated, that will likely be of most physiological significance.

In this chapter, the cytoplasmic and nuclear ratios of these different APA isoforms are compared to identify these APA events that result in differentially regulated APA isoforms. This approach is described in Section 1.10.2, and centres on the premise that if APA isoforms from the same gene differ in either their cytoplasmic stability or nuclear export rate, this will result in a change in the ratio of those two isoforms in the cytoplasm compared to the nucleus.

Two different approaches were used here to compare the subcellular APA landscapes. Firstly 3'end sequencing, specifically 3'READS, was used to compare the nuclear and cytoplasmic ratios of APA isoforms from HEK293 and HBL melanocyte cells. Subsequently,

using the DaPars algorithm (Section 1.10.3), an extensive analysis was performed on publicly available RNA-seq datasets from subcellular fractions of seven human cell lines that were generated as part of the ENCODE project (Djebali et al., 2012).

4.2. RESULTS

4.2.1. *Differential Subcellular APA Isoform Representation*

From the 3'READS data from the HEK293 subcellular fraction generated in Chapter 3, comparison of the ratios of the proximal APA isoform to the distal APA isoform was performed at a transcriptome-wide level. This was done by plotting the $\log_2$ ratio of the proximal APA isoform abundance between the cytoplasmic and nuclear samples versus the $\log_2$ ratio of the distal pA isoform between the two samples for each gene (Figure 16A). For UTR-APA analysis, this was restricted to the two most abundant 3'UTR pA isoforms. Genes which showed significant ($p < 0.01$, Fisher's exact test) differential representation of the APA isoforms were shown in red if the distal isoform was overrepresented in cytoplasm (lengthening in the cytoplasm) or shown in green if the proximal isoform was overrepresented in the cytoplasm (shortening in the cytoplasm) (Figure 16A).

This approach identified ~350 events in which the proximal APA isoform was overrepresented in the cytoplasm, as exemplified by *CLAO1* (Figure 16C), and ~100 events in which the distal APA isoform was overrepresented in the cytoplasm, as exemplified by *EMC8* (Figure 16D). However, while this approach identifies ~450 events where there is a significant differential representation of the APA isoforms, it is important to realise that the large majority (~90%) of the genes which do have multiple UTR-APA isoforms, fall into the grey region along the diagonal. This means that the different APA isoforms are represented at the same ratios in the different subcellular fractions indicating that there are no major

differences in either cytoplasmic stability or nuclear export rates between these APA isoforms.

(Figure on the following page)

Figure 16: Differential representation of UTR-APA isoforms in subcellular HEK293 fractions

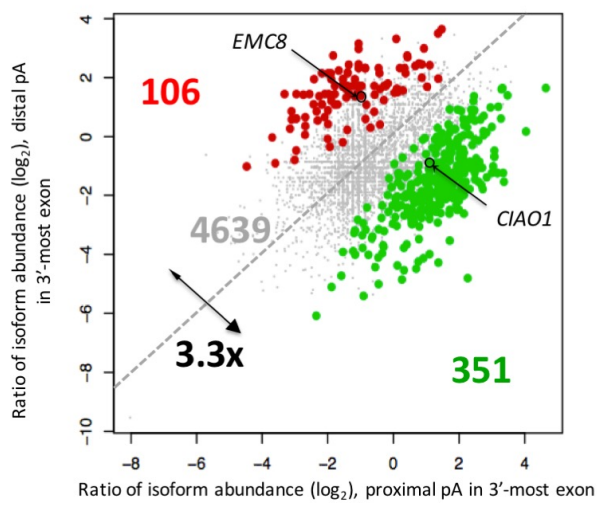
Panel A - Genes with UTR-APA isoforms that differ significantly in relative representation of their two most highly expressed APA isoforms in the cytoplasm compared to the nucleus. Genes are highlighted when the shorter APA isoforms (green) or longer APA isoforms (red) have a significantly higher relative representation in the cytoplasm compared to the nucleus. Significant differentially represented APA isoforms are those in which $P \leq 0.01$ (Fisher's exact test) and the abundance change is $>5\%$. The ratio of significant events moving in either direction is indicated by the double-ended arrow on the plot.

Panel B - As in A, but comparing the whole-cell and nuclear APA profiles.

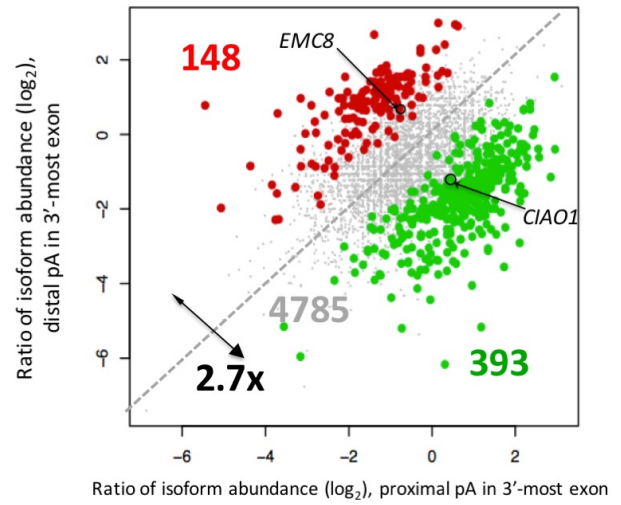
Panel C - Screenshot of *CIAOI* as an example of a gene showing an overrepresentation of its shorter APA isoform in both the cytoplasmic and whole-cell fractions relative to the nucleus. Read numbers are shown on the left, and proximal and distal pA sites are as indicated.

Panel D - *EMC8* gene as an example showing an overrepresentation of distal APA isoform in the cytoplasmic and whole-cell fractions. The positions of *EMC8* and *CIAOI* on the scatter plots are highlighted in Panels A and B.

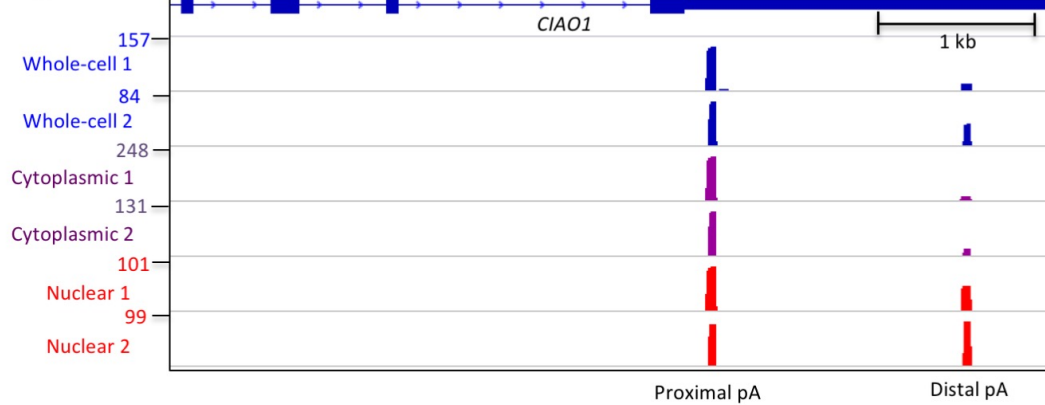
A Cytoplasmic vs Nuclear



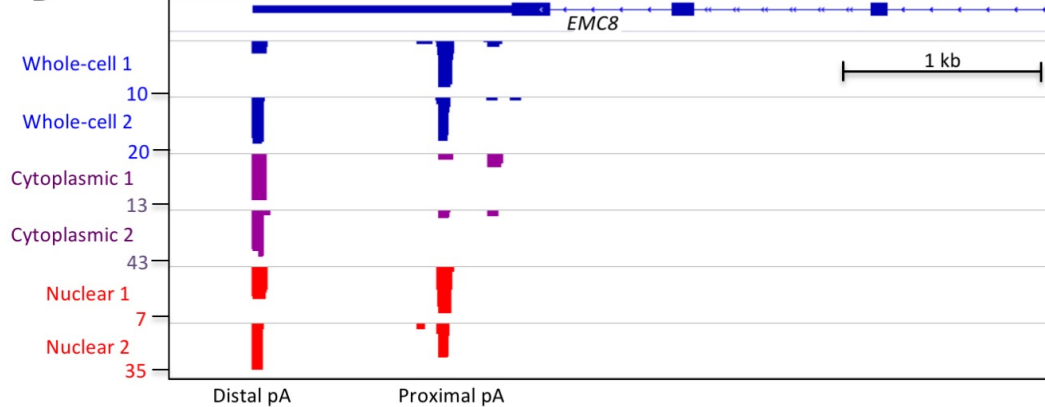
B Whole-cell vs Nuclear



C *CIAO1*



D *EMC8*



A similar approach was then also used to look at the differential subcellular distribution of CR-APA isoforms (Figure 17A). This was done by combining all reads upstream of the terminal exon and comparing these to the combined reads within the 3'UTR of that gene. The $\log_2$ ratio of the combined upstream pA sites between the two samples being analysed was compared to the $\log_2$ ratio of the 3'UTR pA sites between the two samples, as was done in the original 3'READS paper (Hoque et al., 2014). This analysis showed a similar pattern to that seen for UTR-APA with the distal APA isoforms being underrepresented in the cytoplasmic fractions when compared to the nucleus. This underrepresentation of the distal APA isoform in the cytoplasm is exemplified by *ETNK1* (Figure 17C) while *GEMIN8* shows an example of where the distal APA isoform is overrepresented in the cytoplasm of HEK293 cells (Figure 17D).

(Figure on the following page)

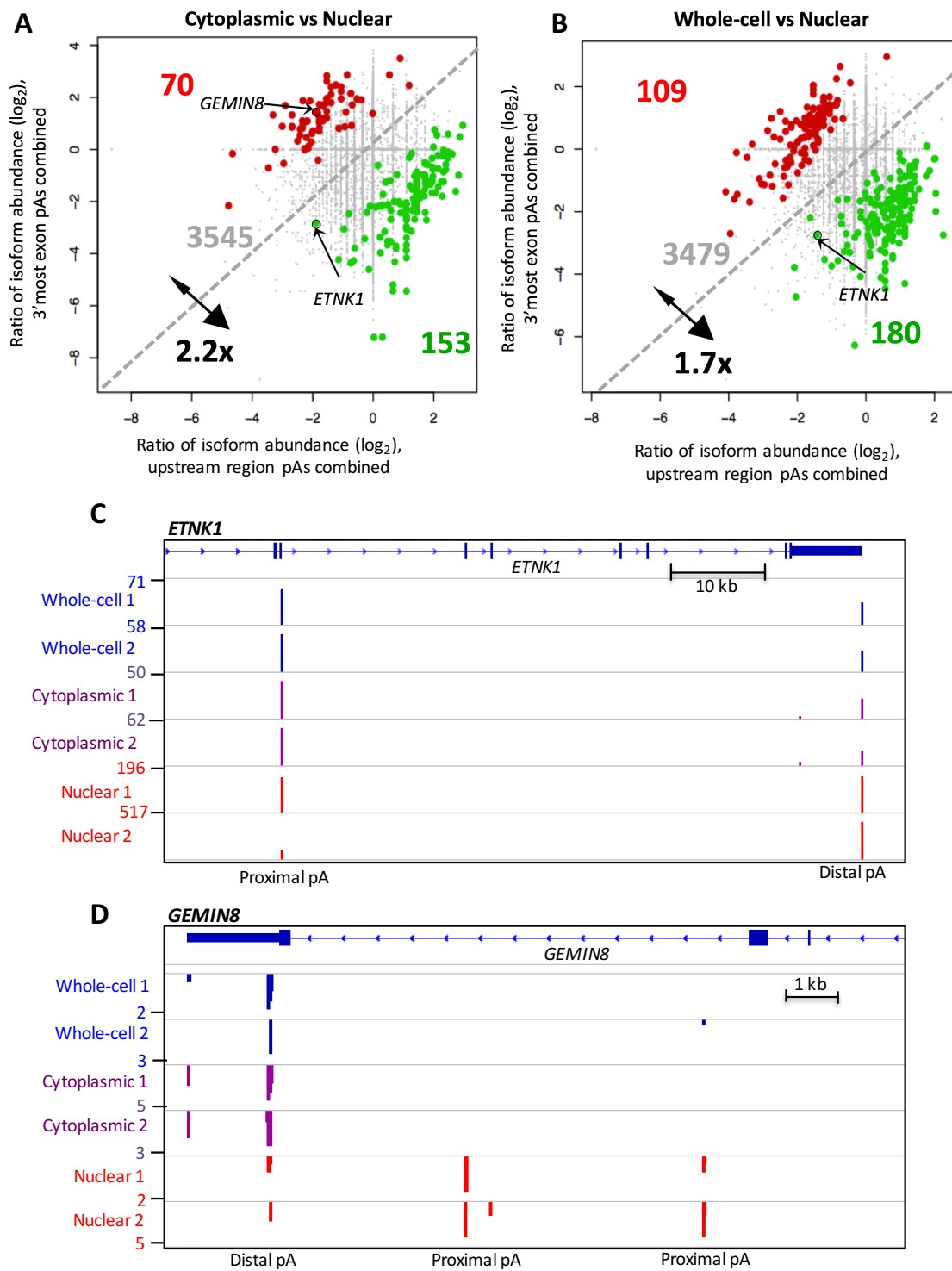
Figure 17: Differential representation of CR-APA isoforms in subcellular HEK293 fractions

Panel A – Genes with CR-APA that significantly differ in their representation of upstream-region (intronic and exonic) APA isoforms relative to 3'-most-exon APA isoforms between the cytoplasmic and the nuclear fractions. Genes are highlighted when the upstream APA isoforms (green) or 3' UTR isoforms (red) have a significantly higher relative representation in the cytoplasm compared to the nucleus. The ratio of significant events moving in either direction is indicated by the double-ended arrow on the plot.

Panel B – As in A, but comparing whole-cell fraction relative to the nuclear fraction.

Panel C – Screenshot of *ETNK1* as an example of a gene that shows an overrepresentation of the upstream APA isoform in both the cytoplasmic and whole-cell fractions relative to nucleus.

Panel D – Screenshot of *GEMIN8* as an example of overrepresentation of the UTR-APA isoform in the cytoplasmic fraction compared to the nuclear pool. Positions of *ETNK1* and *GEMIN8* on the scatter plots are highlighted.



4.2.2. *Reproducibility of subcellular APA distribution*

To address the robustness and reliability of these observations, this pipeline was applied to the two biological replicates of HEK293 cells, as well as in an HBL melanocyte cell line (Sánchez-Laorden et al., 2009). In both of these cell lines, there was a distinct pattern towards an overrepresentation of the proximal APA isoform in the cytoplasm, which would fit with the model that the 3'UTR is repressive and that the longer APA isoforms are generally more unstable in the cytoplasm, or being exported from the nucleus at a lower rate. To compare the relative levels of differential regulation, the read depth was controlled to 1 million reads per sample using the SAAP method, as described in Section 2.14.

The number of significant differential representation events using this SAAP method is shown for the two replicates of HEK293 cells and HBL cells in Figure 18. This analysis showed a consistent pattern between the three sample sets, confirming a general overrepresentation of the proximal pA isoforms in the cytoplasm of both cell lines.

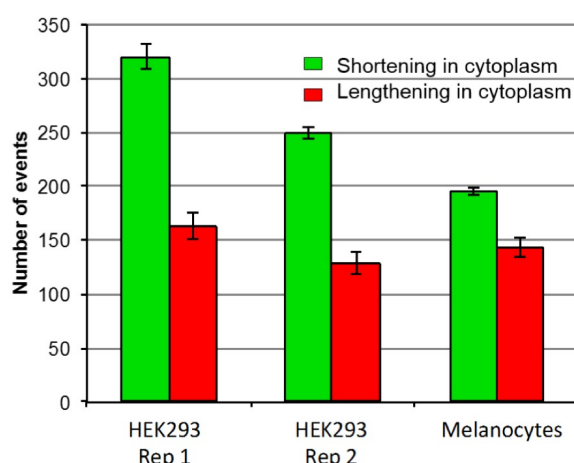


Figure 18: Differentially represented APA isoforms controlled for sequencing depth

The bar chart shows the number of genes that show a significant differential representation of UTR-APA isoforms in the cytoplasmic and nucleus based on the Significance Analysis of Alternative Polyadenylation (SAAP) method with FDR < 10%. This analysis controls read depth to 1 million reads per sample. The error bar is the standard deviation based on 20 values. Green bars represent the number of genes showing a higher representation of proximal APA isoforms in the cytoplasmic fractions. Red bars represent the number of genes in which distal APA isoform is overrepresented in the cytoplasm.

4.2.3. Corroboration of 3'READS data using northern blotting and RT-PCR

To verify the data obtained using the 3'READS method a number of different wet lab techniques were used to look at subcellular RNA distribution on an individual gene level. Firstly, 3'READS data showed the significant nuclear enrichment of *IQGAP1* mRNA relative to the cytoplasmic fraction (Figure 19A, lower panel). To confirm this subcellular distribution of *IQGAP1* in HEK293 cells, northern blotting was used with a probe targeting the terminal exon of *IQGAP1* (Appendix Table 1). This northern blot detects a band at a length corresponding to the 7.2 kb major *IQGAP1* mRNA isoform annotated on the Ensembl database. The differential abundance between the HEK293 cytoplasmic, nuclear and whole-cell RNA fractions on the northern blots corroborates the 3'READS data that shows *IQGAP1* is massively enriched in the nucleus (Figure 19A).

Reverse-transcription PCR was also used to confirm the subcellular distribution of *DIDO1* as measured by 3'READS (Figure 19B). To do this, 1 µg of RNA isolated from each subcellular fraction was used in the reverse transcription with an oligo(dT) primer (AP primer – Appendix Table 1).

Subsequently, to confirm the relative abundances of different APA isoforms, northern blot analysis of *GLUL* was performed. *GLUL* was chosen as it has few splice isoforms, and similar abundances of its two APA isoforms. To generate a strong enough signal from the northern blot, RNA isolated from the whole-cell and subsequently polyA purified using the PolyAtract System. Two micrograms of polyA purified RNA was then used for northern blotting. From these blots, two major bands are seen at ~1 kb and ~3 kb which would correspond to the proximal (annotated size 1.1 kb) and distal (annotated size 3.2 kb) APA isoforms, as detected by 3'READS. To ensure that this was the case, probes to both the cUTR and aUTR were used to distinguish between the proximal and distal APA isoforms (Figure 19C). As expected, both bands are detected when targeting the cUTR but only the upper band, corresponding to the distal APA isoform, was detected when using a probe targeted to the aUTR. The relative abundance of the two bands confirms the ratio of APA isoforms as measured by 3'READS.

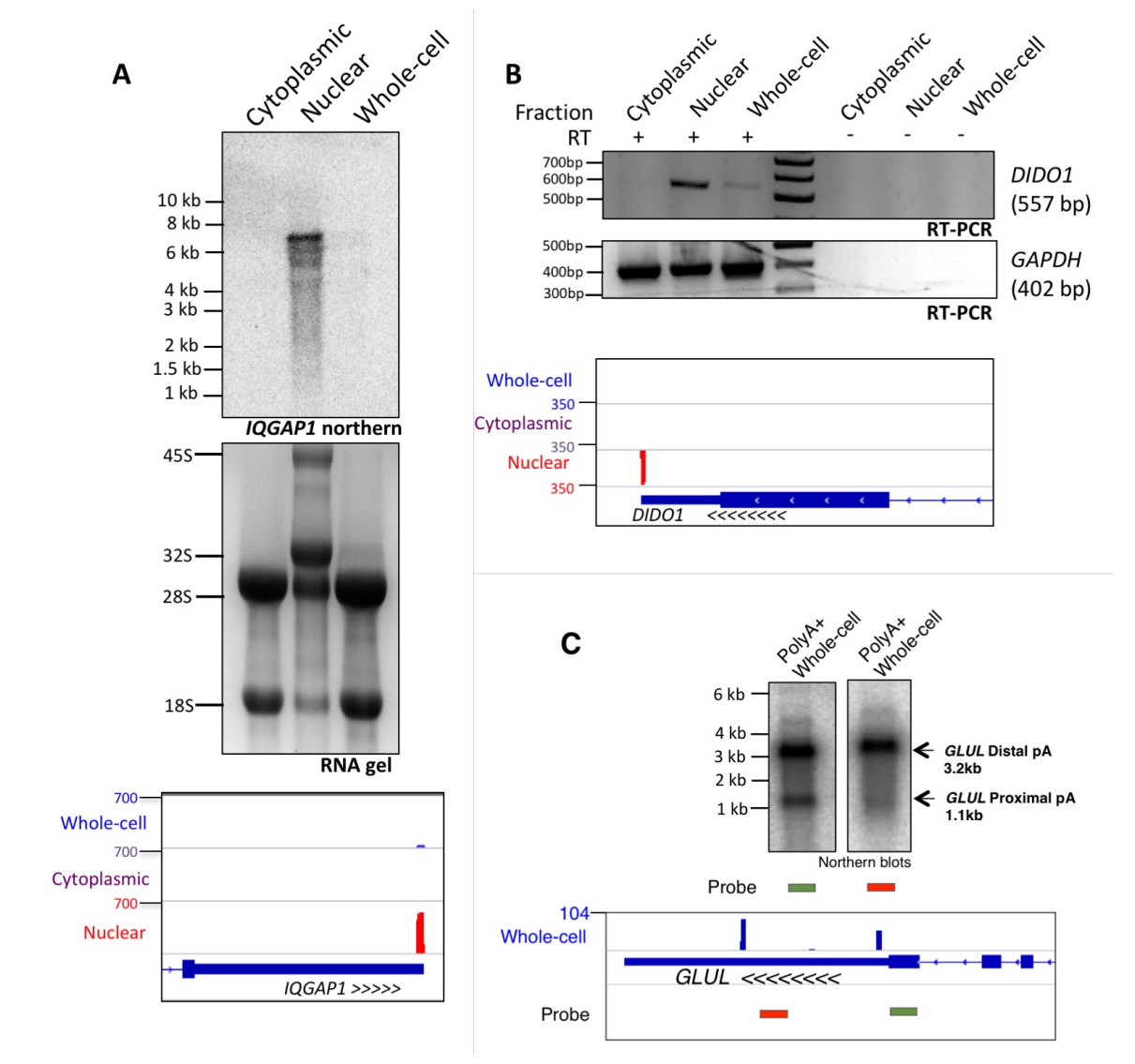


Figure 19: Confirmation of mRNA distribution as measured by 3'READS

Panel A - Upper panel shows a northern blot confirming the subcellular distribution of *IQGAP1* mRNA as measured by 3'READS (lower panel). The middle panel acts as a loading and fractionation control showing the distribution of ribosomal RNAs on the denaturing agarose gel prior to northern blotting, where 15 ug of RNA from each subcellular fraction was loaded.

Panel B - Reverse-transcription PCR confirming the subcellular distribution of *DIDO1*, as measured by 3'READS. The distribution as measured by 3'READS is shown in the lower panel. RT-PCRs of *GAPDH* were used as loading controls.

Panel C - Northern blot of whole-cell polyA+ HEK293 RNA corroborating the 2 UTR-APA isoforms of *GLUL* identified by 3'READS and their respective abundances. The 3'READS trace for whole-cell HEK293 RNA is shown in the lower panel. The probes used for each northern blot either target the common UTR (green) or the alternative UTR (red) to confirm the identity of the two APA isoforms. The positioning of the targets of the two northern probes is shown underneath the schematic of the *GLUL* gene.

Finally, to investigate the differential subcellular distribution of APA isoforms, the *CIAO1* gene was used. 3'READS data from this gene showed that the distal APA isoform is underrepresented in the cytoplasm, compared to the nucleus (Figure 20B). To investigate this northern blots with probes targeting the cUTR and aUTR of *CIAO1* were used (Figure 20A). The cUTR probe detects four major bands in the nuclear fraction of HEK293 cells. To try and distinguish these, the probe targeting the aUTR was used to isolate the APA isoforms using the *CIAO1* distal pA site. Using this aUTR probe, the sole band detected in the nuclear sample is between 4 and 6 kb. This was found to correspond to the distal APA isoform with the penultimate intron being retained (further discussion in Section 4.2.10). The proximal, fully spliced APA isoform would correspond to the ~1.5 kb band detected when using the probe targeting the cUTR (Figure 20A). With this assignment of these bands to the described isoforms, the northern blot shows an identical pattern of the distal APA isoform being present at a much higher level in the nucleus and almost completely absent in the cytoplasm (Figure 20A).

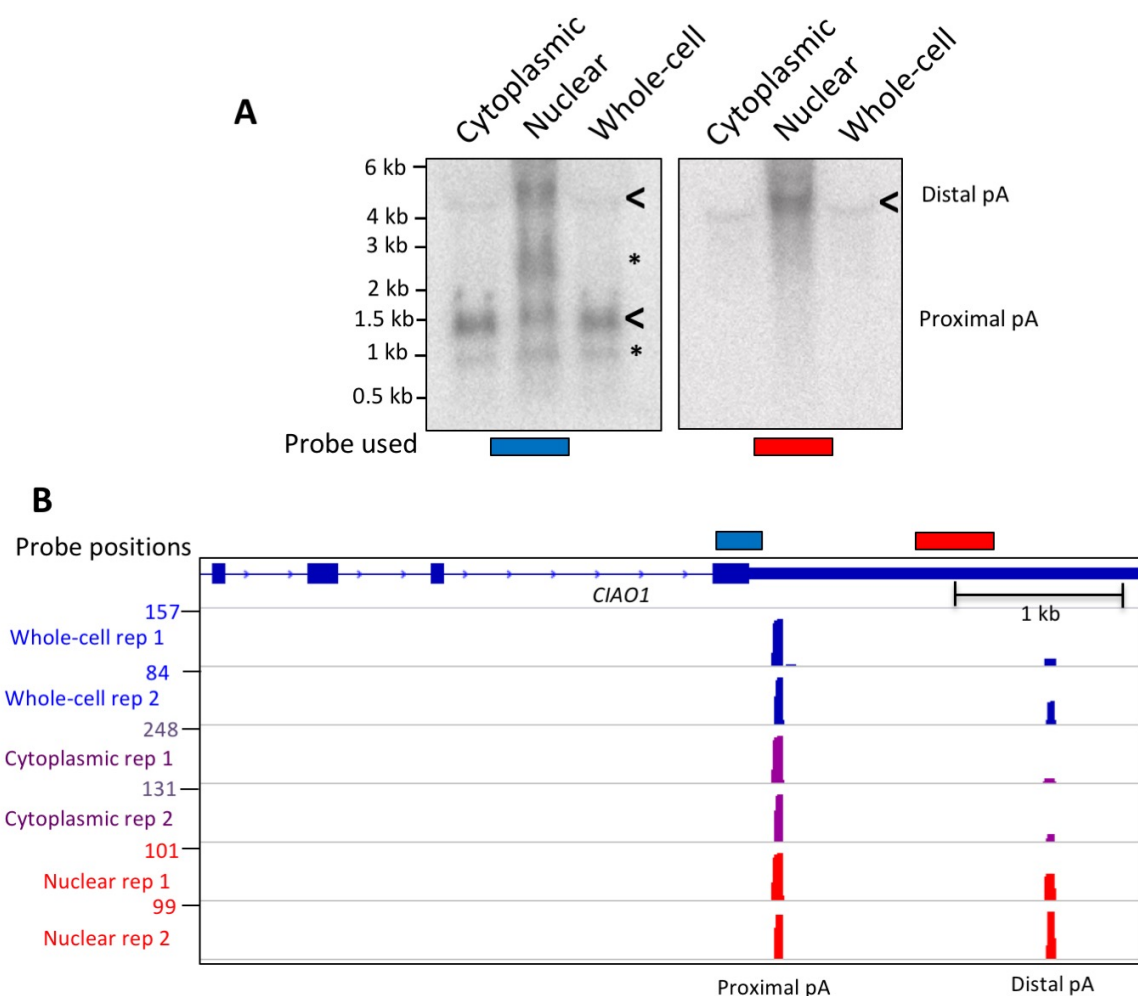


Figure 20: Northern blot analysis of *CIAO1* APA isoforms

Panel A - Northern blot of *CIAO1* in HEK293 subcellular fractions. Probes for constitutive UTR (left), alternative UTR (right) of *CIAO1* were used to corroborate the 3'READS data (lower panel). The assigned proximal and distal APA isoforms are indicated (<) with bands of unknown identity also indicated (*). Molecular size markers are indicated on the left of the blots.

Panel B - 3'READS tracks of subcellular fractions of HEK293s over the *CIAO1* gene. The proximal and distal pA sites are indicated below the tracks and the positioning of northern probes used in Panel A are shown above the schematic of the *CIAO1* gene.

4.2.4. *DaPars analysis to assess subcellular APA isoform distribution*

To investigate whether this bias towards proximal APA isoforms being overrepresented in the cytoplasm was consistent in other cell types, the publicly available DaPars algorithm (Masamha et al., 2014) was utilised which analyses the RNA-seq coverage over the length of annotated 3'UTRs (see Section 1.10.3). This was used in conjunction with the ENCODE datasets (Djebali et al., 2012) that were produced from full RNA-sequencing from polyA+ purified RNA. Datasets were only used if they contained sequencing data from both the nuclear and cytoplasmic RNA fractions. All datasets used, with associated accession numbers and read statistics are detailed in Appendix Table 4.

Despite fractionation controls being performed during the production of the data by looking at the relative levels of the precursor rRNA, an additional control was introduced of a minimum 20-fold enrichment of *MALAT1* in the nuclear fraction relative to its equivalent cytoplasmic fraction. This limited the analysis to seven human cell lines, as outlined in Table 11, with two biological replicates for all but HeLaS3s and K562s.

Table 11 - Cell lines used for cytoplasmic-nuclear comparison of APA profiles

Cell line	Tissue	Background
Mcf7	Cancerous breast tissue	A cell line derived from an invasive breast ductal carcinoma from a 69-year-old Caucasian female (Soule et al., 1973).
K562	Leukaemia cell line	Immortalised myelogenous leukaemia cell line isolated from a 53-year-old female (Lozzio and Lozzio, 1975).
IMR90	Lung fibroblast	A cell line derived from lung tissue of a Caucasian 16-week-old foetus, and have a finite lifespan (Nichols et al., 1977).
HepG2	Liver carcinoma	An immortal cell line derived from the liver tissue of a 15-year-old Caucasian American male with a well-differentiated hepatocellular carcinoma (Knowles BB, Howe CC, 1980).
SK-N-SH	Brain (neuroblastoma)	Derived from a metastatic site (bone marrow) from a 4-year-old female with neuroblastoma (Biedler et al., 1973).
HeLaS3	Cervix (adenocarcinoma)	A clonal derivative of its parent cell line, HeLa cells, which are derived from cervical cancer cells taken from a 31-year-old African American woman (Puck et al., 1956).
HUVEC	Umbilical vascular endothelium	A cell line derived from the endothelium of veins from the umbilical cord often used to study angiogenesis (Jaffe et al., 1973).
2b2 HEK293s	Embryonic kidney cells	A cell line derived from HEK293 cells containing an inducible integrated shRNA that targets <i>DICER1</i> mRNA (Schmitter et al., 2006).
HBL	Melanocytes	A human melanoma cell line (Deschodt-Lanckman et al., 1990)

One significant drawback of the DaPars approach is that it predicts the position of the proximal pA sites by looking at the point in which coverage over the 3'UTR changes between two samples (Masamha et al., 2014). The accuracy of these predictions is, therefore, dependent on the quality of the sequencing data and the ability to isolate events where there is a true biological difference in the relative abundances of different APA isoforms, as opposed to technical variation.

Therefore, to minimise the number of false events, the DaPars script was modified to include an additional filter, whereby the predicted proximal pA site had to lie within 250 nucleotides of an experimentally mapped pA site (Section 2.16). This 250 nucleotide cut-off was determined by using a range of different cut-offs and subsequently looking at the events that were excluded as the window narrowed. Narrowing from 500 to 250 nucleotides primarily resulted in the exclusion of events that looked to be potential artefacts, while narrowing from 250 nucleotides to 100 nucleotides removed a significant number of what looked to be true events.

The BAM alignment files were downloaded for all datasets that passed quality control, and converted to the BedGraph format required by the DaPars script (Masamha et al., 2014). Data from each cell line was then run through the modified DaPars script, combining biological replicates where available. A Δ PDUI of greater than 0.25 was required for proximal and distal APA isoforms to be considered to be significantly differentially regulated between samples.

Comparing the cytoplasmic and nuclear fractions, there is a very similar profile to that seen in the 3'READS data in six out of seven of the cell lines considered. These six show a significant overrepresentation of the proximal APA isoform in the cytoplasm when compared to the equivalent nuclear fraction (Figure 21). Interestingly, the neuroblastoma-derived SK-N-SH cell line shows completely the opposite pattern with the distal APA isoforms generally being overrepresented in the cytoplasm when compared to their equivalent nuclear samples.

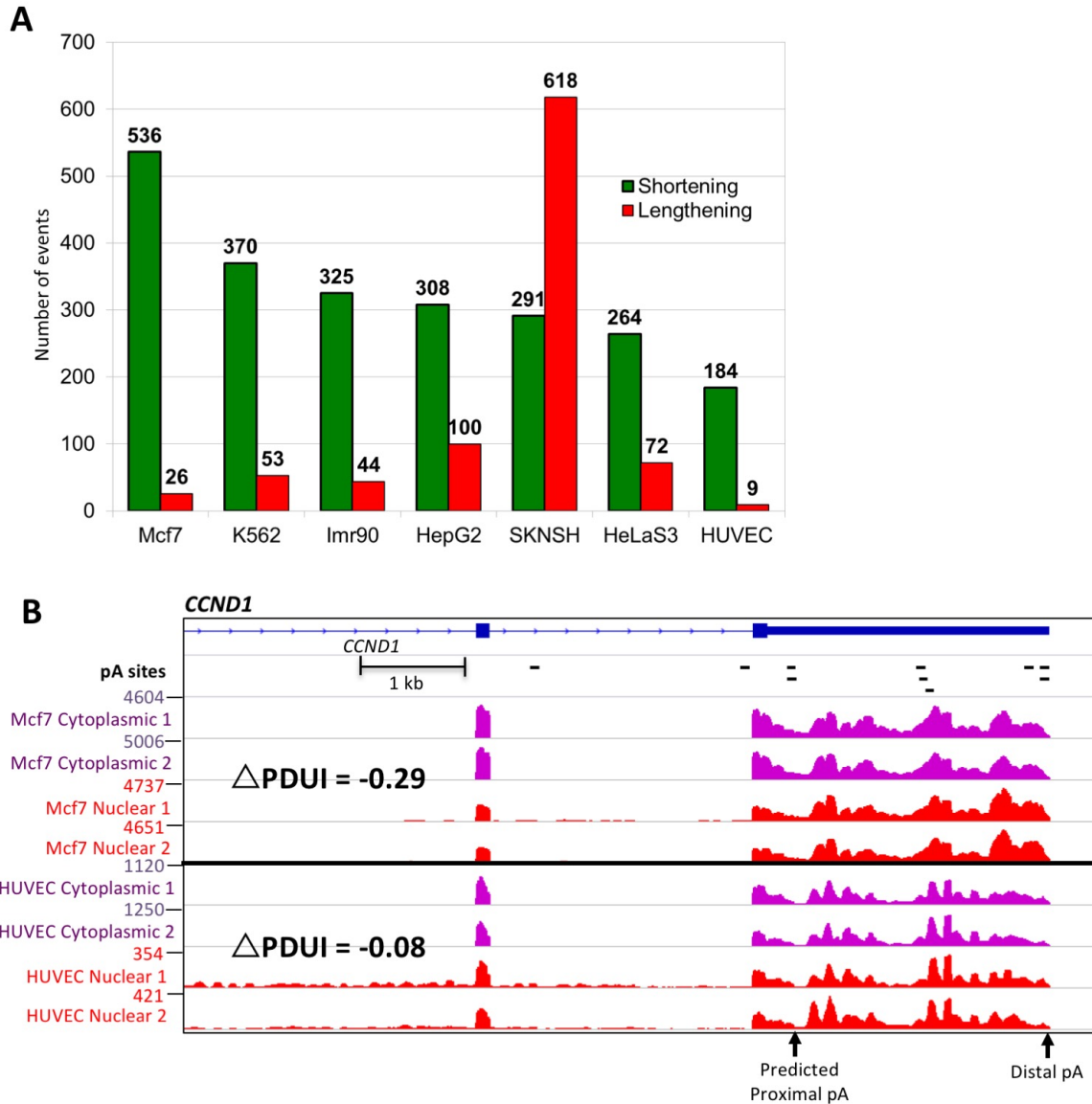


Figure 21: DaPars analysis of RNA-seq from subcellular fractions

Panel A - Summary of the number of genes that show differential APA isoform representation in the subcellular fractions of seven different human cell lines using the DaPars algorithm. Green bars represent events in which there is a higher representation of the proximal APA isoform in the cytoplasm. Red bars show a higher representation of the distal APA isoform in the cytoplasm. The RNA-seq datasets analysed here were downloaded from the ENCODE database (Djebali et al., 2012).

Panel B – RNA-seq tracks of the nuclear and cytoplasmic polyA+ RNA fractions from the Mcf7 and HUVEC cell lines over the *CCND1* gene. Biological replicates of each condition are shown, and the calculated average change in percentage distal usage index between the subcellular fractions is shown for each cell line. The positions of verified pA sites are shown below the schematic of the *CCND1* gene. The RNA-seq datasets shown here have been downloaded from the ENCODE database (Djebali et al., 2012).

4.2.5. *Isolated genes in which APA results in differentially regulated APA isoforms*

The main aim of this subcellular fractionation approach is to isolate the individual genes in which APA produces differentially regulated APA isoforms, as it is these events in which shifting between these APA isoforms will likely alter the genes overall expression. Therefore, it is important to look at exactly which genes have been isolated and which biological processes these genes are involved in.

Looking firstly at individual genes, several genes that have previously been well characterised, regarding the generation of differentially regulated APA isoforms, were picked up across the different cell lines. Specifically, the distal APA isoform of *CCND1* was identified as being underrepresented in the cytoplasm of Mcf7 cells, but not in the other cell lines (Figure 21B, the HUVEC cell line is shown as an example of a cell line where no differential regulation is seen). In previous studies, the differential regulation of the *CCND1* APA isoforms has been well characterised (Masamha et al., 2014; Mayr and Bartel, 2009; Wiestner et al., 2007). Furthermore, several other examples of genes known to produce differentially regulated APA isoforms were also identified in this analysis, including *VMA21* (Masamha et al., 2014), and *ELAVL1* (HuR) (Al-Ahmadi et al., 2009).

4.2.6. *Conservation of Subcellular Distribution of APA Isoforms Across Different Cell Types*

To investigate whether the differential subcellular distribution of APA isoforms in different cell lines involves the same genes or cell-line specific genes, the genes isolated in each of the different cell lines were compared. Interestingly, the majority of differential representation events do appear to be cell type specific, with only eleven genes that show shortening in all seven ENCODE cell lines (Figure 22A). These 11 genes are shown in Figure

22B. *HEMK1* is shown as an example of one of these genes which show differential regulation of APA isoforms in all cell lines analysed (Figure 23). To confirm that this lack of overlap between cell lines is due to true biological differences, rather than technical variation, the two Mcf7 biological replicates were analysed separately and the overlap of events between these replicates was compared. This analysis showed that ~70% of events that occur in replicate 1, also occur in replicate 2 (Figure 22C), indicating that the majority of cases are true cell type specificity rather than technical variation.

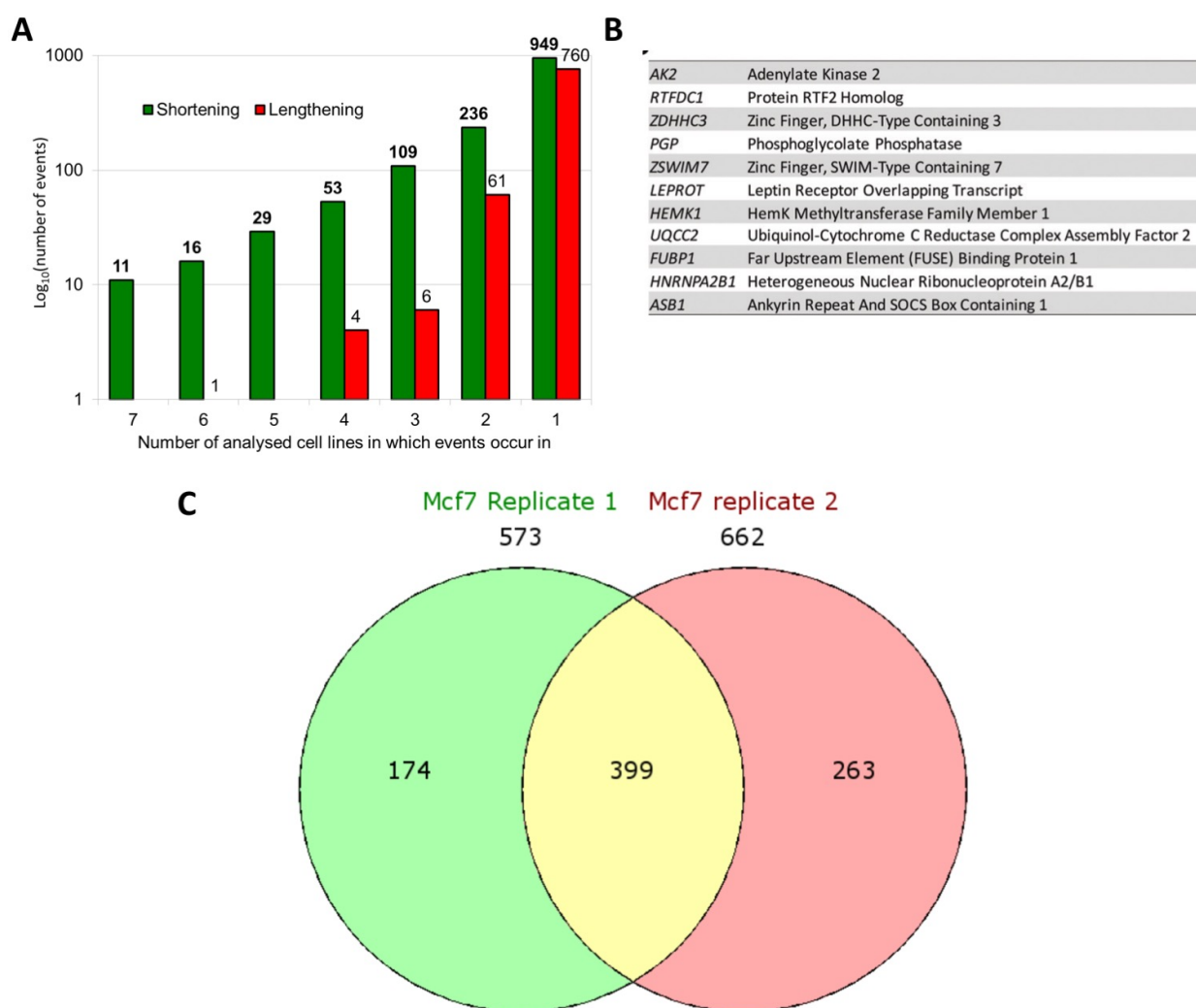


Figure 22: Cell-type specificity of differentially regulated APA isoforms

Panel A - Analysis of the number of genes that show the same pattern of differential regulation across the ENCODE cell lines (Djebali et al., 2012) analysed by DaPars. Genes that show significant differential APA isoform representation using DaPars analysis were cross-compared between cell lines. The number of events that occur in x number of cell lines was plotted on a log10 scale. Events in which the proximal APA isoform are overrepresented in the cytoplasm are shown in green, whereas events in which distal APA isoforms are overrepresented in the cytoplasm are shown in red. The absolute number of events is shown above each bar.

Panel B - Identity of the 11 genes that show an overrepresentation of shorter APA isoforms in the cytoplasm in all seven cell lines analysed.

Panel C – Overlap of differentially regulated APA isoforms in two biological replicates of Mcf7 when analysed separately using the DaPars algorithm. The RNA-seq datasets analysed here were downloaded from the ENCODE database (Djebali et al., 2012).

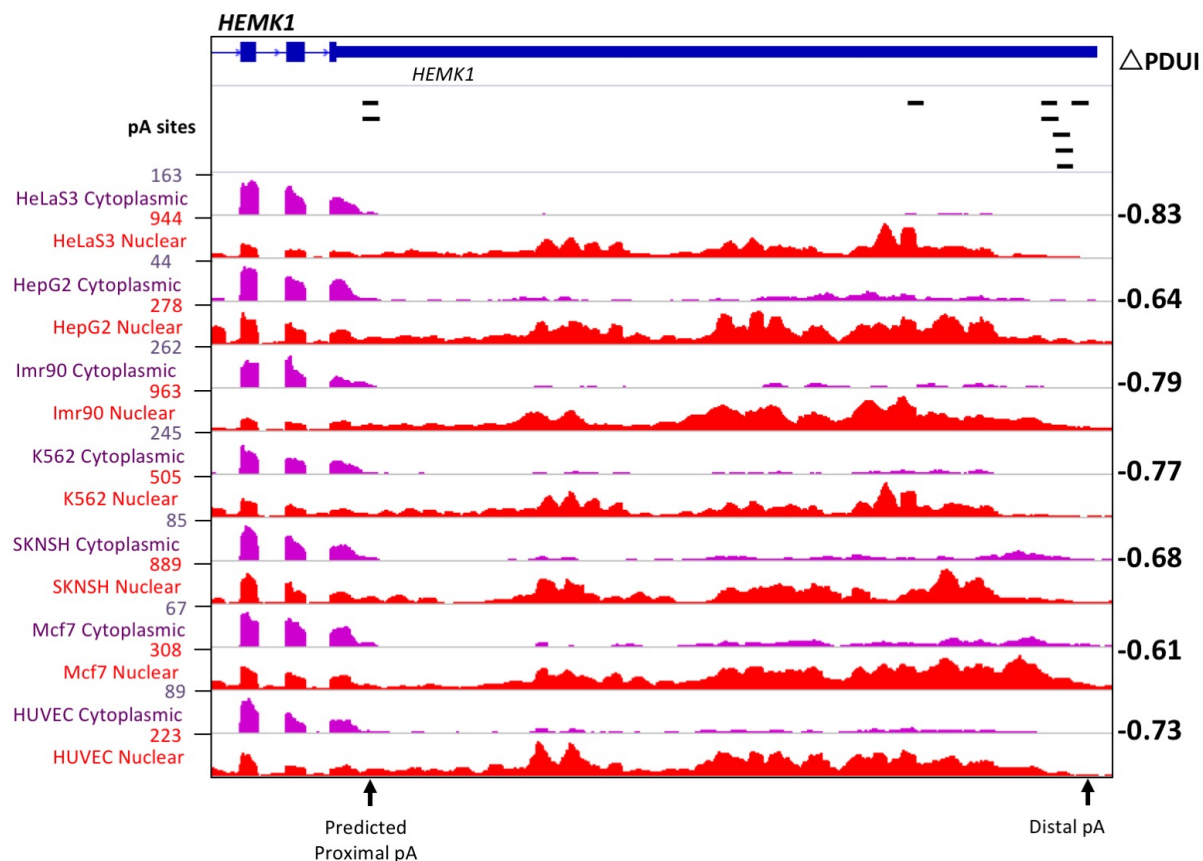


Figure 23: Screenshot of ENCODE RNA-seq data of *HEMK1*

HEMK1 is an example of a gene showing an overrepresentation of the shorter APA isoform in the cytoplasm in all ENCODE cell lines analysed. The change in percentage distal usage index between the subcellular fractions of each cell line is shown on the right with a negative value indicating an overrepresentation of the proximal APA isoform in the cytoplasm. Verified pA sites are shown below the schematic of the *HEMK1* gene. The RNA-seq datasets shown here have been downloaded from the ENCODE database (Djebali et al., 2012).

4.2.7. *Gene ontology enrichment of genes which show differential regulation of APA isoforms*

To determine whether the genes from each cell line, with differentially regulated isoforms, were involved in similar biological processes or pathways, gene ontology (GO) analysis was used. This was performed separately on each ENCODE cell line used and solely focussed on the genes in which the distal APA isoform was underrepresented in the cytoplasm of that cell line. The GO analysis was performed using the PANTHER (Mi et al.,

2016) looking for enrichment in the “biological processes” classifications. Figure 24 shows the enriched GO categories in each cell line ($p < 0.05$). Significant enrichment of particular categories was only found in three of the seven cell lines analysed (Figure 24). Interestingly, the only GO term that is also found to be enriched in all three of these cell lines are genes coding for proteins involved in RNA processing.

SK-N-SH	Shortening in cell line	Expected	Fold Enrichment	P-value
mRNA splicing, via spliceosome	15	2.29	6.55	1.24E-04
mRNA processing	18	3.83	4.7	6.43E-04
RNA processing	27	7.49	3.6	8.24E-05
mRNA metabolic process	20	5.5	3.64	6.81E-03
RNA splicing, via transesterification reactions with bulged adenosine as	15	2.29	6.55	1.24E-04
RNA splicing, via transesterification reactions	15	2.33	6.45	1.52E-04
RNA splicing	17	3.23	5.26	3.15E-04
macromolecular complex assembly	33	12.56	2.63	2.89E-03
cellular component assembly	41	19	2.16	1.53E-02
cellular component organization	82	50.63	1.62	5.07E-03
single-organism organelle organization	43	20.67	2.08	2.08E-02
organelle organization	58	30.19	1.92	2.75E-03
Mcf7				
RNA processing	32	13.1	2.44	3.41E-02
organelle organization	89	52.79	1.69	2.32E-03
cellular protein metabolic process	100	61.99	1.61	2.58E-03
HepG2				
oxidative phosphorylation	8	0.71	11.33	5.94E-03
cellular metabolic process	119	75.1	1.58	6.51E-08
purine ribonucleoside metabolic process	13	2.61	4.97	2.35E-02
cofactor biosynthetic process	11	1.75	6.27	1.60E-02
mRNA metabolic process	18	4.97	3.62	2.50E-02
cellular macromolecule metabolic process	95	60.01	1.58	2.73E-04
RNA processing	22	6.77	3.25	1.11E-02
organelle organization	52	27.28	1.91	1.40E-02
cellular nitrogen compound biosynthetic process	58	31.15	1.86	5.63E-03
organic substance biosynthetic process	74	42.15	1.76	6.05E-04

Figure 24: Gene ontology analysis of genes that show a differential representation of APA isoforms

All gene ontology “biological process” categories with significant enrichment ($p < 0.05$) of the genes that showed an overrepresentation of proximal APA isoforms in the cytoplasm are shown for each ENCODE cell line analysed. Gene ontology analysis was performed using PANTHER Gene Ontology tools with the Statistical Overrepresentation Test using default settings.

An example of one of these genes in this “RNA-processing” GO group is *HNRNPA2B1*, which is one of the 11 genes which is found to be differentially regulated in all of the cell lines analysed (Figure 25).

These GO enrichments indicate that differentially regulated APA isoforms are found more commonly in these classes of genes, but these enrichment levels, although significant, are minimal and generally the genes which produce differentially regulated APA isoforms are involved in a wide range of biological processes and are not restricted to a specific subset.

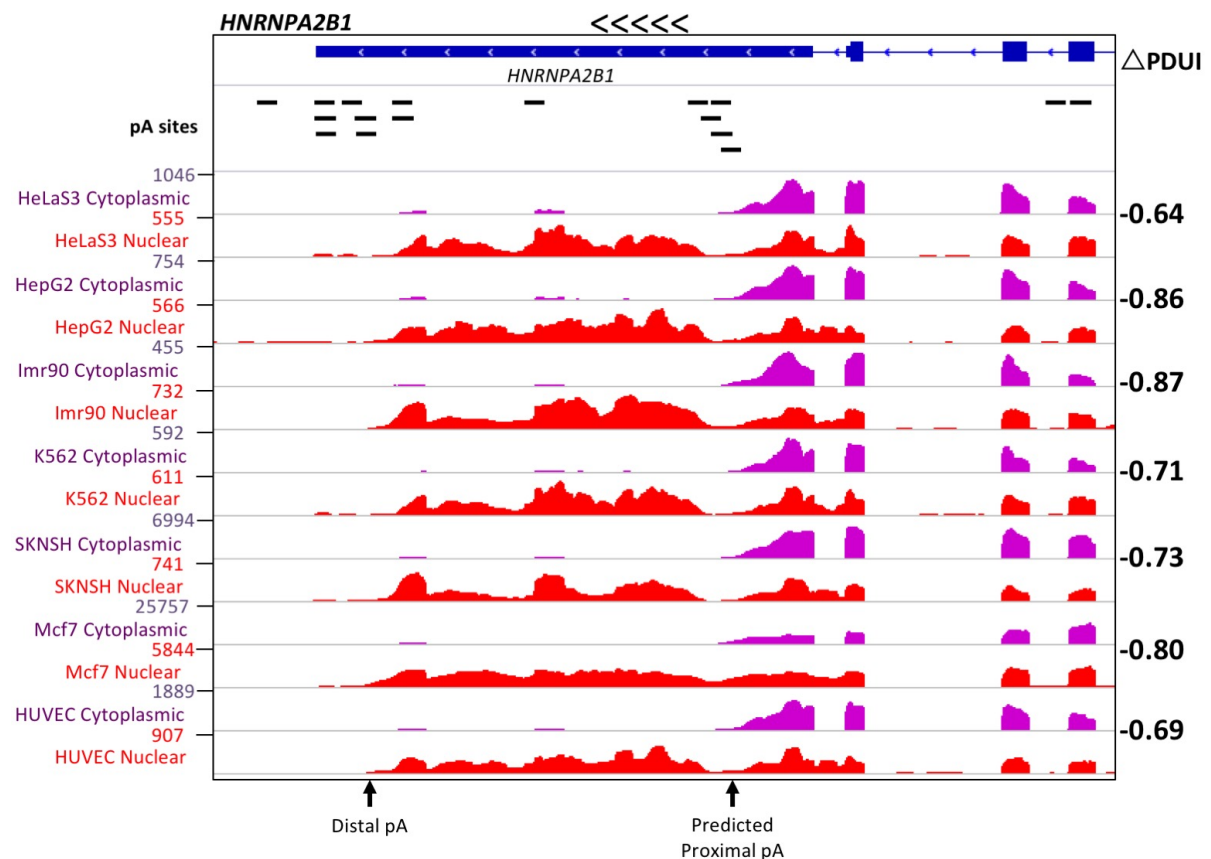


Figure 25: Screenshot of RNA-seq data of the *HNRNPA2B1* gene

HNRNPA2B1 is an example of a gene showing an overrepresentation of the shorter APA isoform in the cytoplasm in all ENCODE cell lines analysed. The change in percentage distal usage index between the subcellular fractions of each cell line is shown on the right with a negative value indicating an overrepresentation of the proximal APA isoform in the cytoplasm. The RNA-seq datasets shown here have been downloaded from the ENCODE database (Djebali et al., 2012).

4.2.8. *Motif discovery in the aUTRs of genes that show 3'UTR shortening in the cytoplasm*

The next step was to try and identify post-transcriptional mechanisms that may be contributing to the general pattern of overrepresentation of the proximal APA isoforms in the cytoplasm. The general mechanism behind the post-transcriptional regulation through UTR-APA is thought to be that the aUTR provides a platform whereby the distal APA isoform can be regulated independently of the proximal APA isoform (Di Giammartino et al., 2011).

In an attempt to identify *cis*-elements in the aUTRs that are potentially leading to differential regulation of the APA isoforms, the aUTR sequences of the genes that showed an overrepresentation of the proximal APA isoforms in the cytoplasm were extracted. This was done for each cell line and the extracted aUTRs were then run through the DREME motif analysis tool. This tool was chosen as it identifies motif enrichment in a given set of sequences relative to a control set of sequences. The control sequences used here were aUTRs from genes that did not show any differential regulation in any cell line. This approach is necessary, as there are a number of common motifs that are abundant in 3'UTRs, such as the *cis*-elements that direct cleavage and polyadenylation, as well as a general richness in A/T (Legendre and Gautheret, 2003). If a non-discriminatory motif analysis tool were used, such as MEME, it would likely pick up these motifs that are common to all UTRs, even though these motifs would likely not be responsible for the differential representation of the APA isoforms being analysed.

Figure 26 shows the top 3 motifs found in the aUTRs of genes that showed an overrepresentation of the proximal APA isoforms in the cytoplasmic fraction of in each cell line. For each motif in Figure 26, there is a corresponding E-value, which represents the enrichment p-value (Fisher's Exact Test) multiplied by the number of candidate motifs tested, as calculated by the DREME tool. Thus, the E-value essentially represents the probability of observing such results by chance.

These motifs were then compared to the motifs that are bound by known RNA-binding proteins as determined by the RNAcompete method (Ray et al., 2013). This comparison was done using the TOMTOM motif comparison tool (Gupta et al., 2007). The p-value cut off for a similarity to known binding motifs was 0.05. The RNA-binding protein that most closely matched the overrepresented motif is shown in Figure 26.

As seen in Figure 26, a large number of distinct motifs were identified in different cell lines; each potentially targets for different RNA-binding proteins. Interestingly, several of the motifs that are identified by the DREME analysis are similar to the motifs bound by several destabilising RNA-binding proteins, as classified in Ray et. al. (Ray et al., 2013). These destabilising RNA-binding proteins are highlighted in red in Figure 26, and stabilising proteins are highlighted in blue.

If these destabilising RNA-binding proteins were indeed targeting the aUTRs of the genes that show an overrepresentation of the proximal isoform in the cytoplasm, they could well be contributing to the observed distribution of APA isoforms.

The next step to take this part of the study would be to deplete the RNA-binding proteins in the cell lines where they are thought to play a role in differential regulation of APA isoforms. If these RNA-binding proteins were targeting these aUTRs and contributing to the observed nuclear-cytoplasmic APA isoform ratios, upon their depletion it would be expected to see a recovery of the distal APA isoforms in the cytoplasm.

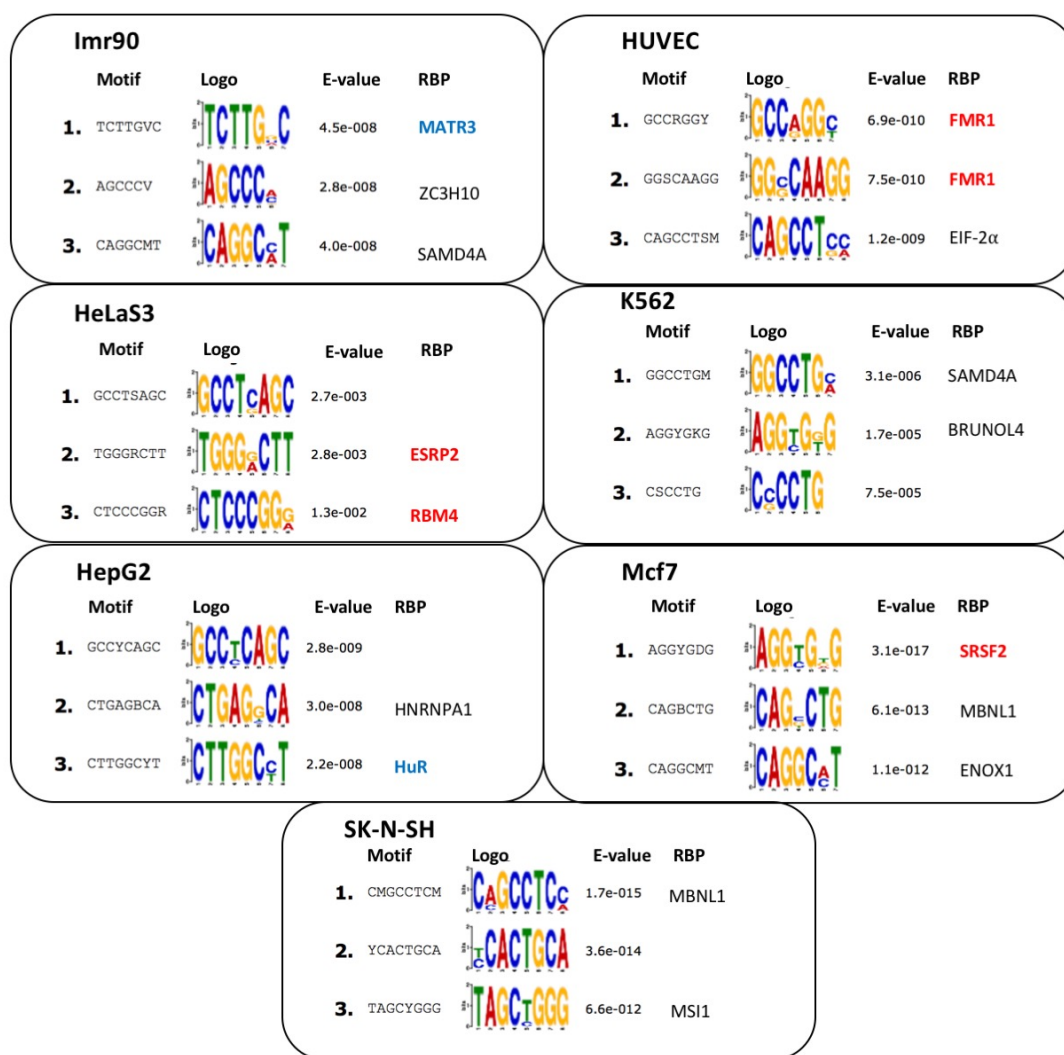


Figure 26: Motif analysis of aUTR regions of differentially regulated APA isoforms

The aUTRs of all genes that showed an overrepresentation of the shorter APA isoform in the cytoplasm in each individual cell line were extracted and subjected to motif analysis using the DREME tool. The top 3 motifs for each cell line are shown as well as corresponding RNA-binding proteins that potentially bind these motifs (Ray et al. 2013). RNA-binding proteins are highlighted in red or blue if they have previously been classified as exhibiting a destabilising or stabilising effect to their target RNAs, respectively (Ray et al. 2013).

4.2.9. *miRNA target sites in aUTRs that show shortening*

In addition to RNA-binding proteins, another major post-transcriptional regulatory pathway that has traditionally thought to be a major player in the differential regulation of APA isoforms is the targeting of the aUTR by miRNAs (Boutet et al., 2012; Di Giammartino et al., 2011). miRNA expression and potency are highly dependent on cell type and cell state

(Krol et al., 2010). Therefore, since the ENCODE datasets analysed here are from different cell types, it would be expected that there is a distinct battery of miRNAs acting in each cell line. Furthermore, depending on what miRNAs are expressed, and the positioning of the corresponding target sites, the APA isoforms being differentially regulated would also be expected to change in a consequent manner.

The next stage was, therefore, to look whether there was any enrichment of specific miRNA target sites in the aUTRs of genes that showed an overrepresentation of the proximal APA isoform in the cytoplasm. This was done using a modified approach that had been applied previously (Gruber et al., 2014). The miRNA-target site predictions used were from the EIMMo database (Gaidatzis et al., 2007), with only target sites with a “high confidence” being considered (EIMMo score > 0.8).

Firstly, the number of target sites were counted for each miRNA that fell within the aUTRs of the genes which show an overrepresentation of the proximal APA isoform in the cytoplasm for each cell type. Subsequently, the labels of each miRNA target sites were randomised so that the position of the target site no longer matched the identity of the miRNA. This was then repeated 500 times and, from this, a z-score for each miRNA was calculated to provide a measure of its enrichment in the aUTRs being analysed relative to what would be expected from a randomised dataset. The z-score is a measure of how many standard deviations away from the mean the measured number of target sites found in the aUTRs is, compared to the randomised dataset. A positive z-score indicates the aUTRs are enriched for the corresponding target sites. Indeed, this analysis showed a significant enrichment of several miRNAs in these aUTRs. This analysis provides some interesting insights into the miRNAs that may be contributing to the observed subcellular distribution of APA isoforms seen.

This analysis was repeated for each cell line analysed and the top 4 miRNAs that had enrichment of target sites in the aUTRs analysed are shown in Figure 27. Overall, this

analysis shows that in each cell line a distinct set of miRNAs are found to have significantly enriched target sites in their aUTRs of genes where the proximal APA isoform is significantly overrepresented in the cytoplasm. It is, therefore, these miRNAs that are most likely to be playing a significant role in the differential regulation of APA isoforms in each of these cell lines.

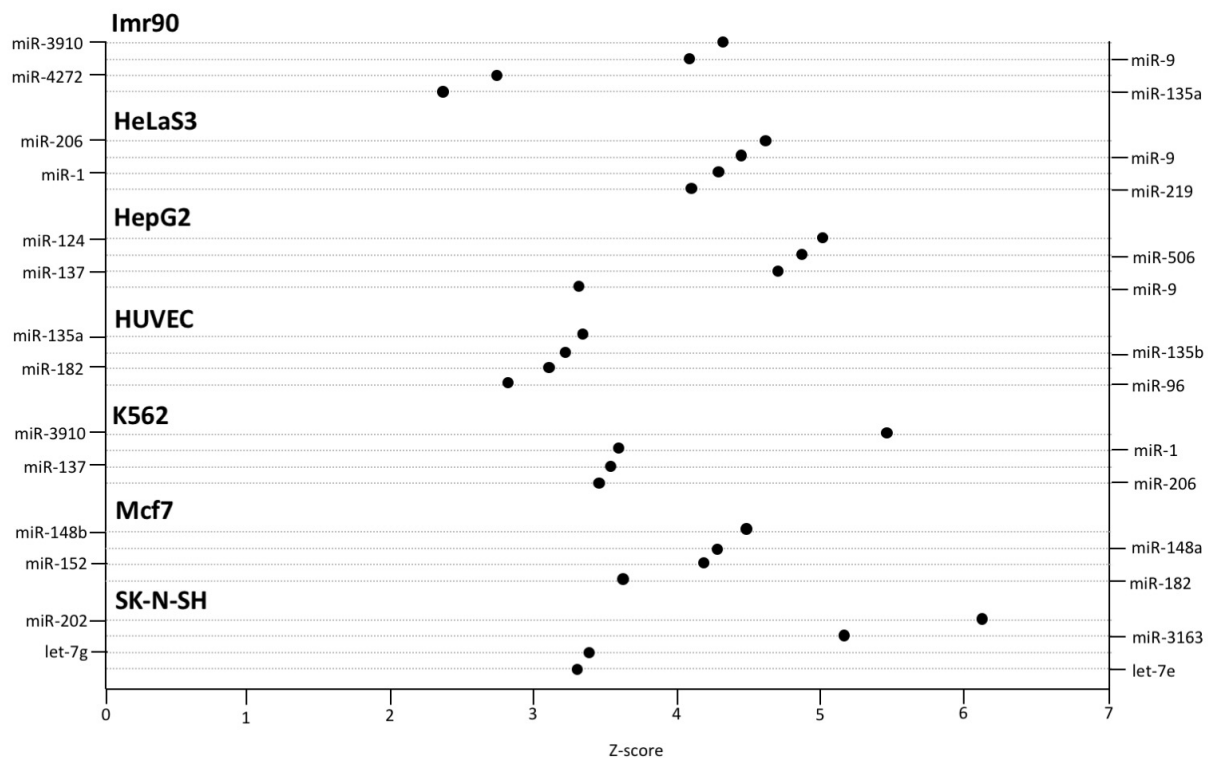


Figure 27: Enrichment of miRNA target sites in the alternative UTRs of differentially regulated APA isoforms

The x-axis represents the z-score of enrichment of miRNA target sites in the alternative UTRs (aUTRs) of the genes that show an overrepresentation of the proximal APA isoform in the cytoplasm of each ENCODE cell line analysed. This was calculated by comparing the observed frequency of target sites with what would be expected from random permutations of target sites across the set of alternative 3'UTRs. A positive z-score indicates that the frequency of miRNA target sites is greater than if sites were placed randomly in alternative regions of 3'UTRs. Only the top four miRNAs that have enriched target sites in these aUTRs for each cell line are shown.

4.2.10. *Intron retention in CIAO1 APA isoforms*

These miRNA and motif analyses are only effective in isolating post-transcriptional regulatory pathways if these pathways are acting on multiple genes. Therefore, some of the more focussed post-transcriptional regulatory pathways targeting small numbers of genes would be overlooked. To do this, each gene needs to be considered individually.

In Section 4.2.3, northern blotting was used to confirm the 3'READS data that showed the differential subcellular APA isoform representation of the *CIAO1* gene in HEK293 cells. To identify whether this pattern was due to differential stability of the APA isoforms, an actinomycin D approach was used. Northern blotting of *CIAO1* was performed using RNA isolated from the whole-cell after being subject to 10 ug/ml actinomycin D for the time periods indicated (Figure 28). Despite the differential APA isoform representation seen in *CIAO1*, both the proximal and distal APA isoforms did not appear to be obviously different in their relative stabilities (Figure 28).

During this initial northern blot analysis, it was noticed that the distal APA isoform detected was ~600 nt longer than expected (Figure 29A). Examination of standard polyA+ RNA-seq data from nuclear and cytoplasmic fractions of several different cell lines from the ENCODE project showed high-level retention of the intron between exons 5 and 6 (Figure 29B). This intron is 609 bp, and therefore it was hypothesised that retention of this was potentially resulting in the size discrepancy seen in the distal pA isoform.

To investigate this, a northern probe was designed against this intronic region, which detected a single band, limited to the nucleus, at the same size of that assigned as the distal APA isoform (Figure 29A). This, therefore, confirmed that the distal APA isoform detected by northern blotting included a retained intron between exons 5 and 6. This was a very intriguing result as any transcripts that contained this intron would be expected to be retained in the nucleus (Boutz et al., 2014). As this intron retention event seems to be limited to the

distal APA isoform, this could, therefore, explain why the distal APA isoform is underrepresented in the cytoplasm.

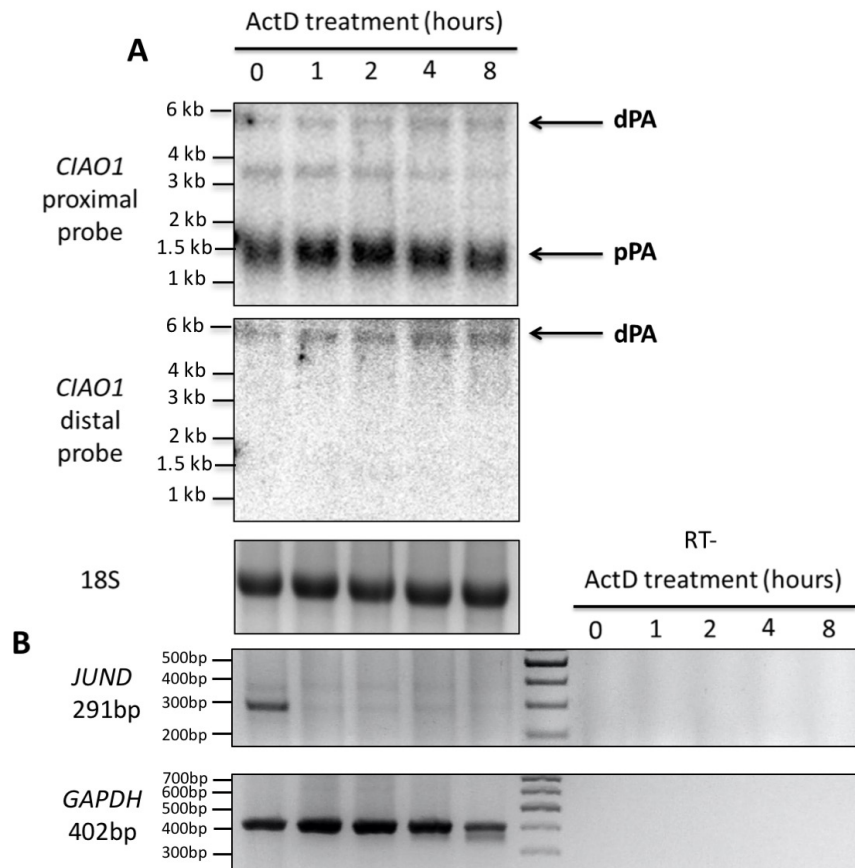


Figure 28: Stability experiment for *CIAO1* APA isoforms

HEK293 cells were exposed to 10 ug/ml Actinomycin D (ActD) for the times indicated. Total RNA was isolated at each time point and subjected to northern blotting with radioactive probes that detect both proximal and distal APA isoforms (proximal probe) and distal APA isoforms only (distal probe) respectively. 18S is shown as a loading control for the northern blot.

Panel B: An RT-PCR experiment showing the rapid degradation of *JUND* upon treatment with ActD, confirming the ActD is effectively stopping transcription. *GAPDH* is shown as a loading control. Reverse transcriptase negative controls are shown on the right of the DNA ladder in each panel. The expected sizes and size of the markers are shown to the left of each RT-PCR panel.

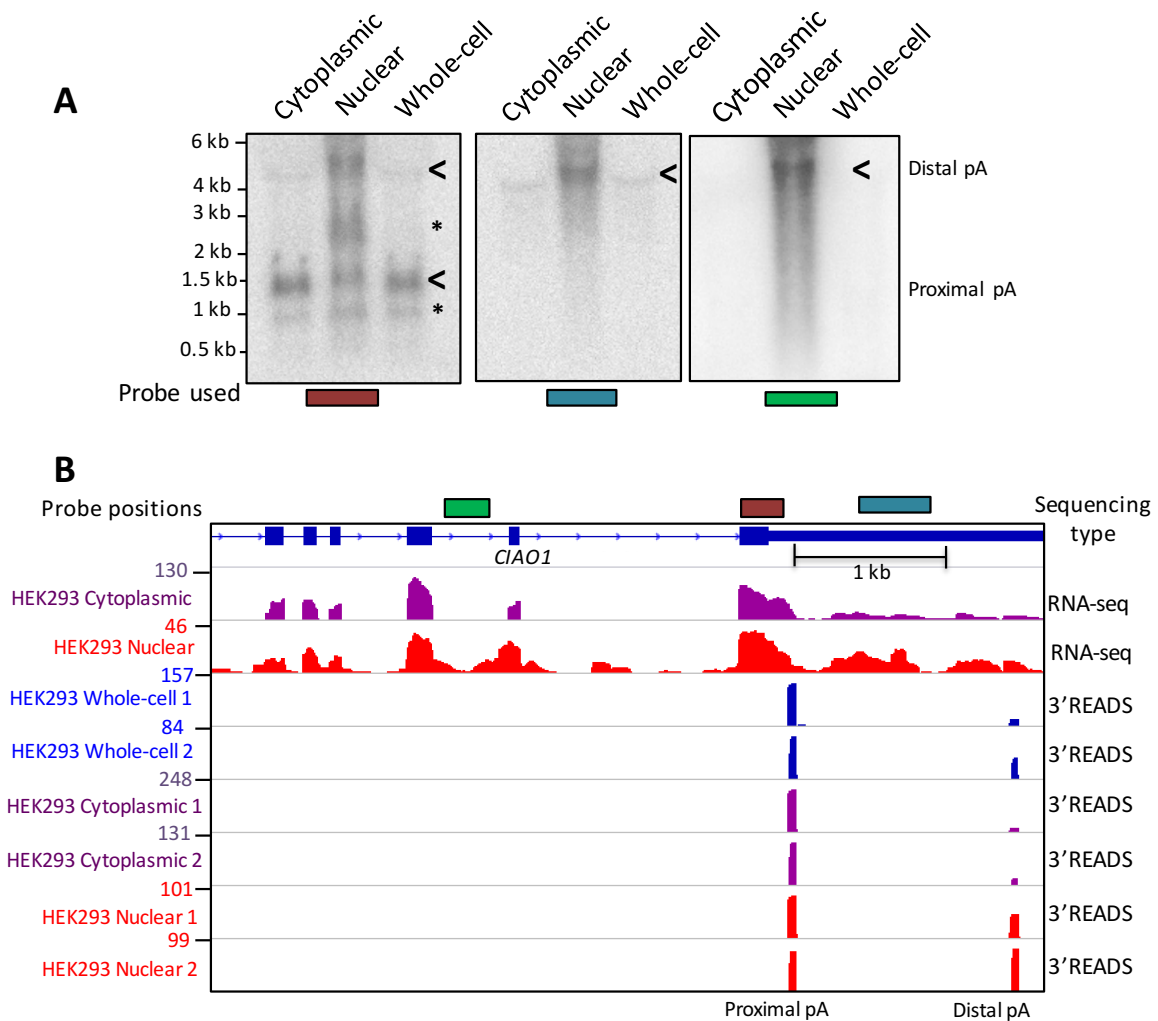


Figure 29: Overrepresentation of *CIAO1* distal APA isoform in the nucleus is coupled to upstream intron retention.

Panel A - Northern blot analysis of *CIAO1* APA isoforms in HEK293 subcellular fractions. Probes targeting the constitutive UTR (left), alternative UTR (middle), and the upstream intron (right) were used to corroborate the 3'READS and RNA-seq data.

Panel B – RNA-seq and 3'READS sequencing tracks over the *CIAO1* gene. The top two tracks show RNA-seq of nuclear and cytoplasmic RNA fractions from HEK293 cells. Below, 3'READS tracks of subcellular fractions from two biological replicates of HEK293s over the *CIAO1* gene are shown. The cell line is indicated on the right of the tracks, and positioning of northern probes used in Panel A are shown above the *CIAO1* gene schematic.

4.2.11. *Intron retention-coupled to specific APA isoforms*

While visually inspecting the examples of differential representation of APA isoforms in the subcellular fractions of the HEK293 datasets, alongside RNA-seq data from the same cells, several genes showed a similar example of the RNA-seq profile as that of *CIAOI*, including *FBXO22*, *TNPO1* and *UQCC2* (Figure 30). These three examples all show extensive intron retention in their terminal intron, as well as an overrepresentation of the proximal APA isoform in the cytoplasm.

Although reads mapping to the intronic regions in nuclear polyA⁺ RNA is not unusual (Braunschweig et al., 2014) it does raise the possibility that this intron retention in these specific APA isoforms may well be contributing to their nuclear retention. However, the ultimate fate of the nuclear-retained transcripts is unknown but are likely either undergo subsequent post-transcriptional processing or nuclear degradation. Furthermore, as this data is derived from fragmented RNA it is not possible to clarify whether these intron retention events are coupled to specific APA isoforms in the same way as in *CIAOI*, and these will all have to be verified on an individual gene basis.

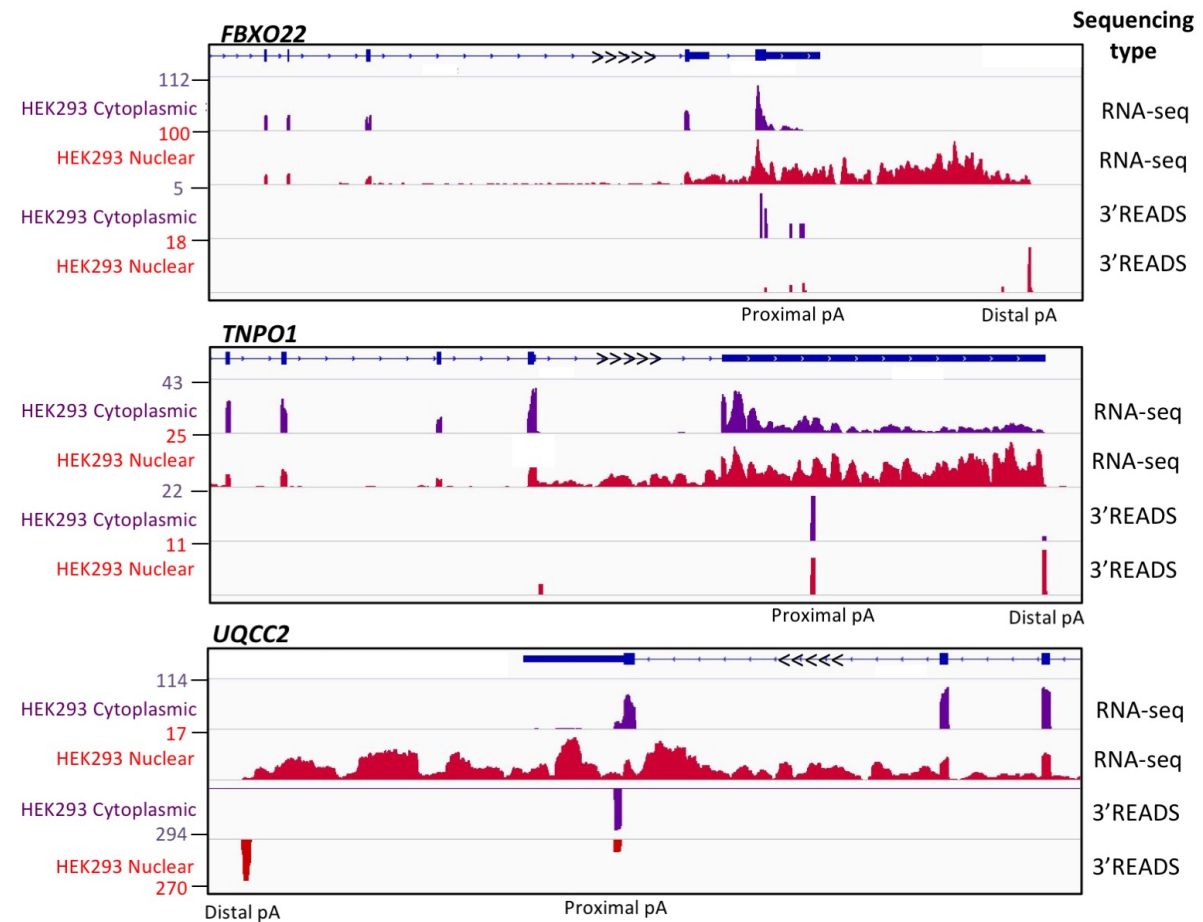


Figure 30: Intron retention is seen at multiple genes the display differential subcellular APA representation

RNA-seq and 3'READS from HEK293 cells is shown for three separate genes. These all show distal APA isoforms being restricted to the nuclear fraction alongside an upstream intron retention event. For each event, proximal and distal pA sites are shown as confirmed by 3'READS in the lower tracks of each panel.

4.3. DISCUSSION

4.3.1. *Methods for the identification of differentially regulated APA isoforms*

In this chapter, comparison of subcellular APA profiles is established as a novel and effective approach for isolating APA isoforms that are differentially regulated in either their cytoplasmic stability or nuclear export rates. Previous studies that have attempted to identify differentially regulated APA isoforms have been surprisingly limited considering the level of attention the APA field has been given as a whole. To date, the studies that have aimed to

identify differentially regulated APA isoforms have done so by using transcriptional arrest (Gupta et al., 2014; Spies et al., 2013), the limitations of which are discussed in Section 1.10.2.

Compared to these previous studies, the data produced here through our subcellular fractionation approach are surprisingly consistent considering these studies were performed using different methods in different organisms. All these studies find that, in general, the minority (10-15%) of APA events generate differentially regulated APA isoforms (Neve et al., 2016; Spies et al., 2013). The pattern of the proximal APA isoform generally being overrepresented in the cytoplasm is also consistent, suggesting that these proximal APA isoforms have a higher cytoplasmic stability or a higher nuclear export rate (Djebali et al., 2012; Neve et al., 2016). Despite this, all studies also find a considerable number of events whereby the distal APA isoform is more stable than the corresponding proximal APA isoform suggesting that stabilising elements in aUTRs may be playing a more significant role than originally thought (Gruber et al., 2014; Spies et al., 2013).

4.3.2. *The isolation of differentially regulated APA isoforms*

In this chapter, two different techniques were used for comparing subcellular APA profiles: 3'READS and DaPars. The advantages and disadvantages of these approaches are discussed in Section 1.10.3. The combined use of these approaches enables the direct examination of APA landscapes in a given RNA sample, through using 3'READS, while DaPars allows the utilisation of pre-existing data sets and reanalysing them in the context of APA. Through combining these methods, the subcellular APA profiles in 9 different human cell lines have been analysed.

Together, this has isolated a large number of differentially regulated APA isoforms, with a large degree of cell type specificity (Figure 22A). The main aim of this approach is to

isolate those APA events which do result in these differentially regulated APA isoforms, as an APA event becomes somewhat irrelevant if the resulting APA isoforms have the same properties (Neve and Furger, 2014). It is those cases where the APA isoforms are differentially regulated, which are going to be the most interesting regarding potentially having an impact on gene expression.

An example of this is the identification of *CCND1* as being differentially regulated in Mcf7 cells (Figure 21B). High levels of the CCND1 protein have previously been found to correlate with poor prognosis in breast cancer patients (Roy et al., 2010). Furthermore, preferential use of the proximal pA site in *CCND1* is seen cancerous cell lines (Mayr and Bartel, 2009), with the proximal APA isoform producing higher levels of the CCND1 protein (Masamha et al., 2014). This higher protein output is thought to be through the increased stability of the shorter APA isoforms, due to the exclusion of specific miRNA target sites and AREs (Mayr and Bartel, 2009).

ELAVL1 (HuR) is another gene known to produce differentially regulated APA isoforms (Al-Ahmadi et al., 2009), which has also been identified through the DaPars approach here. Additionally, ELAVL1 itself is a well-known RNA-binding protein that regulates APA (Table 3), including its autoregulation (Dai et al., 2011). This autoregulation of RNA-binding proteins that influence APA appears to be a common theme, with other examples including CstF77 (Luo et al., 2013), TARDBP (Bembich et al., 2014) and FPA in plants (Duc et al., 2013). In these examples, it is the pA site choice that is being regulated, and for this regulation to have any physiological effect, the resulting APA isoforms have to be differentially regulated.

This observation is corroborated by the fact that in the GO analysis of the genes that were identified as producing differentially regulated APA isoforms, genes involved in RNA processing were one of the primary GO terms that were found to be enriched (Figure 24). Despite finding some enriched subclasses of genes that produce differentially regulated APA

isoforms, in general, there is a wide array of different genes that have been isolated that are involved in a range of biological processes.

4.3.3. *Differential regulation of APA isoforms via RNA-binding proteins*

The range of *cis*-elements contained within the aUTR that can potentially lead to differential regulation of APA isoforms is extensive and is discussed at length in Section 1.1.1. Briefly, these *cis*-elements can affect mRNA stability, localisation or translational efficiency. The *cis*-elements are thought to act predominantly through the binding of either RNA-binding proteins or miRNAs. Here, DREME analysis was chosen to isolate motifs that were enriched in the aUTRs of genes that showed an overrepresentation of the proximal APA isoform in the cytoplasm. This algorithm looks at the enrichment of specific motifs in these sequences versus a given set of control sequences. This, therefore, means that any motifs that are common to 3'UTRs, but not specifically to those of the subset being analysed, will not be picked up by the motif search.

Previous studies (Gruber et al., 2014) have incorporated evolutionary conservation into the identification of *cis*-elements that are thought to bind to aUTRs. This approach was considered but, APA as a whole is extremely lowly conserved (see Section 1.6), and is generally considered species-specific (Ara et al., 2006). Therefore, a mechanism by which an RNA-binding protein binds specifically between two pA sites would unlikely be conserved if the pA sites themselves have a low probability of being conserved.

The DREME approach did indeed identify motifs similar to those that had previously been identified as binding sites for RNA-binding proteins, including some that are known to destabilise targeted mRNAs. The database of RNA-binding protein motifs that the identified motifs were compared to was from a method known as RNAcompete (Ray et al., 2013). RNAcompete is an *in vitro* approach which incubates RNA-binding proteins with a complex

set of oligonucleotides, after which the RNA-binding protein-oligo complex is recovered and the bound oligos are then analysed using microarray analysis (Ray et al., 2013). Despite this being in non-native conditions, with the absence of co-factors and surrounding sequence elements, this is currently the largest collection of binding motifs for RNA-binding proteins. Future studies would aim to utilise actual RIP-seq data, which would define the actual binding sites of these RNA-binding proteins in specific cellular conditions. This would allow a much more accurate annotation of precisely what RNA-binding proteins were binding the aUTRs being considered.

4.3.4. *Differential regulation of APA isoforms via miRNAs*

The targeting of miRNAs to aUTRs has long been proposed as a method of differential regulation of APA isoforms (Majoros and Ohler, 2007). There are a significant number of individual gene studies that experimentally confirm this differential regulation, as discussed in Section 1.5.1 and Table 5. However, the majority of studies to date, including this one, use computational methods which rely on the predictions of miRNA target (Gruber et al., 2014; Nam et al., 2014; Wehrspaun et al., 2014).

miRNA-mRNA target predictions are notoriously difficult to accurately predict due to the multi-faceted nature of their interactions. They are dependent on a large range of factors including, target site accessibility, binding energy, seed complementarity and cellular conditions. As a result, there are no perfect databases that can accurately identify all miRNA-mRNA interactions in specific cell type or condition. The existing prediction algorithms vary in both their method and reliability (Tabas-Madrid et al., 2014). The database chosen here, EIMMo, uses a combination of seed pairing, site number and target site conservation (Gaidatzis et al., 2007). This database was chosen as, in a study comparing the reliability of

different prediction databases, EIMMo was found to contain the highest enrichment of experimentally validated interactions (Tabas-Madrid et al., 2014).

This approach identified a number of miRNAs in each cell that had significantly enriched target sites in the aUTRs of genes that showed an overrepresentation of the proximal APA isoforms in the cytoplasm (Figure 27). The premise of this approach was that if miRNAs were targeting the aUTR and resulting in the destabilisation of solely the distal APA isoform, this could be resulting in this observed APA isoform distribution. Interestingly, several of the miRNAs that were found to have significantly enriched target sites in these aUTRs have previously been identified as being associated with that particular cell line or disease type.

Firstly, in the breast cancer derived Mcf7 cell line the top 3 miRNAs with the highest enrichment of target sites in the analysed aUTRs are the miR-148a, 148b and 152, which comprise the miR-148/152 family. Interestingly, specifically miR-148b has previously been shown to be upregulated in the plasma of breast cancer patients (Cuk et al., 2013). Furthermore, miR-182, another miRNA with significant enrichment of its target sites in the Mcf7 aUTRs analysed, has been shown to play a significant role in promoting metastasis in general (Huynh et al., 2011), and specifically breast cancer metastasis (Lei et al., 2014). The identification that target sites for specifically these miRNAs are found to be significantly enriched in the aUTRs of the genes that show relatively reduced distal APA isoforms in the cytoplasm of Mcf7s, suggests that some of the differential APA isoform regulation, via these miRNAs, may well be an important in establishing the steady-state transcriptome in these cells.

Enriched miRNA target sites in the other cell lines analysed also identified interesting links to potential phenotypic drivers. miR-9 was identified as the miRNA with the second most significantly enriched target sites in the aUTRs of those genes that show an overrepresentation of the proximal APA isoform in the cytoplasm of the cervical cancer-

derived cell line HeLaS3. Interestingly, miR-9 has been found to be overexpressed with potential oncogenic function in cervical cancer (Wilting et al., 2013) and has been proposed as a biomarker for cervical cancer prognosis (Hu et al., 2010).

Despite these general trends, it is important to consider that some of the very specific regulatory pathways, such as a single miRNA targeting one or two genes, will not be identified using these methods. This is because these approaches rely on the same motif occurring multiple times in the sequences being analysed. For these genes, the only way in which the mechanism for the differential regulation of these genes can be determined is by examination of the aUTRs on a single gene basis. This point is also applicable when looking at motif searches for potentially targeting of RNA-binding proteins to aUTRs.

4.3.5. *Differential regulation of APA isoforms via retained introns*

One of the less studied mechanisms of post-transcriptional regulation is the retention of introns. The majority of introns are spliced out co-transcriptionally (Pandya-Jones, 2011), but there are a very defined set of introns which are spliced post-transcriptionally, which have recently been classified as detained introns (Boutz et al., 2014). These detained introns can have half-lives of over an hour and result in partially spliced transcripts which, consequently, remain in the nucleus (Boutz et al., 2014).

One of these detained introns identified in this study (Boutz et al., 2014), is the penultimate intron in *CLAO1*, which has been independently identified here (Figure 29). The northern blots confirming the retention of the intron here interestingly showed that the intron retention events appear to be specific for the distal APA isoform (Figure 29A). The mechanism by which this occurs is unclear, but the coupling between splicing and polyadenylation, in general, is a well-described phenomenon (Proudfoot et al., 2002).

This distally polyadenylated intron-containing APA isoform is almost completely restricted to the nucleus (Figure 29A). Since the differential distribution of the APA isoforms is not thought to be due to differential stability (Figure 28), this would suggest that the *CIAOI* distal APA isoform never leaves the nucleus. Alternatively, a less well-founded possibility could be that a subsequent post-transcriptional splicing event, to remove the detained penultimate intron, also coupled to a second cleavage and polyadenylation event at the proximal *CIAOI* pA site. This reaction would, therefore, result in the initial co-transcriptional processing producing an incompletely spliced distally polyadenylated *CIAOI* isoform, which is then post-transcriptionally processed into a fully spliced proximally polyadenylated *CIAOI* isoform. The regulation of this post-transcriptional event could then act as a method of regulating gene expression as it is controlling the level of *CIAOI* transcripts that are leaving the nucleus.

However, it must be emphasised that this is solely a speculative mechanism, and examples of two sequential polyadenylation events occurring within a single transcript have rarely been described. However, studies on transcriptional termination have indeed shown that multiple cleavage events can occur at the 3' end of an RNA, with the first cleavage event occurring within a terminator region and a subsequent cleavage event within the upstream pA site (Dye and Proudfoot, 2001). Importantly, since the proposed mechanism would be a post-transcriptional processing reaction, the processing machinery would have to be coordinated independently of the C-terminal domain of the Rpb1 subunit of RNAPII, and would likely require coupling to the post-transcriptional splicing of the retained intron.

Alternatively, the *CIAOI* aUTR may simply contain a *cis*-element that results in its nuclear retention where it is degraded. Regardless, the *CIAOI* APA isoforms do appear to be differentially regulated, and, depending on the pA site choice, different levels of corresponding CIAO1 protein would be expected.

Interestingly, previous studies have identified *CIAO1* as a UTR-APA event where the use of the proximal pA site is favoured in small intestinal neuroendocrine tumours (SI-NETS) (Rehfeld et al., 2014). *CIAO1* is thought to interact with the tumour-suppressor protein, Wilm's tumour 1 (WT1) and regulate its transcriptional activation activity, thereby also influencing cell growth and differentiation (Johnstone et al., 1998). Therefore, through solely the shift to the proximal pA site in SI-NETS, it would be expected that the expression of *CIAO1* would be increased, assuming that the differential regulation was also seen in these cells.

4.3.6. *Coding region APA*

So far only UTR-APA has been discussed. However, the analysis in Figure 17 also showed extensive CR-APA, which also showed a similar pattern to UTR-APA in that the proximal APA isoforms tended to be overrepresented in the cytoplasm. Contrary to UTR-APA, CR-APA isoforms will unequivocally result in a difference in the proteome as the APA isoforms, by definition encode for different proteins. The overrepresentation of the proximal APA isoforms in the cytoplasm means that, as well as coding for different proteins, these APA isoforms have different cytoplasmic stabilities and/or nuclear export rates.

4.3.7. *Conclusions*

Throughout this chapter, a novel approach of comparing subcellular APA profiles has been introduced to isolate differentially regulated APA isoforms. This analysis exposes a list of genes which need further individual analysis to identify the exact biochemical mechanism that governs their differential regulation of the resulting APA isoforms and whether this differential regulation has a key role in regulating gene expression. Here, individual gene studies on *CIAO1* have suggested a putative mechanism of differentially regulating APA isoforms via intron retention in specific APA isoforms. The following chapters will go on to

applying this subcellular fractionation approach firstly to try and directly identify which APA events are being differentially regulated through miRNA-targeted destabilisation, and then apply this approach to look at the differential regulation of APA isoforms in colorectal cancer progression.

5. CHAPTER 5 – THE IMPACT OF ***DICER1* KNOCKDOWN ON** **SUBCELLULAR APA PROFILES**

5.1. INTRODUCTION

In the previous chapter, it was found that different cell lines generally had a relatively distinct subset of genes that produced differentially regulated APA isoforms, and each subset of genes was involved in a wide range of different cellular processes. Motif and miRNA target site enrichment analysis were used to try and identify common post-transcriptional regulatory mechanisms that resulted in the observed differential subcellular distribution of APA isoforms.

To take a more direct approach, methods were considered of compromising the major post-transcriptional pathways which may be contributing to the observed differential subcellular abundances of specific APA isoforms. Since miRNA-targeted destabilisation is generally considered the major pathway for differential regulation of APA isoforms (An et al., 2013), it was decided to target the miRNA biogenesis pathway. This was with the aim to identify the events where a differential subcellular representation of APA isoforms was due to miRNA-mediated destabilisation. An overview of the differential regulation of APA isoforms via miRNAs is discussed in Section 1.5.1.

5.1.1. Methods of compromising miRNA-targeted degradation

The general miRNA biogenesis pathway is outlined in Figure 31. Along with DROSHA, DICER1 is a member of the highly conserved RNase III family of proteins. These proteins generally comprise a double-stranded RNA-binding domain and an RNase III

domain. In the canonical miRNA biogenesis pathway, DICER1 catalyses the essential step of cleavage of the precursor miRNA into the mature miRNA.

Previous studies that have involved compromising the miRNA biogenesis pathway as a whole have usually done so by targeting DICER1. Several different approaches have been used to ablate or reduce DICER1 activity including generating a HEK293 stable knockdown cell line (Bail et al., 2010; Schmitter et al., 2006), as well an inducible genetic ablation of the RNase III domain in mouse embryonic stem cells (mESCs) (Nesterova et al., 2008; Tan et al., 2015) and mouse embryonic fibroblast cells (MEFs) (Gantier et al., 2011).

5.1.2. *DICER1 inside of the nucleus*

In addition to the canonical post-transcriptional gene silencing pathway, DICER1 and DROSHA both participate in the non-canonical nuclear RNAi pathway. Three types of nuclear RNAi have been described: transcriptional gene silencing (Verdel, 2004), RNA-directed DNA-methylation, and RNA activation. These nuclear RNAi pathways all function through either small-interfering RNAs (siRNAs) or nuclear miRNAs.

Furthermore, DICER1 has been found to interact directly with chromatin, independently of small RNAs. It was first shown to physically associate with a subset of protein-encoding genes in fission yeast, triggering a silencing response (Woolcock et al., 2011). Here it was found that Dicer was localising to euchromatic regions around stress response genes to inhibit the initiation of transcription in the absence of stress (Woolcock et al., 2012). Subsequently, upon the onset of stress, such as heat stress, Dicer is exported and the stress-responsive genes are then able to be freely expressed (Woolcock et al., 2012).

In addition to regulating transcription initiation, DICER1 has also been implicated in the regulation of RNAPII termination. The association of DICER1 with transcriptional termination was identified as being mediated through R-loop formation (Skourti-Stathaki et

al., 2014). R-loops form as a result of the invasion of the DNA duplex by the nascent transcript and are found to be enriched at GC-rich regions, including CpG island promoters and GC-rich terminator elements (Ginno et al., 2013, 2012).

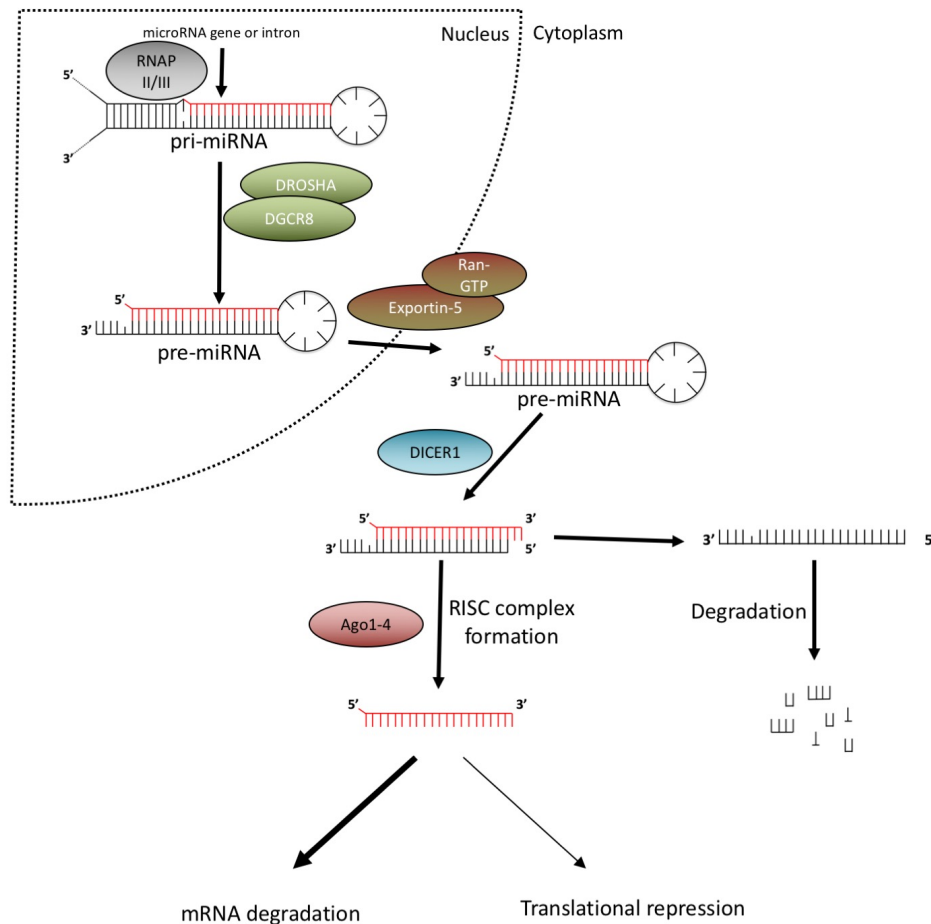


Figure 31: microRNA biogenesis pathway

The canonical pathway of microRNA (miRNA) biogenesis. Usually, transcription of miRNAs is performed by RNA polymerase II, but some miRNAs are transcribed via RNA polymerase III. Transcripts form a secondary hairpin structure creating the primary miRNA (pri-miRNA). The pri-miRNA is then processed by the microprocessor complex, which includes DROSHA and DGCR8. This then forms the precursor miRNA (pre-miRNA) which is then exported from the nucleus through the Exportin-5 pathway. In the cytoplasm, the pre-miRNA is further processed by DICER1 to form the mature miRNA. The functional strand of the mature miRNA is then loaded onto to the RNA-induced silencing complex (RISC). The miRNA will then guide RISC to its target binding sites which will result primarily in destabilisation, but also translational repression of the target RNAs.

Specifically, at terminator elements, R-loop formation is known to result in the formation of antisense transcripts. These antisense transcripts are then thought to bind the nascent transcripts and form the double-stranded RNA (dsRNA) that acts as the substrate for DICER1 (White et al., 2014). The processing of this dsRNA results in the production of endogenous siRNAs which can subsequently result in the formation of heterochromatin at this terminator region. Ultimately, the heterochromatic formation reinforces RNAPII pausing and efficient termination (Skourti-Stathaki et al., 2014).

5.1.3. *Overview of findings*

To investigate the effect of compromising the miRNA biogenesis pathway, two different approaches were used. Firstly, this was done using the inducible DICER1-knockdown HEK293 cells, and measuring the subcellular APA landscapes using 3'READS before and after DICER1 depletion. Secondly, the RNA-sequencing data from fractionated mESCs before and after genetic Dicer1 ablation (Tan et al., 2015) was analysed using the DaPars algorithm.

Surprisingly, both sets of data showed that reduction or loss of DICER1 activity affected the nuclear APA profiles more significantly than the cytoplasmic profiles. Subsequent individual gene studies, using a range of ChIP experiments alongside qRT-PCR, suggests that nuclear DICER1 is influencing pA site choice through its effect on the chromatin structure surrounding pA sites.

5.2. RESULTS

5.2.1. *DICER1 Knockdown Using a Tet-Inducible System*

In general, miRNA-targeting is considered the major pathway for differentially regulating APA isoforms (An et al., 2013), and miRNA-mediated regulation predominantly acts via degradation of targeted transcripts rather than translational inhibition (Guo et al., 2010). The post-transcriptional gene-silencing mechanism of miRNA-targeting occurs in the cytoplasm and is thought to be one of the main causes for the overrepresentation of the shorter APA isoforms in the cytoplasm seen in Chapter 4. Therefore, a block in the miRNA pathway would be expected to result in a recovery of the distal APA isoforms in the cytoplasm as they would no longer be being targeted for miRNA-mediated degradation.

Initially, to investigate this an inducible DICER1 knockdown HEK293 cell line was used (Schmitter et al., 2006). These cells were grown in the presence or absence of doxycycline, which, when present, induced the expression of a short-hairpin RNA (shRNA) which targeted the *DICER1* RNA resulting in its depletion. To ensure sufficient depletion of the DICER1 protein a western blot was performed after one and five days induction with doxycycline (Figure 32). This shows around a 2-fold depletion of the DICER1 protein after one day and a ~95% depletion after five days.

Previous studies using this same cell line measure the reduction in levels of several miRNAs after the induction of the *DICER1*-targeting shRNA and find that the majority of these are depleted by between 30-80% after three days (Bail et al., 2010). It was therefore decided that five days would provide a decent balance between ensuring a sufficient drop in the majority of miRNA levels but not leaving the cells too long to develop significant secondary effects (Schmitter et al., 2006).

After 5-days induction of DICER1 depletion, subcellular fractionation followed by RNA extraction was performed alongside a whole-cell protein extraction. The subcellular RNA fraction for both the control (non-induced) and five-day induced samples were run on

an RNA gel to ensure a clean subcellular fractionation as done previously (Figure 32). A western blot was then performed on the protein extract to compare the levels of DICER1 protein in the five-day knockdown relative to the non-induced sample (Figure 32). Two full biological replicates were performed for each sample, with the full quality controls performed each time. RNA fractions were then subject to 3'-end sequencing by 3'READS. Upon sequencing, the relative abundance of reads mapping to the nuclear-retained lncRNA *XIST* was compared, which showed a minimum of a 200-fold enrichment in the nucleus relative to the equivalent cytoplasmic fraction.

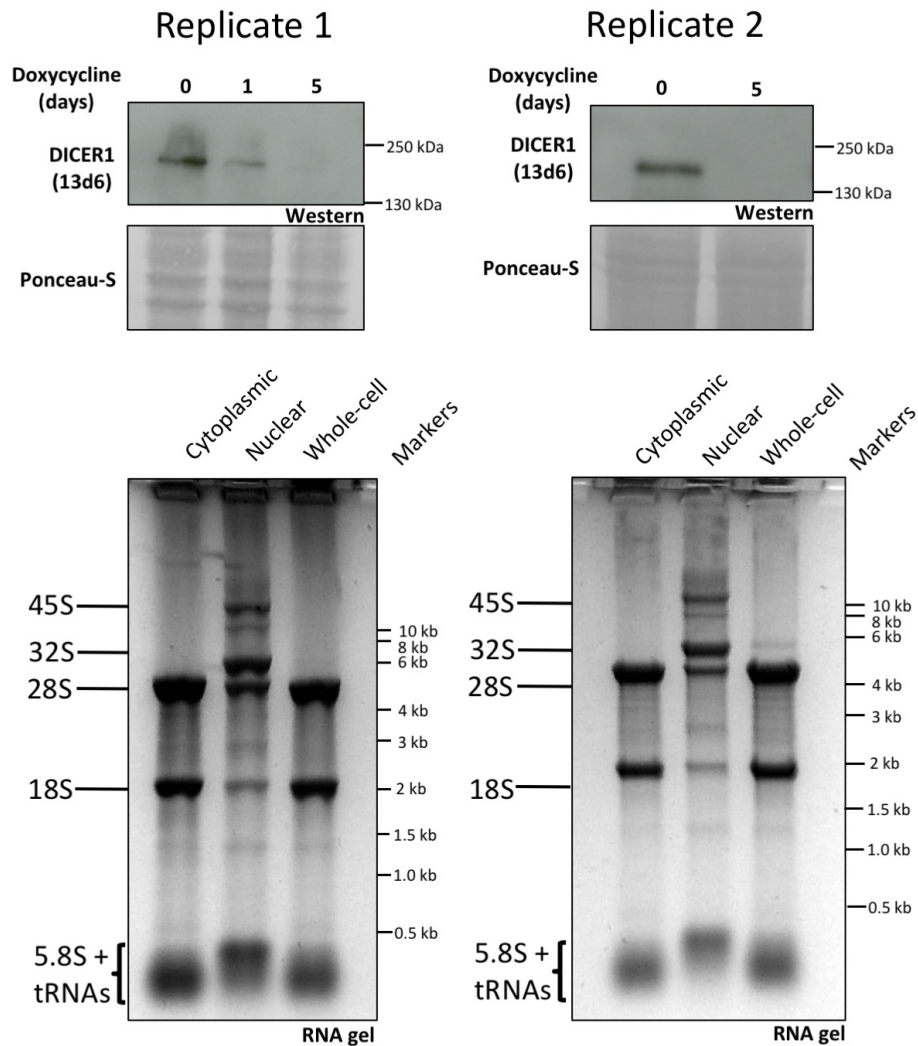


Figure 32: Quality control experiments for DICER1-depletion

The upper panels show western blots of DICER1 before and after induction of the shRNA targeting endogenous DICER1 via the addition of doxycycline. Time periods of DICER1 depletion are shown above each lane. Ponceau-S stains of the western membranes are shown for each blot as a loading control.

The lower panels show denaturing RNA gel with 2 ug of RNA from each subcellular RNA fraction. The identity of each ribosomal RNA band is shown on the left-hand side of the gel with size markers on the right-hand side of the gel. Only the RNA gels for the DICER1 depleted cells are shown, but no difference was seen before DICER1-depletion, and the gel for the first biological replicate of this is shown in Figure 6.

5.2.2. *The Effect of DICER1 knockdown on Subcellular APA Profiles*

The APA profiles of the cytoplasmic, nuclear and whole-cell RNA fractions from the two biological replicates of the DICER1-depleted and non-depleted HEK293 cells, were analysed using 3'READS. The sequencing depth for each sample can be found in Appendix Table 3. Once sequenced, the APA profiles were compared as in Section 4.2.1.

Comparing the UTR-APA profiles in the whole-cell fractions of the DICER1-depleted and non-depleted cells revealed that ~300 APA events occurred as a result of the loss of DICER1 (Figure 33A). In the case of solely miRNA depletion and recovery of the distal APA isoforms upon DICER1 depletion, it would be expected that the large majority of APA events would be due to a recovery of the distal APA isoform, and thus a lengthening in the DICER1 depleted cells. However, what is seen is around an equal number of shifts to proximal and distal APA isoforms upon DICER1 depletion.

One possibility that may increase the abundance of proximal APA isoforms upon DICER1 depletion is that the aUTR in the distal APA isoforms is forming a secondary structure with the cUTR causing miRNA-target site occlusion (Ameres et al., 2007). Therefore, only the proximal APA isoform would be targeted for degradation by the miRNA, and, upon DICER1 depletion, the proximal APA isoform would increase in abundance relative to the distal APA isoform.

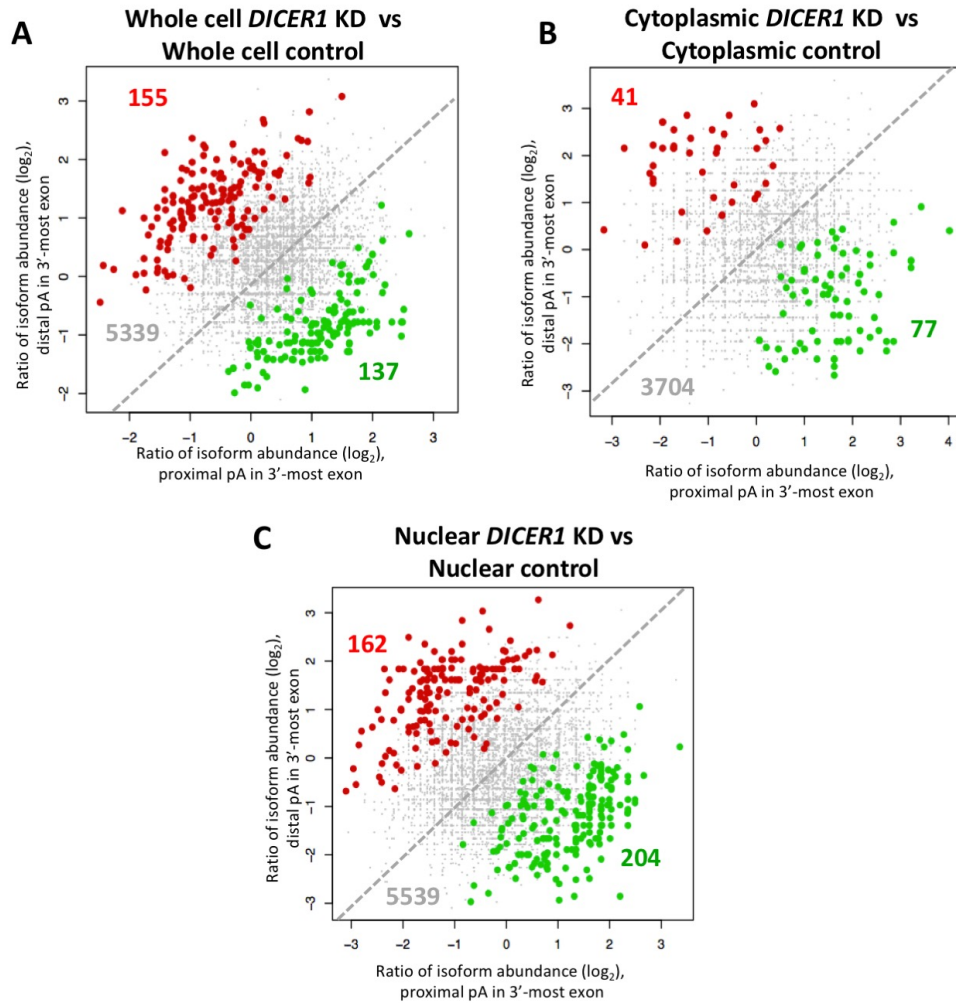


Figure 33: Effect of *DICER1*-depletion of subcellular UTR- APA profiles

Panel A - Genes with UTR-APA isoforms that show a differential representation in the *DICER1*-depleted whole-cell fraction when compared to the non-*DICER1*-depleted whole-cell fraction. Genes are highlighted in green or red if they show a significant (Fisher's exact test, $p < 0.01$) overrepresentation of the proximal or distal APA isoforms in the *DICER1*-depleted sample respectively. The number of events in each cohort is indicated on the plot.

Panel B and C – As in Panel A, but comparing the cytoplasmic fractions (B) and the nuclear fractions (C) of the *DICER1*-depleted cells, relative to the equivalent non-*DICER1*-depleted fractions.

To investigate further, both the APA profiles of cytoplasmic fractions before and after *DICER1* depletion were compared (Figure 33B), as well as the nuclear fractions (Figure 33C). In both of these comparisons, no clear pattern of shifts to either proximal or distal UTR-APA isoforms could be seen. Surprisingly, greater shifts in UTR-APA profiles

were seen when comparing the nuclear fractions before and after DICER1 depletion compared to when cytoplasmic APA profiles were compared. This pattern was consistently seen when the same comparisons were done, but when looking at CR-APA (Figure 34).

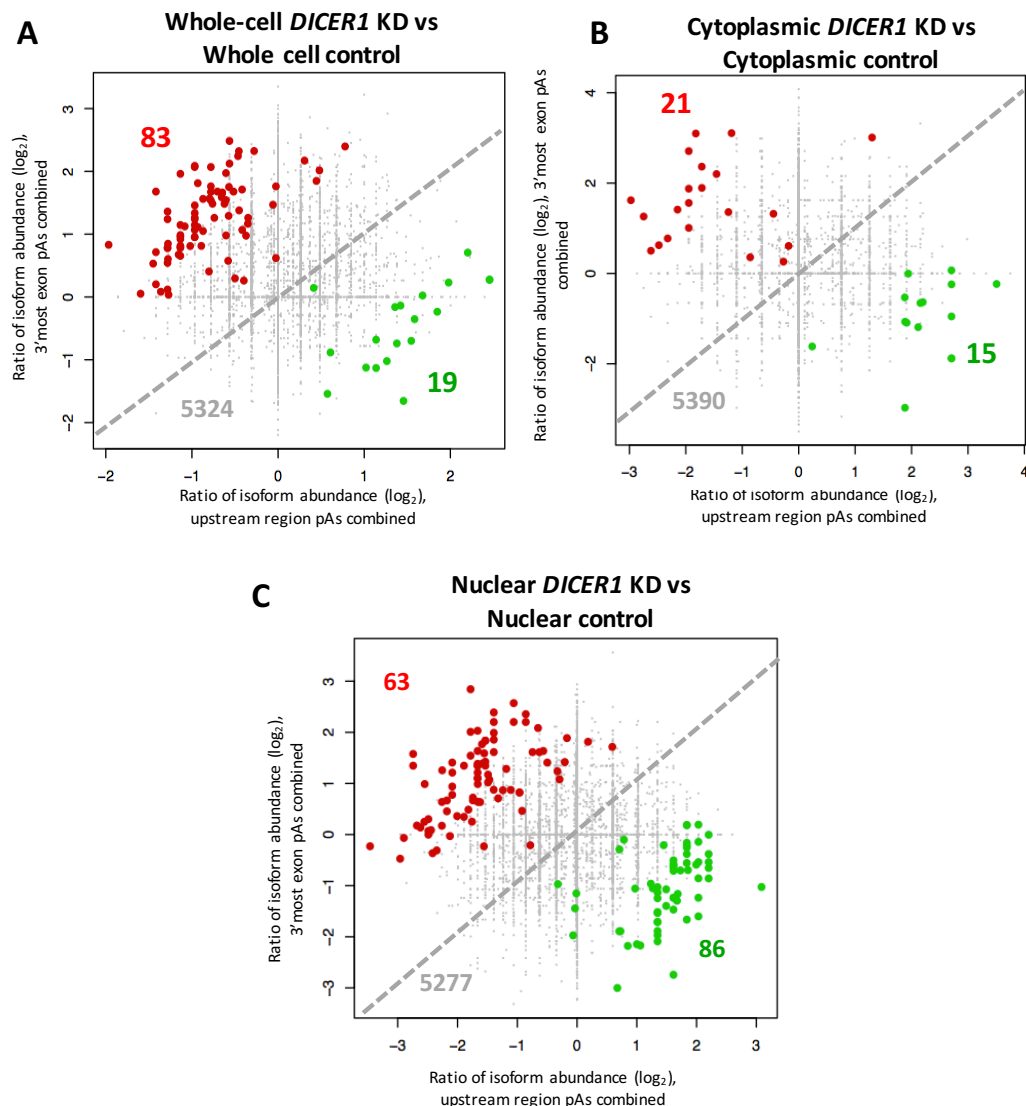


Figure 34: Effect of DICER1-depletion of subcellular CR-APA profiles

Panel A - Genes with CR-APA isoforms that show a differential representation in the DICER1-depleted whole-cell fraction when compared to the non-DICER1-depleted whole-cell fractions. Genes are highlighted in green or red if they show a significant (Fisher's exact test, $p < 0.01$) overrepresentation of the coding-region or UTR-APA isoforms in the DICER1-depleted sample respectively. The number of events in each cohort is indicated on the plot.

Panel B and C – As in Panel A, but comparing the cytoplasmic fractions (B) and the nuclear fractions (C) of the DICER1-depleted cells, relative to the equivalent non-DICER1-depleted fractions.

To ensure that this observed trend was not a function of read depth, SAAP analysis, which controls for read depth, was performed on both of the biological replicates (Figure 35). This analysis showed that for both UTR-APA (Figure 35A) and CR-APA (Figure 35B), in both biological replicates, DICER1 depletion results in more changes in the nuclear APA profiles. Despite being controlled for sequencing depth, comparisons between DICER1-depleted and non-depleted cells in replicate two consistently resulted in a higher level of differential APA isoform levels. It is unknown what caused this, but may be due to different levels of DICER1-depletion. Together, these observations are contrary to what was expected, suggesting that something additional to the reduced level of miRNA biogenesis was contributing to the shifts in APA profiles.

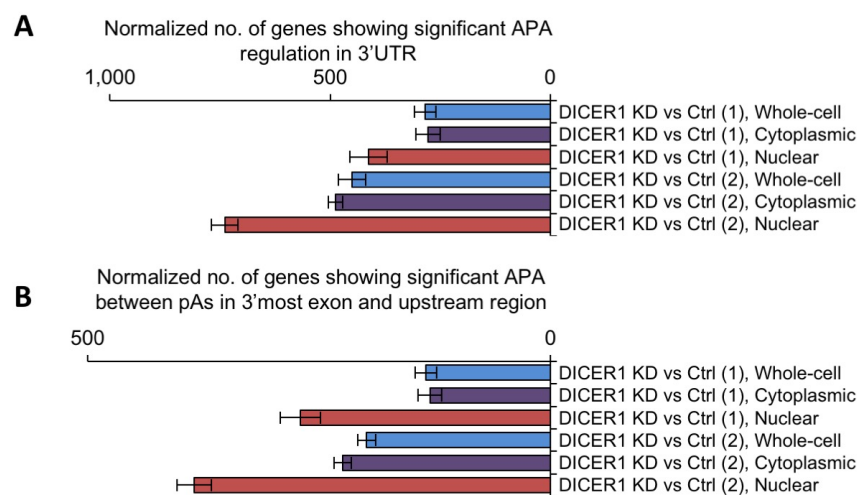


Figure 35: APA events upon DICER1-depleted cells controlled for sequencing depth

Panel A: Number of genes showing a significant differential representation of UTR-APA isoforms in the DICER1-depleted (DICER1 KD) versus non-depleted (Ctrl) subcellular fractions based on the Significance Analysis of Alternative Polyadenylation (SAAP) method with the false discovery rate controlled to less than 10%. Read depth is controlled to 1 million reads per sample. The error bar is standard deviation based on 20 values, as described in Materials and Methods. Both shortening and lengthening events are combined in this analysis.

Panel B: As in Panel A, but considering CR-APA events rather than UTR-APA events.

5.2.3. *Changes in levels of polyA factors upon DICER1 knockdown*

The first step taken was to investigate the potential secondary effects caused by DICER1 depletion that may be contributing to the altered APA profiles. These secondary effects would include, for example, the dysregulation of miRNA-regulated factors that influence pA site choice. As 3'READS can be used to accurately quantify gene expression (Hoque et al., 2014), the expression levels of factors known to influence pA site choice were compared between the DICER1-depleted and non-depleted cells (Figure 36). Only modest changes in gene expression were seen suggesting that the nuclear effects on APA profiles upon DICER1 depletion may be direct.

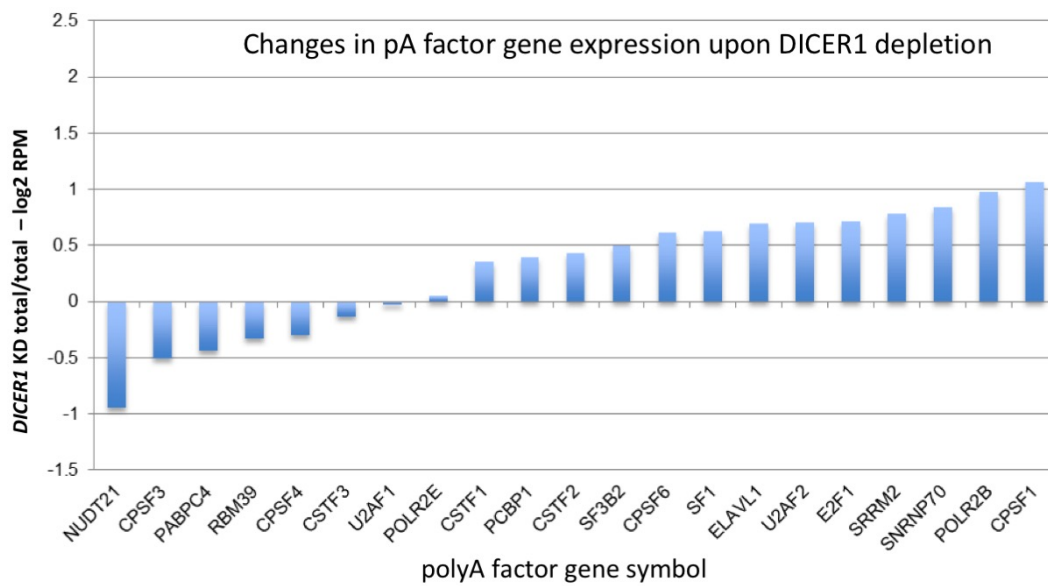


Figure 36: Changes in gene expression of known effectors of pA site choice, upon DICER1 depletion

The gene expression of genes that have previously been linked to affecting alternative polyadenylation was compared in control and *DICER1* KD whole-cells fractions. Genes were only included if at least one of the samples had greater than ten reads per million.

Beginning of collaboration with Gullerova Lab.

Disclaimer: All experiments, analysis and interpretation were performed by members of the Gullerova Lab

5.2.4. *DICER1 ChIP in the ETNK1 Locus*

Since a nuclear role for mammalian DICER1 has recently been proposed (White et al., 2014), and specifically, as having a role in transcriptional termination (Skourti-Stathaki et al., 2014), a more direct role of DICER1 in influencing pA site choice was investigated. Initially, *ETNK1* was identified as a gene that undergoes a large-scale shift in pA site choice, from the proximal to the distal pA site upon depletion of *DICER1* (Figure 37B). This particular event is a CR-APA event in which the proximal pA site is located in the intron between exons 2 and exon 3.

Firstly, to confirm the 3'READS data that showed loss of the proximal APA isoform upon DICER1 depletion, qRT-PCR was used. As before, DICER1 depletion was triggered using induction of the integrated shRNA using doxycycline. qRT-PCR before and after DICER1 depletion was performed using amplicons targeting both APA isoforms (amplicons 2 and 4), solely the proximal APA isoform (amplicon 4+6) and solely the distal APA isoform (amplicons 8 and 10), as seen in Figure 37A. This analysis confirmed that in the amplicons that would include the proximal APA isoform there is a significant drop in levels ($p < 0.05$), whereas there is no significant drop in levels were seen when using amplicons targeting solely the distal APA isoform (Figure 37C). This, therefore, corroborates the 3'READS data that DICER1 depletion results in a loss of the proximal, but not the distal APA isoform of *ETNK1*.

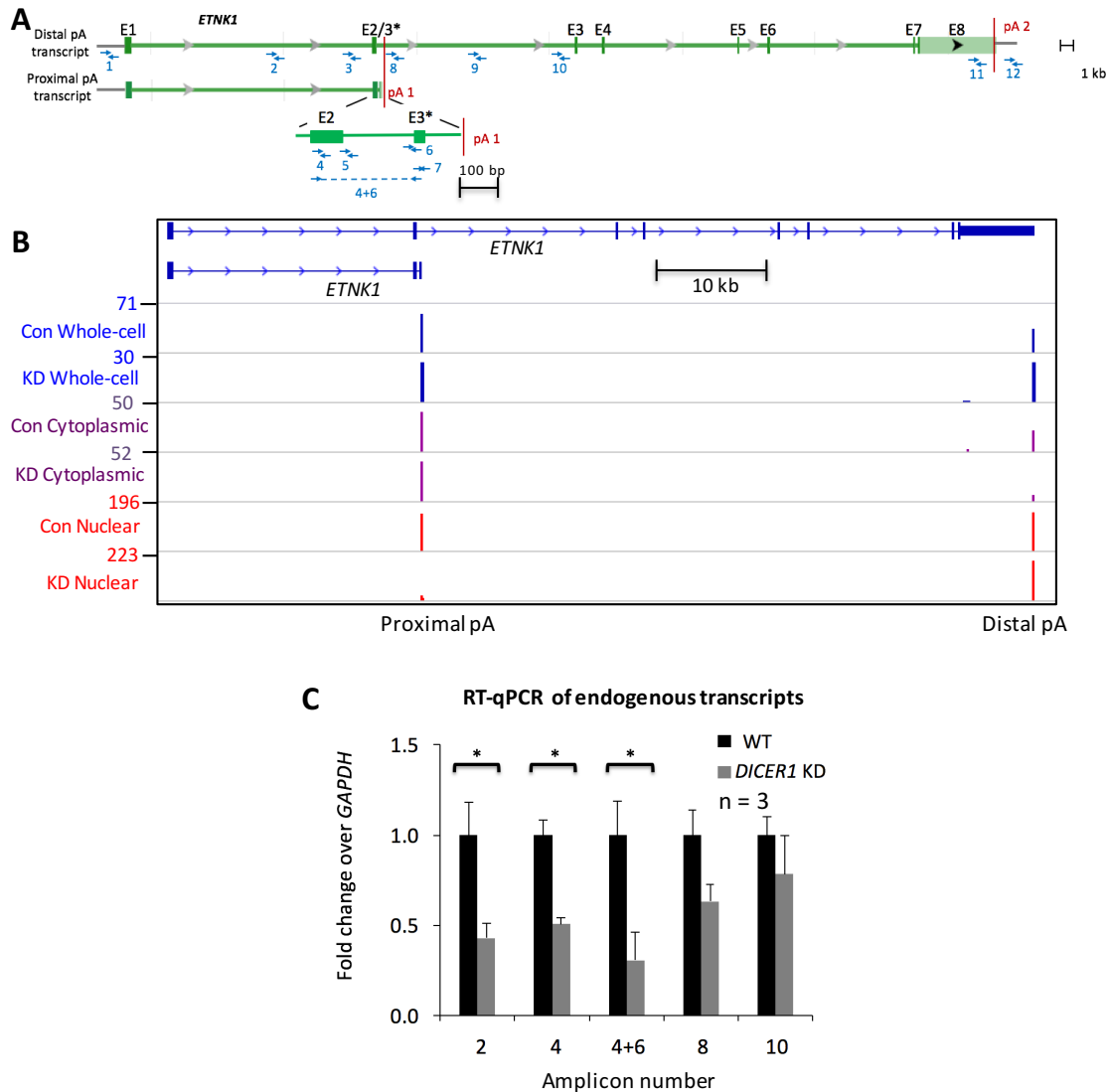


Figure 37: Confirmation of 3'READS results at the *ETNK1* locus upon *DICER1*-depletion

Panel A – Schematic showing positions of ChIP amplicons and RT-qPCR probes (1–12) on the human *ETNK1* locus around proximal and distal pA sites 1 and 2: (E3*) alternative exon 3.

Panel B: Screenshot of 3'READS results for the whole-cell, cytoplasmic and nuclear fractions before (Con) and after (KD) DICER1 depletion.

Panel C: Quantitative reverse transcriptase PCR (RT-qPCR) across the *ETNK1* locus. Signals from DICER1 depleted (DICER1 KD) and non-depleted (WT) HEK293 cells are shown as fold over wild type and normalised to *GAPDH* sense transcript levels. Mean values and standard deviations were calculated from at least three biological replicates: (*) P < 0.05 by two-tailed Student's t-test. The positions of the amplicons are shown in Panel A.

Disclaimer: All experiments, analysis and interpretation were performed by members of the Gullerova Lab.

To explore whether DICER1 was potentially localising to defined regions throughout the *ETNK1* genes ChIP experiments were used. Using an antibody targeting endogenous DICER1 (13d6), and primers targeting amplicons throughout the *ETNK1* gene an enrichment over the amplicons six and eight were found (Figure 38B). Interestingly, both of these amplicons lie within close vicinity to the *ETNK1* proximal pA site (Figure 37A).

Despite these findings, signal intensity using antibodies targeting endogenous DICER1 were low. Therefore, to further investigate the chromatin binding of DICER1, an HEK293 cell line was used that contained a doxycycline-inducible, integrated TAP-tagged *DICER1* allele (Neve et al., 2016). In addition to the TAP-tagged DICER1 allele, this cell line also expresses the doxycycline-inducible shRNA that specifically targets the endogenous *DICER1* mRNA, but not the TAP-tagged *DICER1* mRNA. Thus, when these cells are exposed to doxycycline, it results in the expression of TAP-tagged DICER1 and depletion of endogenous DICER1.

Figure 38D shows a western blot with the A-2 antibody targeting endogenous DICER1 before and after doxycycline induction of these cells shows the shift in molecular weight from the detection of the endogenous DICER1 protein in the non-induced cells to the TAP-tagged DICER1 in the induced cells. A subsequent western blot targeting the Protein A on the TAP-tagged DICER1 shows the presence of this TAP-tagged DICER1 only upon induction with doxycycline (Figure 38D).

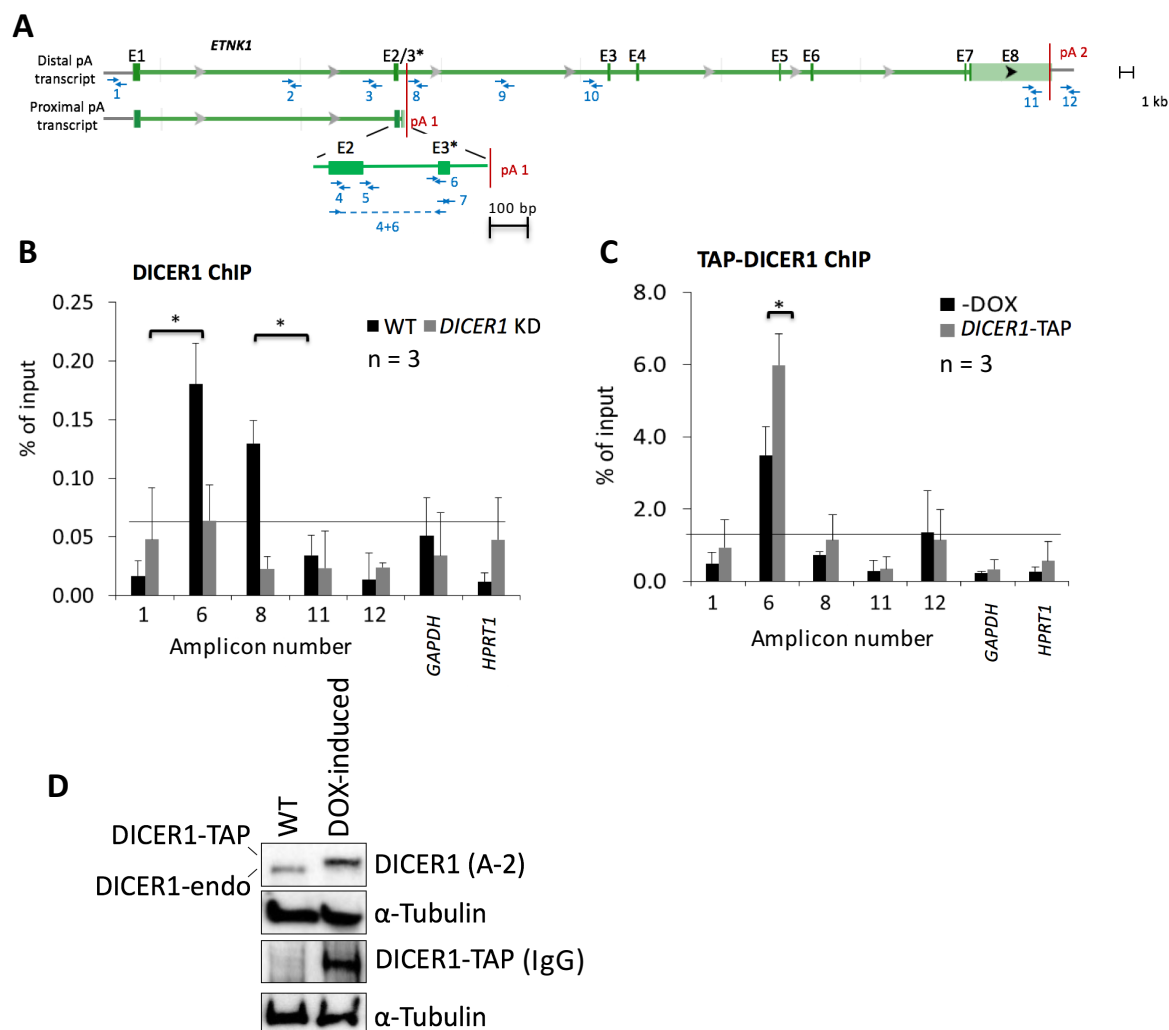


Figure 38: DICER1 localisation to the *ETNK1* locus

Panel A – Schematic showing positions of ChIP amplicons and RT-qPCR probes (1–12) on the *ETNK1* locus around proximal and distal pA sites 1 and 2 (pA 1 and 2): (E3*) alternative exon 3.

Panel B – ChIP analysis of endogenous DICER1 throughout *ETNK1* locus. Values from non-DICER1 depleted (WT) and DICER1-depleted (*DICER1* KD) HEK293 cells are shown as a % of input. Mean values and standard deviations were calculated from at least three biological replicates. $p < 0.05$ (*) by two-tailed Student's *t*-test.

Panel C - ChIP analysis of TAP-DICER1 throughout *ETNK1* locus. Values from non-induced (-DOX), and induced to express TAP-tagged DICER1 (*DICER1*-TAP) HEK293 cells are shown as a % of input. Mean values and standard deviations were calculated from at least three biological replicates. $p < 0.05$ (*) by two-tailed Student's *t*-test.

Panel D - Western blot showing reconstitution with recombinant, TAP-tagged DICER1 upon doxycycline induction (DOX-induced) and reduction of endogenous DICER1 (DICER1-endo). α -tubulin is shown as a loading control.

Disclaimer: All experiments, analysis and interpretation were performed by members of the Gullerova Lab.

Subsequent ChIP analysis using affinity purification with Sepharose IgG beads in doxycycline-induced cells showed a significant enrichment of TAP-tagged DICER1 specifically at amplicon 6 (Figure 38C). This corroborates the ChIP data targeting endogenous DICER1 with the discrepancy that DICER1 enrichment is no longer detected at amplicon 8. It is unclear what causes this discrepancy. In the absence of doxycycline induction, there still appears to be an enrichment of *DICER1* signal at amplicon 6, though to a lower degree than when induced. This is thought to be due to transcriptional leakage of the integrated TAP-tagged *DICER1* allele. Together this data indicates that there is a significant enrichment of DICER1 around the *ETNK1* proximal pA site, but not the distal pA site.

5.2.5. Double-stranded RNA at the *ETNK1* locus

The natural substrate for DICER1 is dsRNA, and recent studies have linked the recruitment of DICER1 to chromatin with the formation of dsRNA (Skourti-Stathaki et al., 2014; White et al., 2014). Therefore, to assess whether this was the case with *ETNK1*, the presence of antisense transcript throughout the *ETNK1* was investigated using strand-specific qRT-PCR, before and after DICER1 depletion (see Section 2.10). Interestingly, two of the three regions directly upstream of the proximal pA site showed a significant enrichment of antisense RNA (Figure 39B). This, therefore, raised the possibility that the *ETNK1* sense and antisense transcripts around the proximal pA site may be binding to form dsRNA, resulting in the recruitment of DICER1 to the chromatin.

If dsRNA at the *ETNK1* proximal pA site were indeed responsible for the recruitment of DICER1 to this region, it would, therefore, be expected that loss of this dsRNA would result in the loss of DICER1 recruitment. To test this hypothesis, the isolated chromatin was

subject to RNase treatment before immunoprecipitation. Interestingly, when RNase treatment was used, detection of signal from endogenous DICER1 at amplicons 6 and 8 in the *ETNK1* locus were significantly reduced ($p < 0.05$) (Figure 39C). This, therefore, indicates that the recruitment of DICER1 to this region is dependent on the presence of dsRNA, potentially formed through the hybridisation of the nascent transcript and the antisense transcripts.

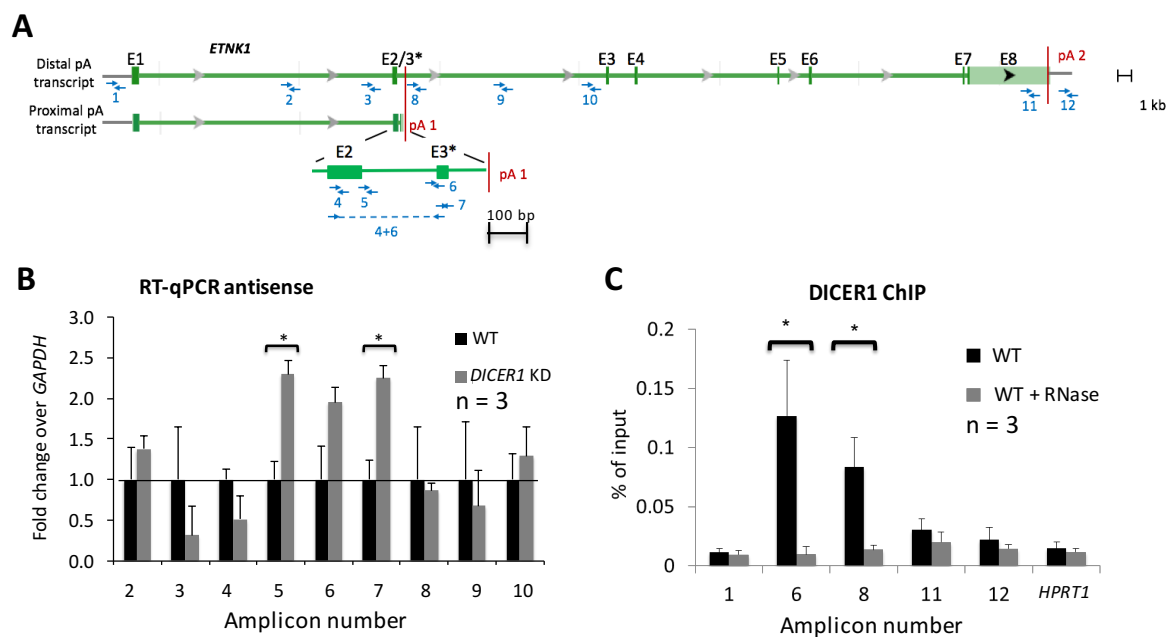


Figure 39: Dependence of DICER1 recruitment on dsRNA

Panel A – Schematic showing positions of ChIP amplicons and RT-qPCR probes (1–12) on the human *ETNK1* locus around proximal and distal pA sites 1 and 2 (pA 1 and 2): (E3*) alternative exon 3.

Panel B – Antisense RNA levels across *ETNK1* locus. Primer pairs specific to intronic and exonic antisense transcripts were used to determine relative changes in antisense RNA levels upon DICER1 KD by RT-qPCR.

Panel C – Endogenous DICER1 ChIP analysis throughout *ETNK1* locus with and without RNase treatment. Mean values and standard deviations were calculated from at least three biological replicates: (*) $P < 0.05$ by two-tailed Student's t-test.

Disclaimer: All experiments, analysis and interpretation were performed by members of the Gullerova Lab.

5.2.6. *Chromatin marks associated with DICER1 recruitment*

The processing of dsRNA into siRNA by DICER1 is discussed in Section 5.1.2. Briefly, the resulting siRNAs bind to their complementary targets, including the regions that they were derived from, resulting in the recruitment of the histone methyltransferase G9a, subsequent deposition of the repressive H3K9me2 mark and the formation of heterochromatin (Skourti-Stathaki et al., 2014). To determine if similar events were occurring in the vicinity of the *ETNK1* proximal pA site, ChIP analysis of G9a and H3K9me2 was performed. Firstly, ChIP analysis of G9a showed a strikingly similar profile to the TAP-tagged DICER1, with a significant enrichment at amplicon six relative to all other amplicons tested in the *ETNK1* locus (Figure 40B). To test whether this recruitment of G9a was dependent on the presence of DICER1, ChIP analysis using an antibody targeting G9a was performed before and after DICER1 depletion. This analysis showed that by depleting DICER1, there is a significant loss in the level of G9a localising to the amplicon six region (Figure 40B).

Subsequent ChIP analysis targeting H3K9me2 showed an enrichment of the chromatin mark over the amplicon six region relative to amplicon one, but surprisingly only reduced by a small level at amplicons downstream of the *ETNK1* proximal pA site (Figure 40C). However, this repressive chromatin mark was not significantly reduced upon the depletion of DICER1, as would be expected. It was hypothesised that even though the recruitment of histone methyltransferase that is producing this mark is reduced, the mark is too stable to see a significant enough effect to be detectable by ChIP analysis. Additionally, since G9a recruitment to the amplicon six region is only reduced, rather than completely lost, there will still be residual levels of H3K9me2 still being produced.

Therefore, to enhance this effect, BIX-01294, a chemical inhibitor of the G9a histone methyltransferase (Kubicek et al., 2007) was used. Use of BIX-01294 both before and after

DICER1 depletion showed a significant loss of H3K9me2 specifically at amplicon six (Figure 40C). A loss of signal was also seen, although to a lesser extent, at amplicon eight, and no significant change was seen at all other amplicons measured (Figure 40C). This, therefore, indicates that the abundance of H3K9me2 at amplicon six in *ETNK1* is dependent upon the activity of G9a.

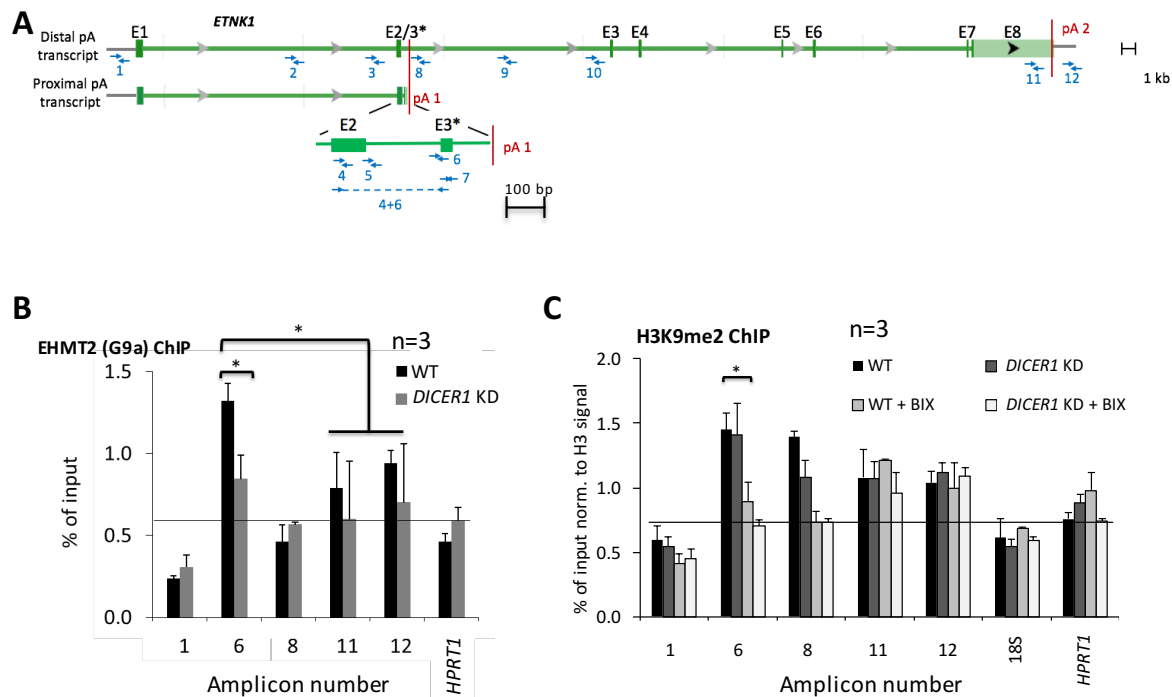


Figure 40: Chromatin modifications across *ETNK1* locus

Panel A – Schematic showing positions of ChIP amplicons and RT-qPCR probes (1–12) on the human *ETNK1* locus around proximal and distal pA sites 1 and 2 (pA 1 and 2): (E3*) alternative exon 3.

Panel B – ChIP analysis of EHMT2 (G9a) histone methyltransferase across the *ETNK1* locus in non-DICER1 depleted (WT) and DICER1-depleted (DICER KD) HEK293 cells.

Panel C – ChIP analysis of the H3K9me2 chromatin mark across the *ETNK1* locus. This was done in all combinations of non-DICER1 depleted (WT) and DICER1-depleted (DICER KD) HEK293 cells with and without treatment with the G9a inhibitor BIX. Mean values and standard deviations were calculated from at least three biological replicates: (*) $P < 0.05$ by two-tailed Student's t-test.

Disclaimer: All experiments, analysis and interpretation were performed by members of the Gullerova Lab.

Combined this data shows that there is an abundance of DICER1 at the amplicon six region in the *ETNK1* gene. Its recruitment to this region is dependent on the presence of dsRNA, the natural substrate of DICER1. Furthermore, with DICER1 recruitment, recruitment of the histone methyltransferase G9a is also seen, which itself can affect the levels of H3K9me2. From this data, it was therefore hypothesised that the loss or depletion of DICER1 may result in a drop in both levels of both G9a and H3K9me2 in the region surrounding *ETNK1*, thereby reducing the repressive nature of the surrounding chromatin.

This potential opening up of the surrounding chromatin could allow RNAPII to pass through this region more freely, reducing the probability of the transcripts being cleaved and polyadenylated at the proximal pA site. This mechanism of DICER1 influencing cleavage and polyadenylation, and thus termination in the at the proximal pA site in *ETNK1* appears to be very similar to that found in the termination of the β -actin gene (Skourti-Stathaki et al., 2014).

5.2.7. *Dependence of pA site selection in ETNK1 on RNAPII processivity*

To further investigate the proposed mechanism of DICER1 influencing pA site choice, the ChIP analysis was used to assess the RNAPII profiles throughout the *ETNK1* locus before and after DICER1 depletion. In non-DICER1-depleted conditions, RNAPII levels are higher around the promoter region and are then generally consistent throughout the rest of the gene. Subsequently, upon DICER1 depletion, the most significant loss of RNAPII occupancy is seen at amplicon 6, with a reduction also seen at amplicon 8 (Figure 41B). This would fit with the model that DICER1 depletion is resulting in a more relaxed chromatin conformation surrounding the *ETNK1* proximal pA site, allowing RNAPII to move through this region faster, resulting in a lower RNAPII occupancy, and a lower probability of using the proximal pA site.

If the pA site choice were indeed dependent on RNAPII processivity, it would be expected that a slow RNAPII would result in the increase in the abundance of the proximal APA isoform. To test this, inducible DICER1 knockdown cells were either transfected with a plasmid expressing an α -amanitin-resistant slow C4 mutant RNAPII Rpb1 subunit or a plasmid expressing an α -amanitin-resistant RNAPII Rpb1 subunit with wild-type (WT) processivity levels (de la Mata et al., 2003). After the transfection with these plasmids, endogenous RNAPII was suppressed by treatment with α -amanitin. To confirm the expression of the recombinant Rpb1 subunits, western blotting targeting Rpb1 was used on mock-transfected, C4-transfected and WT transfected cells after α -amanitin treatment. As α -amanitin treatment results in the degradation of the Rpb1 subunit (Van Nguyen et al., 1996), very little Rpb1 was detected in both mock transfected DICER1-depleted and non-depleted cells (Figure 41D). However, the western blots in C4 and WT Rpb1 transfected cells confirmed the expression of these recombinant subunits in both DICER1-depleted and non-depleted cells (Figure 41D).

(Figure on the following page)

Figure 41: RNA Polymerase II occupancy across *ETNK1* locus

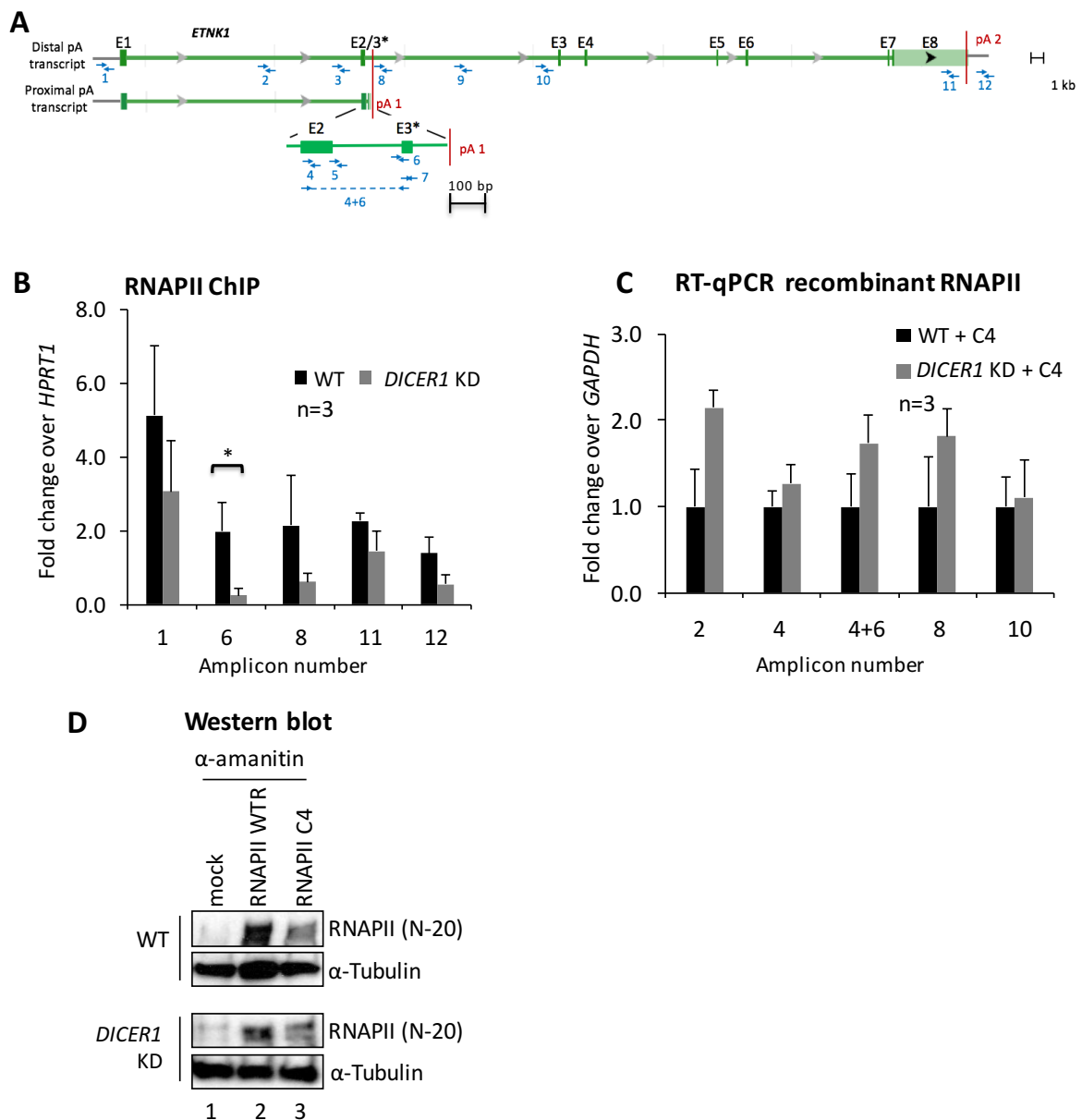
Panel A – Schematic showing positions of ChIP amplicons and RT-qPCR probes (1–12) on the human *ETNK1* locus around proximal and distal pA sites 1 and 2 (pA 1 and 2): (E3*) alternative exon 3.

Panel B –RNAPII occupancy across the *ETNK1* locus. The percentage of input values from non-DICER1-depleted (WT) and DICER1-depleted (DICER1 KD) HEK293 cells are shown as fold change over HPRT1.

Panel C – RT-qPCR analysis of sense transcripts across the *ETNK1* locus upon reconstitution of transcription with either the slow elongating C4 RNAPII mutant or the recombinant wild-type RNAPII. Signals from non-DICER1-depleted (WT) and DICER1-depleted (DICER1 KD) HEK293 cells are shown as fold change over wild type recombinant and normalised to *GAPDH* sense transcript levels. Mean values and standard deviations were calculated from at least three biological replicates: (*) $P < 0.05$ by two-tailed Student's t-test.

Panel D - Western blot showing expression levels of recombinant, α -amanitin resistant RNAPII mutant WTR (recombinant wild type) and C4 (slow elongating mutant) in wild-type and DICER1-depleted (DICER1 KD) HEK293 cells

Disclaimer: All experiments, analysis and interpretation were performed by members of the Gullerova Lab.



Subsequently, RNA was isolated from these cells and subject to qRT-PCR analysis with primers targeting the *ETNK1* gene. This analysis shows that the loss of the *ETNK1* proximal isoform upon DICER1 depletion that was originally seen (Figure 37C) is no longer observed (Figure 41C). This indicates that through the use of an RNAPII with slow processivity, the effects on *ETNK1* pA site choice upon DICER1 deletion are averted.

Together, this supports the model that DICER1 depletion is causing more relaxed chromatin conformation surrounding the *ETNK1* proximal pA site, thus allowing RNAPII to pass through this region faster and reduce the probability of cleaving and polyadenylating at this proximal pA site (summarised in Figure 42). With the slow RNAPII, this pattern is not seen as, despite the relaxed chromatin state, the mutations within Rpb1 do not allow increased processivity through this region.

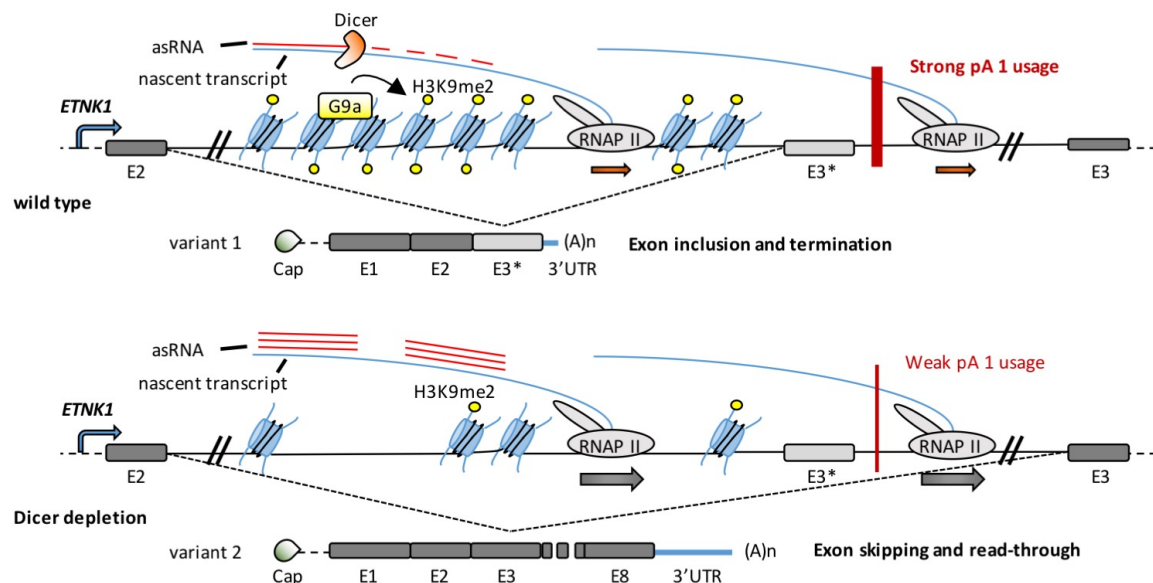


Figure 42: Model for DICER1-dependent termination at the proximal pA site of *ETNK1*

In non-DICER1-depleted cells, antisense RNA complementary to nascent *ETNK1* transcript causes the formation of double-stranded RNA (dsRNA), which is processed by DICER1. Close proximity of DICER1 with RNAPII or processing of dsRNA by DICER1 into endogenous small interfering RNA (endo-siRNA) may recruit histone 3 lysine 9 methyltransferase G9a and cause the formation of facultative heterochromatin. Chromatin condensation impairs RNAPII elongation, thereby favouring the production of *ETNK1* variant 1 by usage of the proximal pA site, the inclusion of alternative exon 3 (E3*) and termination. Upon DICER1 depletion, non-processed antisense RNA around the proximal pA site accumulates. The absence of DICER1 prevents the formation of facultative heterochromatin, which favours RNAPII elongation, promotes read-through at the proximal pA site and skipping of E3* in *ETNK1* variant 2.

Disclaimer: All experiments, analysis and interpretation were performed by members of the Gullerova Lab.

End of collaboration with Gullerova Lab

5.2.8. *Effect of Dicer1 ablation in mouse ES cells*

The described nuclear effect of DICER1 in HEK293 cells was an unexpected, albeit interesting, finding. To further investigate whether this nuclear role of DICER1 was a general mechanism, previous RNA-seq data produced from mouse ES cells before and after *Dicer1* ablation was analysed using DaPars to examine the APA profiles. The data used for this analysis was performed as part of a study looking at the regulation of miRNA levels via lncRNAs (Tan et al., 2015). This study utilised a conditional *Dicer1* knockout system (Nesterova et al., 2008). This is composed of a mouse ES cell line that contains a tamoxifen-inducible *Cre*-recombinase, which drives recombination between the *loxP* sites that flank the RNase III domain of *Dicer1* (Nesterova et al., 2008).

To look at the effects of lncRNA levels upon *Dicer1* ablation, and thus loss of miRNA, the nuclear and cytoplasmic RNA fractions were subject to full polyA⁺ RNA sequencing (Tan et al., 2015). This was done before *Dicer1* ablation and 12 days post-ablation. Two biological replicates from each of condition were combined and analysed and read depths were controlled to 50 million reads.

Analysis of these APA landscapes in each of these samples was performed using the DaPars analysis as previously described in Section 4.2.4. From this analysis, it is once again seen that, upon *Dicer1* depletion, the nuclear APA profile is more affected than that of the cytoplasmic (Figure 43A). *Ddx52* is shown as an example of one of these nuclear changes where there appears to be a shift to the distal pA site upon *Dicer1* ablation, and the change in pA site choice also penetrates through to the cytoplasmic profiles (Figure 43B). Combined, this data indicates that DICER1 indeed plays a nuclear role in both mouse and human cells and appears to be altering APA profiles through influencing the chromatin conformation at specific loci, thereby influencing pA site choice.

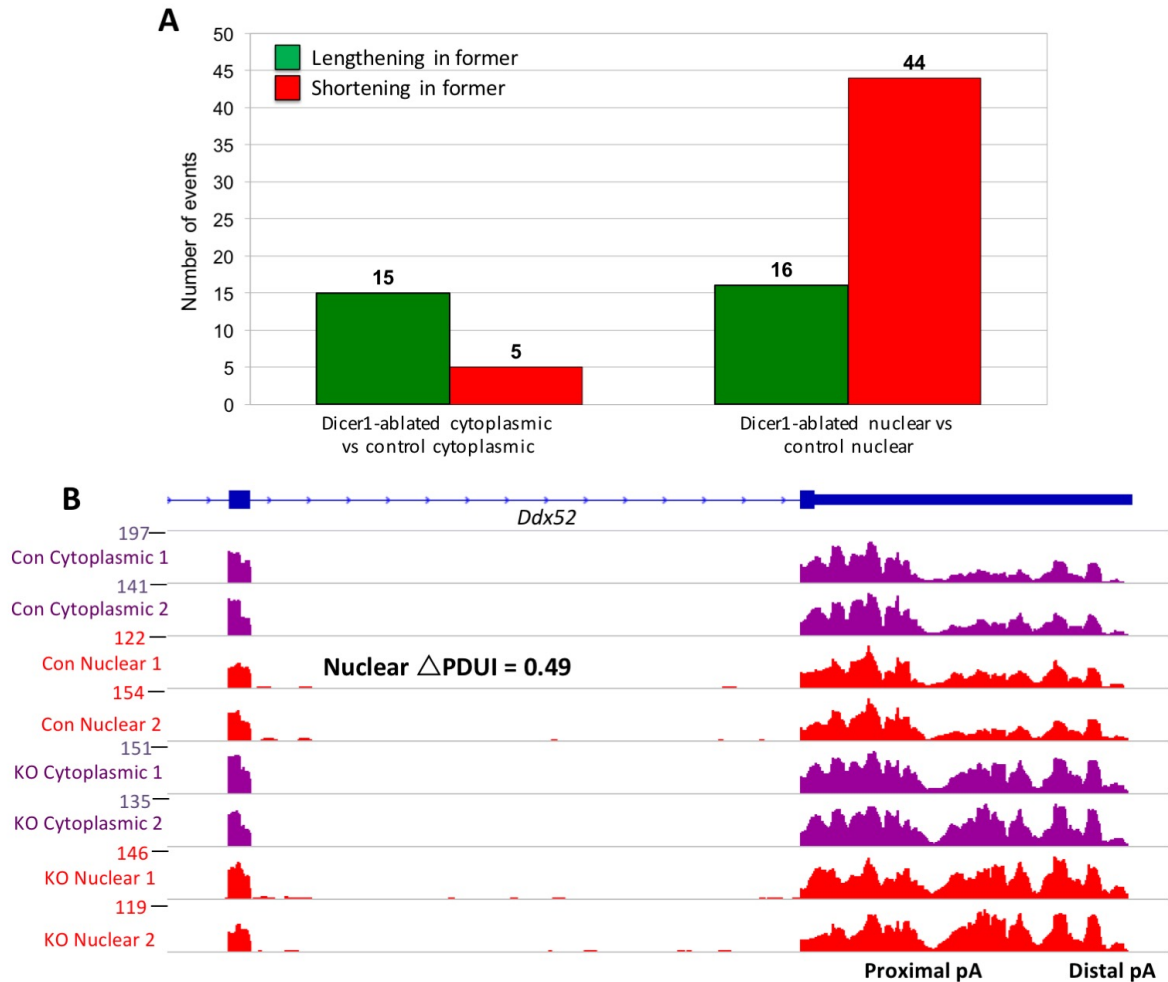


Figure 43: APA analysis of mouse RNA-seq data before and after Dicer1 ablation

Panel A – Number of APA events, as detected by DaPars analysis, when comparing subcellular RNA fractions before or 12 days after genetic Dicer1 ablation in mouse ES cells. Events which show a higher representation of the distal or proximal APA isoform in the Dicer1-ablated samples are shown in green and red respectively.

Panel B – Screenshot of *Ddx52* as an example of a gene that undergoes a shift in nuclear APA profiles upon Dicer1 ablation (KO), relative to the equivalent sample before ablation (Con). The change in the percentage distal usage index (Δ PDUI) between nuclear samples before and after Dicer1 ablation is also shown.

5.3. DISCUSSION

5.3.1. *Differentially regulated APA isoforms mediated by miRNAs*

After the finding that there were a significant number of APA events which resulted in differentially regulated APA isoforms in Chapter 4, the question then became, what are the mechanisms that govern the differential regulation? Sections 4.2.8 and 4.2.9 address this to some extent, and do show some miRNAs and RNA-binding proteins that have enriched target sites in the aUTRs of differentially regulated APA isoforms.

However, to address the extent of miRNA-mediated differential regulation of APA isoforms directly, DICER1 depletion was used as a method to compromise miRNA biogenesis. Upon DICER1 depletion, it would be expected that distal APA isoforms that are being targeted for degradation would recover in the cytoplasm. However, during this initial experimental design and execution, the nuclear role of DICER1 was largely unknown. From this study (Neve et al., 2016), and others (Skourti-Stathaki et al., 2014; White et al., 2014) it is now clear that DICER1 is playing a significant role at the chromatin level. Since this is the case, depletion of DICER1 can have two separate mechanisms of altering the APA profiles:

1. By inhibiting miRNA biogenesis removing a layer of differential APA isoform regulation;
2. By altering the chromatin state surrounding specific pA sites, altering the probability that that pA site will be utilised.

Regrettably, this makes it very hard to pinpoint what is causing the shifts in the cytoplasmic APA profiles, as they may be a result of a change in miRNA levels, or a change in pA site choice that has penetrated through to the cytoplasm. Conversely, nuclear shifts in APA profiles are almost certain to come from a shift in pA site choice, as the general migration of RNA from the nucleus to the cytoplasm means that any shifts in APA profiles that occur in the cytoplasm will be unlikely to directly affect the nuclear APA profile.

5.3.2. *Nuclear function of DICER1*

Our investigation of the nuclear role of DICER1 centred around ChIP analysis of the *ETNK1* gene. This showed an enrichment of DICER1 around the proximal pA site, along with the histone methyltransferase, G9a, and repressive chromatin marks, H3K9me2. Together, this, with other studies (Skourti-Stathaki et al., 2014) has implicated nuclear DICER1 to a role in transcriptional termination. If DICER1 is localising to this subset of transcriptional termination regions, forming more compact chromatin and promoting termination at a nearby pA site, it is somewhat unsurprising that its depletion would result in a shift in the pA site choice. Depletion would, therefore, be expected to allow a transcribing RNAPII past this point with much more ease, reducing the probability of using that pA site, as seen with the proximal pA site of *ETNK1*.

ETNK1 is shown as an example of a shift in nuclear APA profile from a proximal to a distal pA site upon DICER1 depletion (Figure 37B). As happens with *ETNK1*, this would be expected when DICER1 depletion causes a more open chromatin conformation around the proximal pA site increasing the probability of the RNAPII progressing to the distal pA site, resulting in a shift to the distal pA site.

However, when looking at the shifts in nuclear APA profiles upon DICER1 depletion, both shifts to proximal and distal pA sites are seen. A potential explanation for this is that if the DICER1-dependent termination sites are the distal pA sites, rather than the proximal one seen with *ETNK1*, this could result in read-through transcripts which are never cleaved and polyadenylated. This will result in the loss of the distal APA isoforms and the read-through transcripts will never be detected by 3'READS. This would present as a loss of the distal APA isoform and therefore a relative increase in the proximal APA isoform, and therefore be classified as a “shortening” event when comparing the DICER1-depleted nuclear RNA to non-depleted nuclear RNA (Figure 33C).

The involvement of DICER1 in regulating chromatin is not a novel finding. In fission yeast, the RNA transcribed from the centromeric regions acts as a template for the RNA-dependent RNA polymerase complex which forms dsRNA which is subsequently degraded by Dicer into siRNAs. These siRNAs then target the complementary sequences in the centromeric regions and result in the RISC-dependent heterochromatin formation (Colmenares et al., 2007). In the case of RNAi in fission yeast, this is through the CLR4 histone methyltransferase (homologue of G9a) and the spreading of the repressive H3K9 methylation (Volpe et al., 2002).

It would be expected that this mechanism could be targeted to any chromatin region through the specific formation of dsRNA. In the case of *ETNK1*, this appears to be targeted by antisense transcription in the region surrounding the proximal pA site, thus, in non-depleted conditions, increasing the compactness of the chromatin and hindering RNAPII as it passes, therefore resulting in proximal pA site usage. Depletion of DICER1 thereby reduces processing of the dsRNA, resulting in its accumulation, and ultimately a more open chromatin complex allowing RNAPII to pass freely, thereby favouring the usage of the distal pA site (Figure 42).

5.3.3. *Influence of DICER1 on pA site choice*

The fact that the depletion of DICER1 influences pA site choice by no means demonstrates that it is a physiological regulator of APA. Much like some of the core pA factors that have been proposed to regulate pA site choice (Table 2), it is hard to imagine a mechanism by which the cell would regulate the expression levels of DICER1 to control the level of proximal pA site usage in *ETNK1*. However, this does not mean that the presence of DICER1 at these sites, and its effect on pA site choice is not significant. Understanding the factors that influence pA site usage, even if they do not actively regulate it, is key to fully

understanding the mechanisms by which APA profiles are established as well as the physiological impact of APA as a whole.

5.3.4. *Conclusions and future directions*

Despite the failure to identify APA isoforms that were differentially regulated by miRNAs, the finding that DICER1 influences pA site choice is a valuable one. To establish the regulatory pathways which are causing differential regulation of APA isoforms, an alternative approach will have to be adopted. For example, a more defined approach may be better suited, such as the ablation of single miRNAs and RNA-binding proteins that have been identified through the motif analysis approaches seen in Sections 4.2.8 and 4.2.9.

6. CHAPTER 6 – SUBCELLULAR APA **ISOFORM ANALYSIS IN** **COLORECTAL CANCER CELL LINES**

6.1. INTRODUCTION

Chapter 4 demonstrated that subcellular fractionation is an effective and novel approach to isolate APA isoforms that are differentially regulated, either in their cytoplasmic stability or nuclear export rates. This approach was applied to nine different cell lines, the majority of which showed a consistent pattern of the proximal APA isoform being overrepresented in the cytoplasmic fractions. While the application of this subcellular fractionation approach to individual cell lines can isolate individual APA events that are differentially regulated in that cell line, a dynamic system is required to assess the contribution of changes in pA site choice to altered gene expression.

In a dynamic system, such as cancer progression, there are two mechanisms by which APA could effectively regulate gene expression:

1. There is a regulated switch between pA sites in response to a specific signal or conditions, thereby switching APA isoforms. These APA isoforms then need to be differentially regulated at the point of posttranscriptional regulation.
2. There is a change in the expression of a miRNA or an RNA-binding protein, which will then result in the introduction of a layer of differential regulation of APA isoforms. In this situation, the actual pA site choice does not have to change.

Either of these cases will result in an altered level of protein output without altering the transcriptional rate of these genes. To address this, this chapter focuses on applying the

subcellular fractionation approach to identify genes in colorectal cancer progression which fall into either of these categories.

6.1.1. *Cancerous versus non-cancerous colorectal cell lines*

The most well-characterised global shifts of APA profiles have been seen in cancerous versus non-cancerous cells (as discussed in Section 1.3.4). One of these studies involved the characterisation of APA profiles in colorectal cancer patients where, similar to other cancer subtypes, 3'UTR shortening was generally observed upon carcinogenesis (Morris et al., 2012). Specifically, Morris et. al. observed 3'UTR shortening from adenoma to carcinoma progression, whereas the progression from normal mucosa to adenoma, there was no specific shifts to lengthen or shorten 3'UTRs (Morris et al., 2012). On an individual gene level, APA has been proposed to play a role in the post-transcriptional regulation of *COX-2*, a gene which is key in the synthesis of prostaglandin (Young and Dixon, 2010) and *COX-2* overexpression is seen at multiple stages of colorectal cancer carcinogenesis (Wang et al., 2005).

Combined, this data suggested that APA in colorectal cancer may be playing a role in colorectal carcinogenesis. While Morris et al. have shown APA profile changes throughout colorectal cancer progression, they were unable to determine whether these changes have any impact on shaping the gene expression profiles or if they are contributing to carcinogenesis

. Thus, the subcellular fractionation approach was employed to evaluate whether APA, in response to the transition from normal to cancerous colon cells, results in APA mRNA isoforms that are differentially regulated.

Previous studies looking at APA in cancer progression have done so using cell line-based approaches, comparing a cancerous cell line (e.g. Mcf7) with a cultured human “normal” cell line (e.g. Mcf10A) (Fu et al., 2011; Mayr and Bartel, 2009). A drawback of

using cell line based models is that they may not be fully representative of the disease state. In the context of colon cancer however, a study profiling 70 colorectal cancer cell lines, including gene expression, mutations, gene copy number and SNP analysis, showed that human colorectal cancer lines are highly representative of primary tumours and are suitable for the study of colorectal cancer biology (Mouradov et al., 2014). It was therefore decided to adopt a cell line-based approach.

The cell line chosen as a “normal” non-malignant adult-derived human male colonic epithelial cell was the 1CT cell line (Roig et al., 2010). This cell line was immortalised with the non-oncogenic proteins cyclin-dependent kinase 4 (CDK4), allowing the cells to bypass the normal cell culture associated stresses that can lead to senescence, and human telomerase (hTERT), thereby allowing telomeres to be maintained and preventing replicative senescence (Roig et al., 2010). Characterisation of the immortalised 1CT cell line has shown that ~98% of cells display the colon epithelial cell specific marker A33 (Roig et al., 2010).

The cancerous cell lines chosen were the SW480 and SW620 cell lines which were both derived from the same patient: a 51-year old Caucasian male. The SW480 cell line was established from a Dukes’ type B primary adenocarcinoma of the colon and the SW620 cell line was derived from a lymph node after cancer recurred with widespread metastasis (Leibovitz et al., 1976). Since the establishment of these cell lines (Leibovitz et al., 1976), they have undergone long-term cell culture *in vitro*. To confirm that SW480s and SW620s still represent a valid model for studying colorectal cancer metastasis the phenotypic differences between them, such as comparative levels of tumorigenicity, were studied (Hewitt et al., 2000). From this, it was concluded that the long-term *in vitro* cell culture did not alter the phenotypic characteristics that make SW480s and SW620s an ideal model to study the late stages of colorectal cancer progression. The comparison of the 1CT, SW480 and SW620 cell lines, therefore, provide a rare model in which changes in APA profiles throughout the onset and metastases process of colorectal cancer can be evaluated.

6.1.2. *3'end sequencing approach*

This chapter switches the method of 3'end-sequencing, primarily due to a commercially available kit becoming available shortly before the work for this chapter began. The kit used is the QuantSeq 3'mRNA-Seq library kit. This uses an oligo(dT)-based priming approach (Section 1.9.4) rather than the polyA+ capture approach used in 3'READS. Advantages of the QuantSeq kit include a simpler protocol (Figure 5) as well as a much lower level of required starting material (500ng).

Previously, to assess the quantitative abilities of QuantSeq, libraries were prepared from two separate External RNA Controls Consortium (ERCC) RNA spike-in controls and sequenced on the Ion Proton System (Moll et al., 2014). QuantSeq showed a high input to output correlation for two separate ERCC mixes, demonstrating a very high accuracy in gene expression quantification, with a Spearman Correlation of 0.986.

However, the main drawback is that the isolation of pA sites is solely based on the use of an oligo(dT) primer in reverse transcription. Therefore, any RNA containing a genomically encoded sequence that has a stretch of A residues can potentially be reverse transcribed and penetrate through into the final sequencing data. However, this drawback is circumvented here by removing these internal priming events bioinformatically by using the Samtools Intersect function. This removes all sequenced reads that do not partially overlap a 100 nucleotide window upstream of a previously identified pA site that has been verified using either 3'READS here or PolyA-seq (Derti et al., 2012).

6.2. RESULTS

6.2.1. *Subcellular fractionation of 1CT, SW480 and SW620 cells*

As significant variations in subcellular fractionation purity can be seen between cell types, mainly due to different membrane compositions (Holden and Horton, 2009), the

fractionation controls were performed for two biological replicates of 1CTs, SW480s and SW620s. The RNA gels for replicate one are shown in Figure 44, demonstrating a clean subcellular RNA fractionation for each of these cell lines.

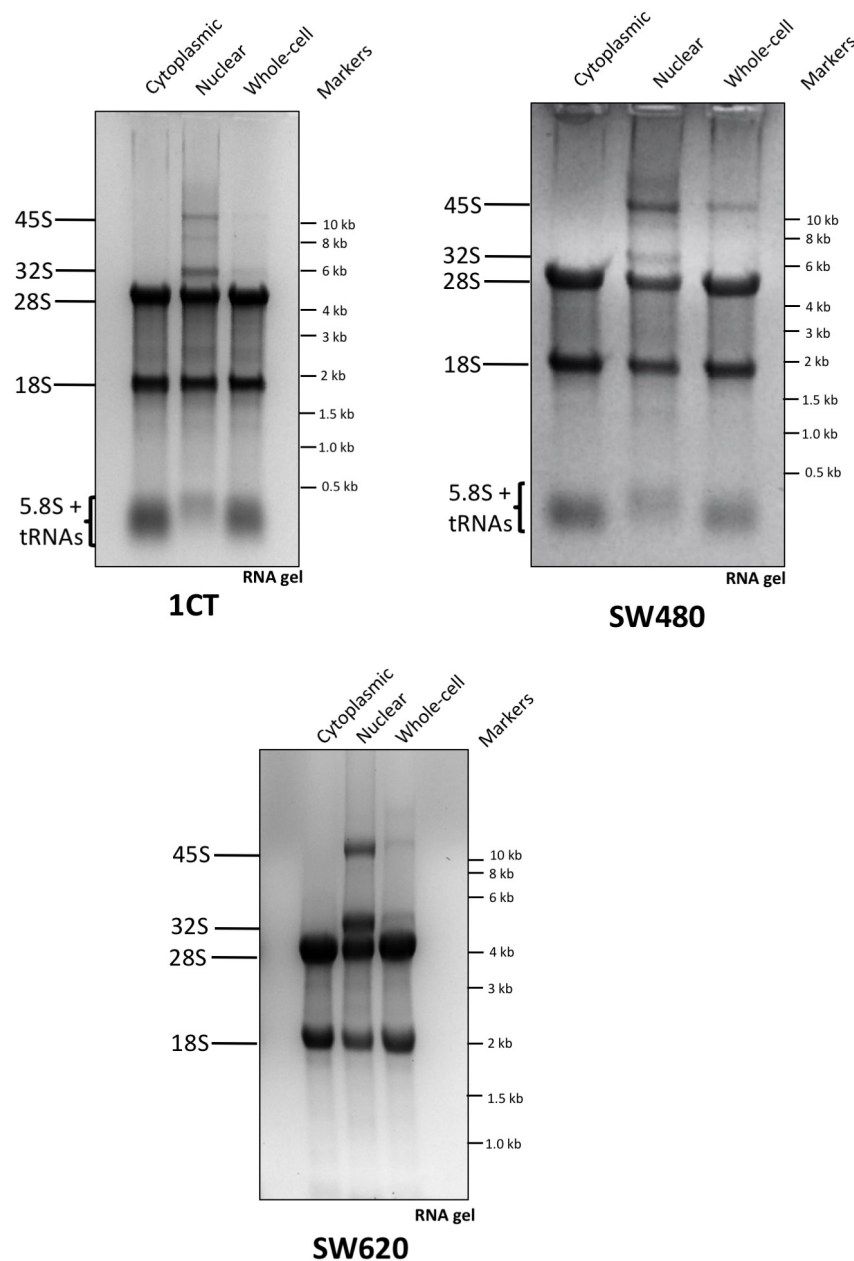


Figure 44: Subcellular fractionation controls of 3 colorectal cell lines

Each panel shows a denaturing RNA gel from each subcellular RNA fraction of the three different colorectal cell lines used in this study: 1CT cells (Roig et al., 2010), SW480 cells and SW620 cells. The identity of each ribosomal RNA band is shown on the left-hand side of the gel with size markers on the right-hand side of the gel.

Subsequently, the cytoplasmic and nuclear RNA samples were used as input for sequencing library preparation with the QuantSeq kit and sequencing on the Ion Proton system. Sequencing produced between 2.8 and 7.4 million reads per sample (Appendix Table 3) and these were then filtered solely for reads that overlapped with previously identified pA sites (see Figure 45 and Section 2.15). The resulting reads were then passed through the same APA analysis pipeline as in Section 2.13, with the modification that reads did not require any non-genomically encoded adenosines to be considered for analysis. This was repeated for two biological replicates of 1CTs, SW480s and SW620 nuclear and cytoplasmic RNA fractions. Using the gene expression data from this sequencing data confirmed an efficient fractionation in all isolations with the nuclear-retained lncRNA *MALAT1* showing a minimum of 380-fold enrichment in the nuclear fraction relative to the corresponding cytoplasmic fraction.

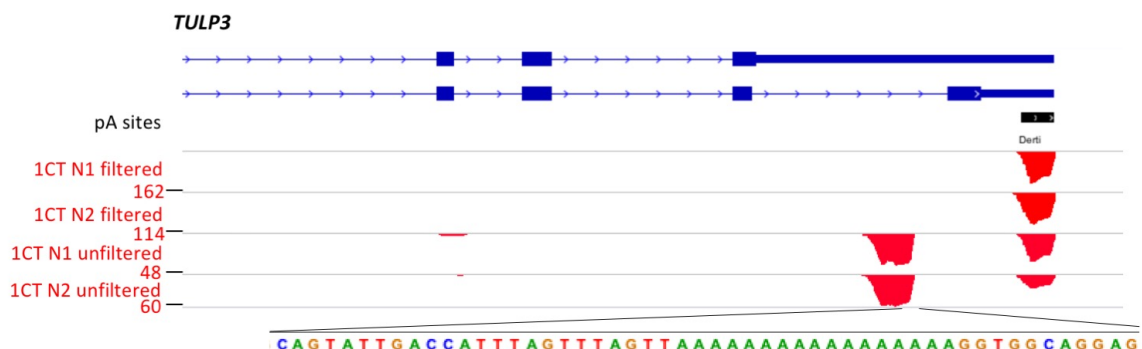


Figure 45: Isolation of reads mapping to true pA sites

A comparison of Bigwig files at the *TULP3* locus generated from BAM files before and after being filtered for internal priming events. Filtering was performed using the SAMTools Intersect function which removed all reads that did not overlap with a region 100 nt directly upstream of a previously identified pA site. These regions are shown in the track directly under the schematic of the *TULP3* gene. The sequence region at the end of the additional peak in the unfiltered sample has been expanded and shows a long stretch of genomically encoded adenosines indicating that these reads are likely produced as a result of internal priming. As these reads do not overlap with any previously identified pA site, these reads are filtered out and not used in the APA analysis.

6.2.2. *1CTs, SW480s and SW620s as a model for colorectal cancer*

Firstly, to ensure that the biological replicates were consistent and reproducible, the Pearson correlation of the calculated gene expression values were determined. As discussed in Materials & Methods (Section 2.13), this is calculated for each sample as the sum of all pA sites mapping to each gene. The correlation matrix is shown in Figure 46A and shows a minimum correlation of 0.946 for direct biological replicates. As a measure of the similarity of the gene expression profiles of the subcellular fractions of 1CTs, SW480s and SW620s, principal component analysis was performed using the FactoMineR R package (Figure 46B). This shows tight clustering of the biological replicates, as well as highlighting that the different samples do display unique gene expression profiles.

In Section 6.1.1, previous studies that included SW480s, SW620s and 1CTs as a model for investigating colorectal cancer were discussed. However, to ensure that the key gene expression biomarkers for colorectal cancer were present in the SW480s and SW620s relative to the 1CTs, the 3'end sequencing data from the cytoplasmic samples was used to compare steady-state cytoplasmic RNA levels. Firstly, the top 50 genes that showed the most significant variation between the cytoplasmic samples were isolated. These top 50 were defined as those the lowest average p-value (Fisher's exact test) for a difference in gene expression across all comparisons between the cytoplasmic samples. The $\log_2(\text{RPM}+1)$ values were calculated for each of these genes and are shown on the heat map in Figure 47. Clustering of the samples shows that, based on the steady-state gene expression levels of these 50 genes, it is the 1CT cells that are the most distinct, with the SW480 and SW620 samples showing the higher similarity to each other.

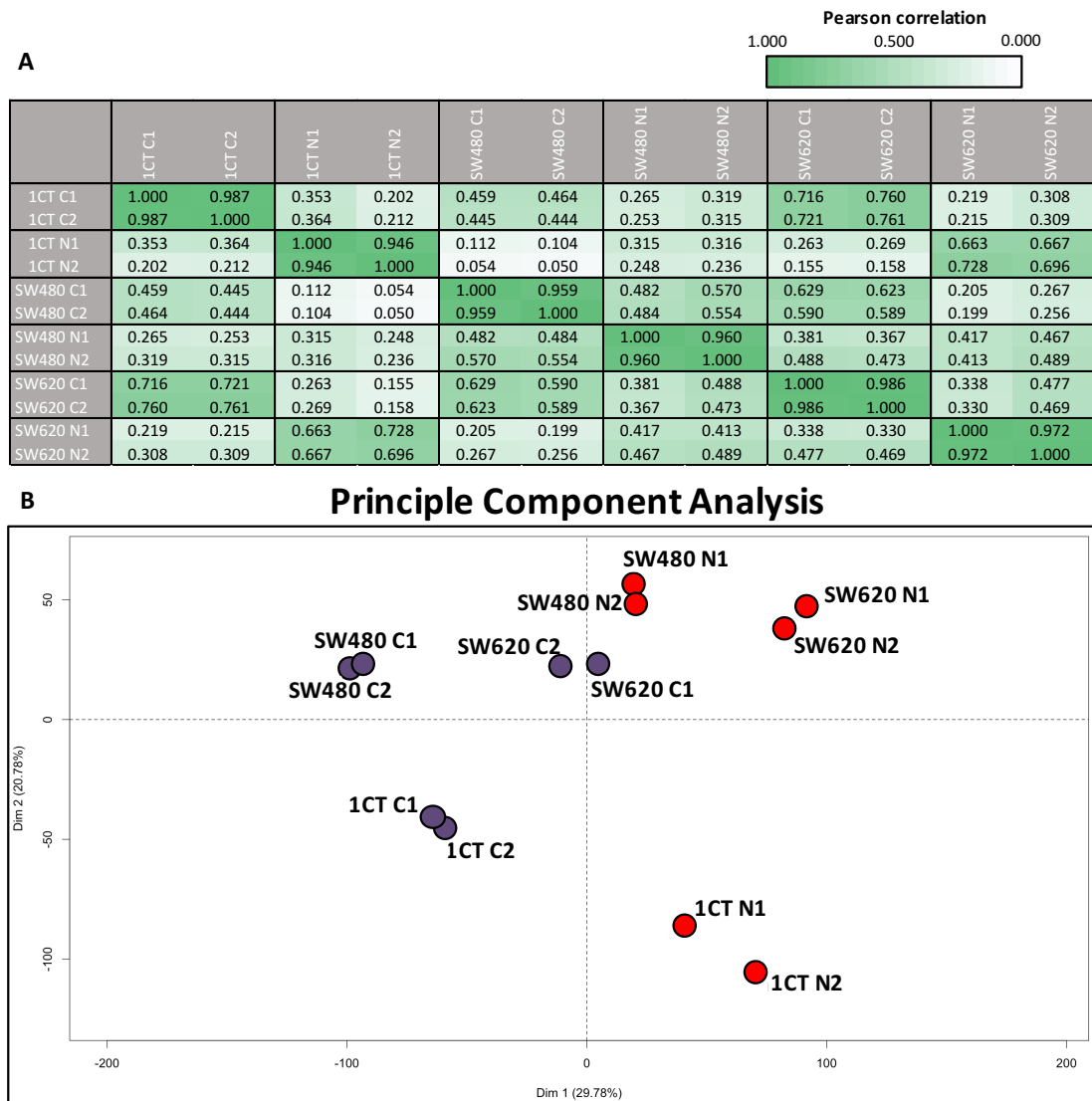


Figure 46: Reproducibility of 3'-end sequencing data from 1CT, SW480 and SW620 subcellular fractions

Panel A – A correlation matrix showing the Pearson correlation between the gene expression values of 1CT, SW480 and SW620 subcellular fractions, as calculated by 3'-end sequencing using QuantSeq 3' mRNA-Seq on the Ion Proton Sequencer.

Panel B – Principal component analysis based on gene expression values calculated for each subcellular fraction. This shows the clustering of the biological repeats of each sample, as well as the distinct expression profiles for each of the subcellular fractions of the cell lines analysed. The principle components along axis X and Y show the proportion of variance encompassed by that principal component.

Genes that are found to be significantly downregulated in both SW480s and SW620s relative to 1CTs, such as *MYL9*, have previously been found to be downregulated in colorectal carcinomas compared to normal mucosa (Cardoso et al., 2007). Furthermore, some of the key genes previously found to be upregulated in colorectal cancer are also included in these 50 genes. The top 5 most significantly upregulated in SW620s relative to 1CTs are shown in Table 12. Four out of five of these overexpressed genes have previously been identified as being overexpressed in colorectal cancer samples relative to their non-cancerous equivalents, thus further demonstrating that using the use of 1CTs and SW620s are a suitable model for colorectal cancer.

Table 12 - Most significantly overexpressed genes in SW620 relative to 1CT cells

Gene Symbol	Description	References
<i>KRT19</i>	Member of the keratin family which are responsible for the structural integrity of epithelial cells. <i>KRT19</i> has previously been shown to be overexpressed in 90% of colorectal cancer patients when analysing circulating tumour cells.	(C.-H. et al., 2013)
<i>S100A4</i>	Member of S100 calcium binding family and secreted by tumour cells to promote angiogenesis. Found to be overexpressed in colorectal adenocarcinomas.	(Takenaga et al., 1997)
<i>NQO1</i>	NAD(P)H quinine oxidoreductase 1, which is known to be overexpressed in colorectal cancer tissue samples relative to adjacent non-tumour tissues.	(Ji et al., 2014)
<i>RAN</i>	Ras-related nuclear protein and a member of the Ras oncogene family. Involved in nuclear transport, microtubule nucleation and spindle formation. <i>RAN</i> expression levels have previously been found to have significantly higher expression levels in colorectal cancer cells when compared to normal epithelial cancer cells, and increased levels of expression correlate with poor prognosis.	(Fan et al., 2013)
<i>PFN1</i>	Profilin-1 is a regulator of the microfilament system which has not previously been linked to colorectal cancer. However, increased expression levels have previously been found in gastric cancer, and knockdown of <i>PFN1</i> inhibits cell migration and proliferation through G ₀ /G ₁ arrest.	(Cheng et al., 2015)

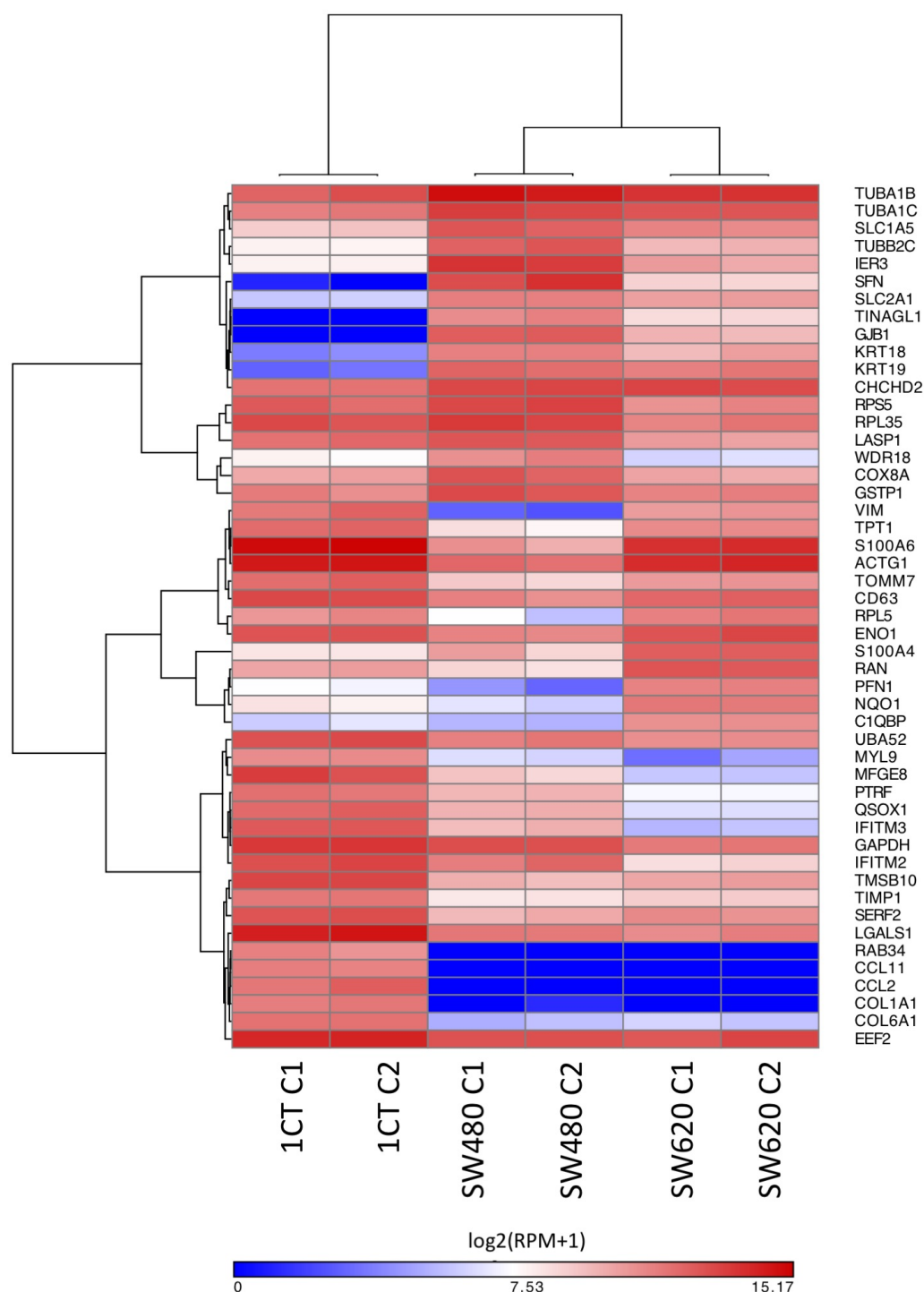


Figure 47: Heat map of genes which show variable expression between 1CT, SW480 and SW620 cell lines

A heat map of the top 50 most differentially expressed genes between the cytoplasmic fractions of 1CT, SW480 and SW620 cells. Genes were selected as the most differentially expressed by those that have the lowest average p-value (Fisher's exact test) of differential expression when comparing the gene expression values between the cytoplasmic samples from the different cell lines. The heat map was generated using the GENE-E software (<http://www.broadinstitute.org/cancer/software/GENE-E>) using the one minus Pearson correlation distance metric for clustering.

6.2.3. *Comparison of Differential Subcellular APA isoform distributions*

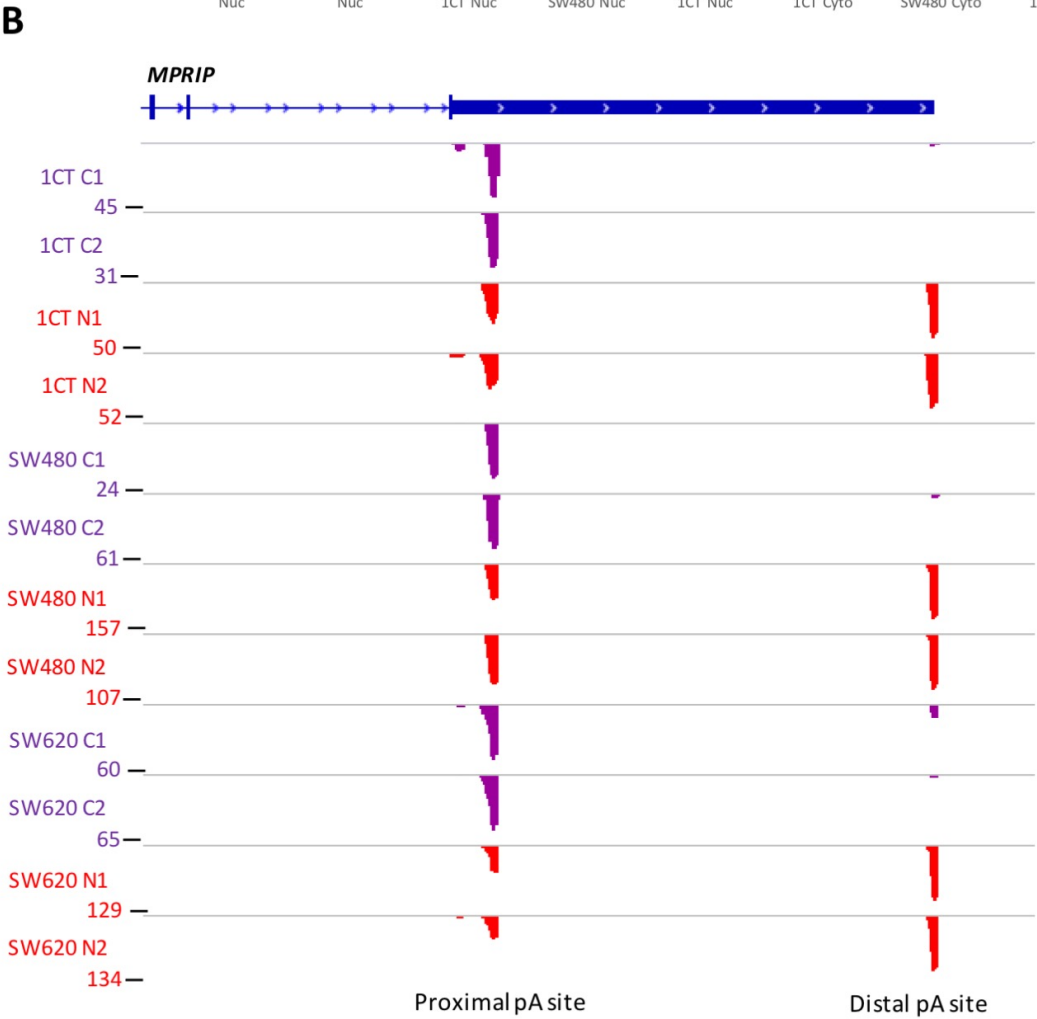
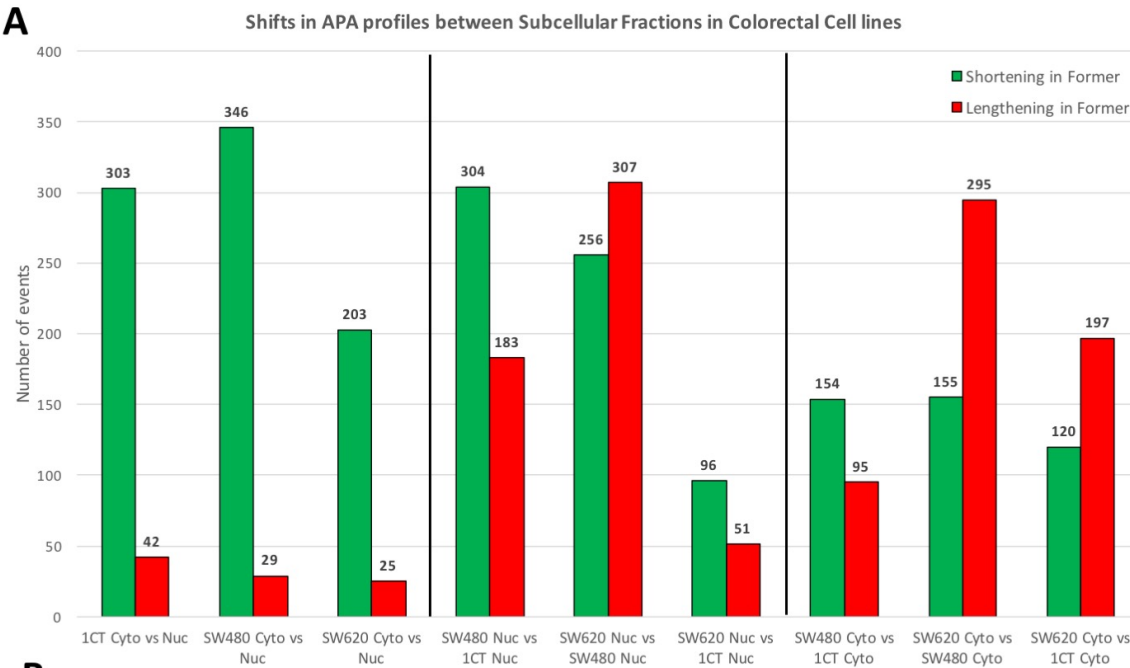
To compare the APA profiles of the subcellular fractions for each cell line, the biological replicates of each sample were combined and analysed for significant differential representation of APA isoforms. As in Section 4.2.2, significant APA events are classified as those where $p < 0.01$ (Fisher's exact test) with a minimum of 5% change in APA isoform abundance. Initially, comparisons between the cytoplasmic and nuclear APA profiles were compared for each of the 1CT, SW480 and SW620 cell lines (Figure 48A, first three comparisons). For UTR-APA, this showed very similar profiles to that seen in the HEK293 and HBL cells in Chapter 3 (Figure 18), with the cytoplasmic fractions generally showing an overrepresentation of the proximal APA isoforms in the cytoplasm. *MPRIP* is shown in Figure 48B, of an example of where the proximal APA isoform is overrepresented in the cytoplasm in all three cell lines.

(Figure on the following page)

Figure 48: APA analysis of the subcellular fractions of colorectal cancer cell lines

Panel A – Bar chart displaying the number of events showing differential APA isoform representation when comparing UTR-APA isoform abundances. The identity of the samples being compared is shown under each comparison with each bar representing events where the shorter APA isoforms (green) or longer APA isoforms (red) have a significantly higher relative representation in the former sample. Significant differentially represented APA isoforms are those in which $P \leq 0.01$ (Fisher's exact test) and the isoform abundance change is $> 5\%$ relative to total APA isoforms associated with that gene.

Panel B - Screenshot of *MPRIP* as an example of a gene that shows an overrepresentation of the proximal APA isoform in the cytoplasmic fraction, relative to the nuclear fraction, in all three colorectal cell lines analysed here.



6.2.4. *Comparison of Nuclear APA profiles in 1CT, SW480 and SW620 cells*

To assess the different APA profiles in 1CTs, SW480s and SW620s at a point more reflective of pA site choice, the nuclear profile was used. Despite still being a steady state profile, the nuclear RNA would be expected to reflect the pA site choice more closely than that of either the cytoplasmic RNA or RNA isolated from the whole-cell (Li et al., 2015; Neve et al., 2016). This is because looking at the nuclear RNA profile mitigates the effects of different cytoplasmic stabilities of the APA isoforms between the two cell types. To this end, the nuclear profiles of 1CTs, SW480s and SW620s were compared.

When comparing the nuclear APA profiles of either SW480s or SW620s to 1CTs, there was a significant bias ($p=4.6 \times 10^{-8}$ and 2.6×10^{-4} respectively, 2-tailed binomial test) towards shifts to proximal APA isoforms (Figure 48A, middle three comparisons). Comparing the nuclear APA profiles in SW620s compared to SW480s, there is a small bias towards shifts to distal pA sites in SW620s ($p=0.035$). One example of a gene that shows a different nuclear APA profile in 1CTs when compared to SW480s and SW620s is *TOLLIP* (Figure 49). In the 1CT nuclear APA profile, the distal APA isoform is the most abundant, whereas the proximal APA isoform is the dominant isoform in the nuclear samples of both SW480s and SW620s. Conversely, the APA profiles of the cytoplasmic fractions are all relatively similar, thus demonstrating the importance of considering nuclear APA profiles as well as the steady-state whole-cell or cytoplasmic APA profiles.

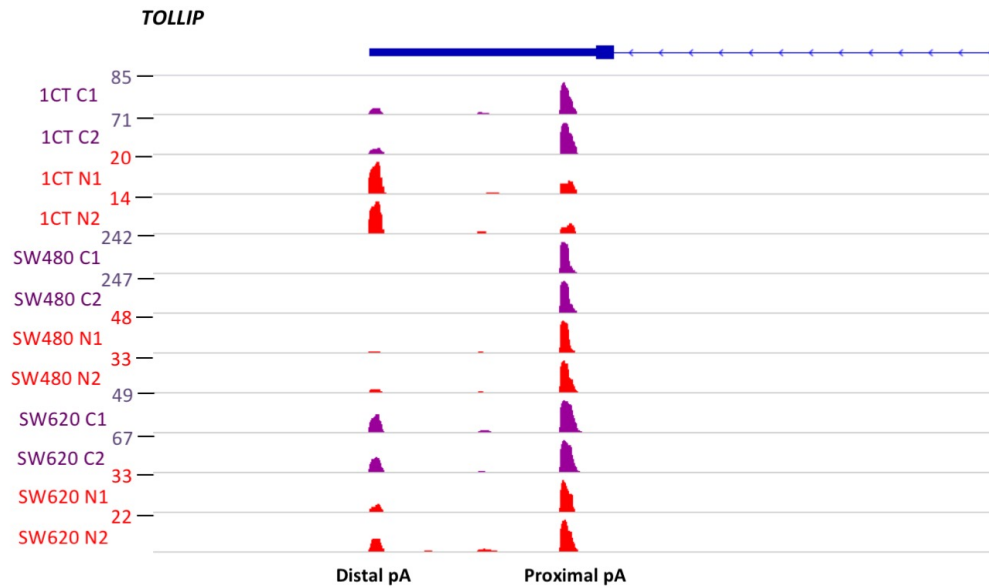


Figure 49: Screenshot of 3' end sequencing reads mapping to the *TOLLIP* locus

Screenshot of *TOLLIP* as an example of a gene that shows different nuclear APA profiles when comparing 1CTs, SW480s and SW620s. The positioning of the proximal and distal pA sites are indicated below the sequencing tracks and the read levels are shown to the left of each track.

6.2.5. Comparison of Cytoplasmic APA profiles in SW620 and 1CT cells

For completion, comparisons of the cytoplasmic RNA profiles were made to identify those genes that differ in their ratio of APA isoforms in their cytoplasmic fractions (Figure 48A, last three comparisons). When considering the SW620 cytoplasmic APA profiles, there does appear to be a slight bias towards shifts to distal APA isoforms compared to 1CTs ($p=1.8 \times 10^{-5}$) and SW480s ($p=4.0 \times 10^{-11}$). However, since these APA profiles are steady-state cytoplasmic levels, it is unclear whether these shifts are occurring as a result of changes in the nuclear APA profile penetrating through to the cytoplasm, or are the result of post-transcriptional changes targeting specific APA isoforms in the cytoplasm.

6.2.6. *Overlap of genes that show differential regulation and pA site shifts*

If shifts in pA site choice were indeed contributing to either carcinogenesis and disease phenotype, it would be expected that the shifts in pA site choice would be occurring in those genes that are also differentially regulated. This is where the subcellular fractionation approach becomes valuable, as this would correspond to the genes where you see a shift in both nuclear APA profiles from the progression of 1CTs through to SW620s, as well as a shift between the nuclear and cytoplasmic APA profiles in the SW480s or SW620s themselves. If a shift in nuclear APA profiles were seen, but not a shift in cytoplasmic to nuclear APA profiles, this would suggest that, even though there is a shift in pA site choice, it is not significantly affecting either the cytoplasmic stability or nuclear export rates. Thus, in this scenario, the shifts between pA sites becomes less likely to influence any changes at the protein level, and consequently unlikely to have a physiological effect.

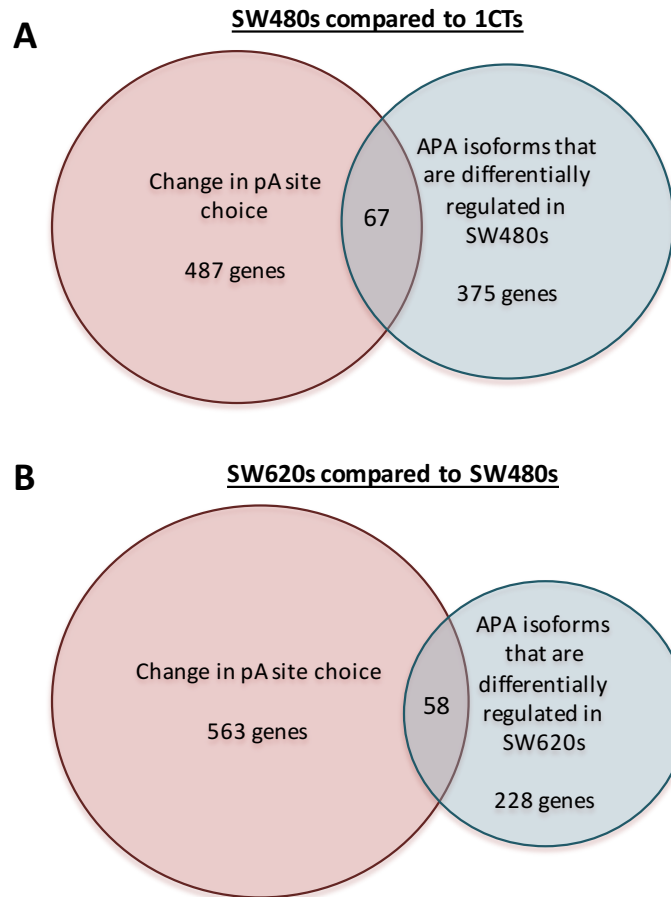


Figure 50: Genes that show both differential regulation of APA isoforms and a shift in nuclear APA profiles through colorectal cancer progression

Panel A – Venn diagram showing the overlap of genes that show both a shift in APA profiles when comparing the nuclear APA profiles of SW480s and 1CTs (classed as “Change in pA site choice”) and a shift between the cytoplasmic and nuclear APA profiles of SW480s (classed as “APA isoforms that are differentially regulated in SW480s”).

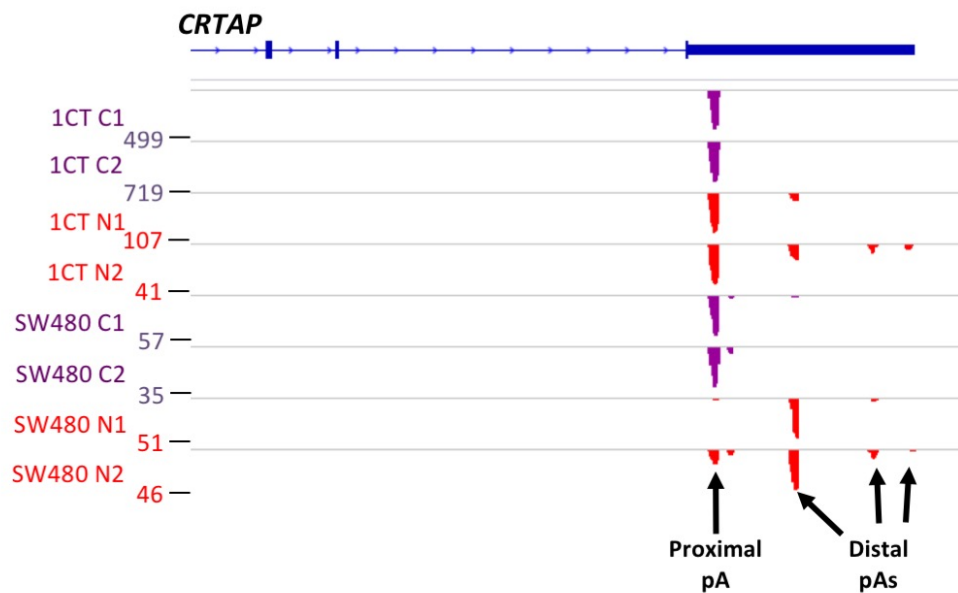
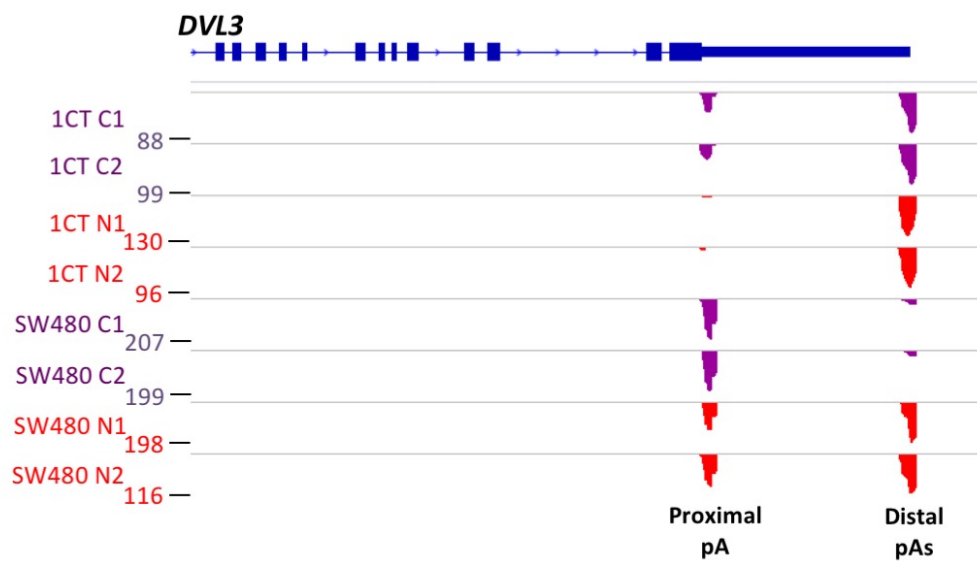
Panel B – Same as in Panel A but comparing SW620s to SW480s.

To this end, those genes that display both a shift in nuclear APA profile through colorectal cancer progression and also show differential subcellular APA isoform distribution in the cancerous cell line are highlighted in Figure 50. This corresponds to ~18% (67) of the genes that show a shift in pA site choice from 1CTs to SW480s, and also show differential regulation of APA isoforms in SW480s. An example of this is seen in *CRTAP* where there is a shift in nuclear APA profile towards distal pA sites in SW480s compared to 1CTs, and these distal APA isoforms are then not seen in the cytoplasmic fraction of SW480s (Figure 51B). This differential representation of the APA isoforms in SW480s indicates that either

the distal transcripts are not being efficiently exported from the nucleus or that they are rapidly degraded in the cytoplasm.

Interestingly, when looking at the read levels in the different samples, it can be seen that, even though the nuclear levels of *CRTAP* are similar in SW480s and 1CTs, the cytoplasmic level in SW480s is ~10 fold less than that in 1CTs. This change in steady-state cytoplasmic RNA levels may well be partially due to the shift to the distal pA site as this would be the exact profile expected if the distal isoform was less stable. To test whether this pattern holds for the other genes in this cohort, genes that lengthen or shorten in the SW480 nuclear fraction relative to the 1CT nuclear fraction, and also show an overrepresentation of the proximal APA isoform in the SW480 cytoplasmic fraction compared to the nuclear, were isolated and the overall cytoplasmic RNA levels were compared between 1CT and SW480 (Figure 51). The comparative RNA levels were either classified as upregulated, downregulated or no change in SW480s ($p < 0.01$ Fisher's exact test).

It would be expected if transcription levels remain unchanged between cell lines, but there was a shift to a more distal pA site, resulting in a less stable APA isoform, this would result in a drop in the cytoplasmic RNA levels in SW480 compared to that of 1CTs. Of the genes that do show differential expression levels, there is indeed a bias towards a drop in cytoplasmic level (15/20, $p < 0.05$, two-tailed binomial test) (Figure 51A). As discussed above, *CRTAP* is an example of one of these cases (Figure 51B). In those examples where there is a shift to proximal pA sites in SW480s compared to 1CTs, stabilisation in the cytoplasm of SW480 would be expected, and thus upregulation of RNA levels in SW480s. Of the genes that do change in expression levels there is a slight, but non-significant, bias towards an increase in cytoplasmic RNA levels in SW480s, as exemplified by *DVL3* (Figure 51C).

A**B****C**

(Figure on the previous page)

Figure 51: Changes in gene expression levels upon shifts in nuclear APA profiles between SW480s and 1CTs

Panel A - Bar chart showing the changes in gene expression levels in the subset of genes that were isolated as shifting nuclear APA profiles when comparing SW480s to 1CTs, and also a differential representation of those APA isoforms between the cytoplasmic and nuclear fractions of SW480s. These genes were separated into those that moved to more proximal pA sites in the nuclear fraction of SW480s, relative to 1CTs (the three bars on the left), and those that moved to more distal sites (those three bars on the right). Changes in gene expression level was defined as a significant change in the steady-state cytoplasmic RNA levels, as measured by 3' end sequencing ($p < 0.01$, Fisher's exact test). * Indicates a significant bias in direction to changes in gene expression level $p < 0.05$ (2-tailed binomial test)

Panel B - Screenshot of *CRTAP* as an example of a gene that shows shifts to the distal APA isoform when comparing the nuclear profiles of SW480s to that of 1CTs, as well as differential representation of APA isoforms in the cytoplasmic and nuclear fractions of SW480s. This is accompanied by a significant decrease in cytoplasmic steady-state RNA levels in SW480 as shown by the read numbers to the left of each track.

Panel C - Screenshot of *DVL3* as an example of a gene that shows shifts to the proximal APA isoform when comparing the nuclear profiles of SW480s to that of 1CTs, as well as differential representation of APA isoforms in the cytoplasmic and nuclear fractions of SW480s. This is accompanied by a significant increase in cytoplasmic steady-state RNA levels in SW480 as shown by the read numbers to the left of each track.

This same analysis was then also done when comparing SW620s to SW480s to observe the changes that occur upon the onset of metastasis. This showed similar numbers of genes that both undergo a shift in pA site choice and that the resulting APA isoforms are also differentially regulated in SW620 cells. In the genes that shift to shorter APA isoforms in SW620s, and also show a differential representation of APA isoforms in the SW620 cytoplasmic fraction compared to the nuclear fraction, there is a significant bias (12/14 genes that show a difference in expression levels) towards an increase in cytoplasmic RNA levels in SW620 (Figure 52A). An example of this is shown by *CCT6A* in Figure 52C. However, when looking at examples of the genes that shift to the distal pA site in SW620s, the analysis does not show the expected pattern of a general reduction in RNA levels in SW620s, with a slight, but non-significant, bias towards an increase in SW620 cytoplasmic RNA levels (Figure

52A). Despite this, some genes do still show a decrease in RNA levels upon shifting to distal APA isoforms in SW620s, as exemplified by *SMAD3* in Figure 52B.

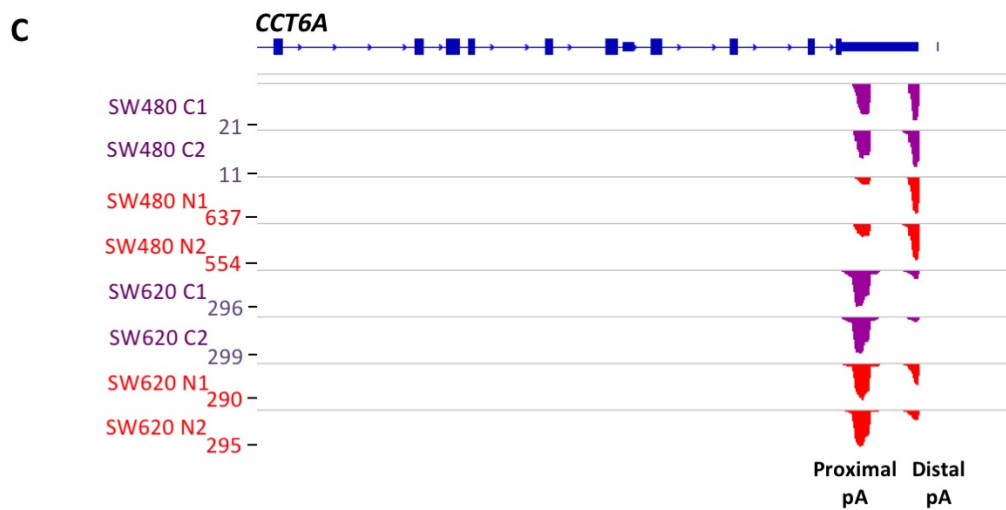
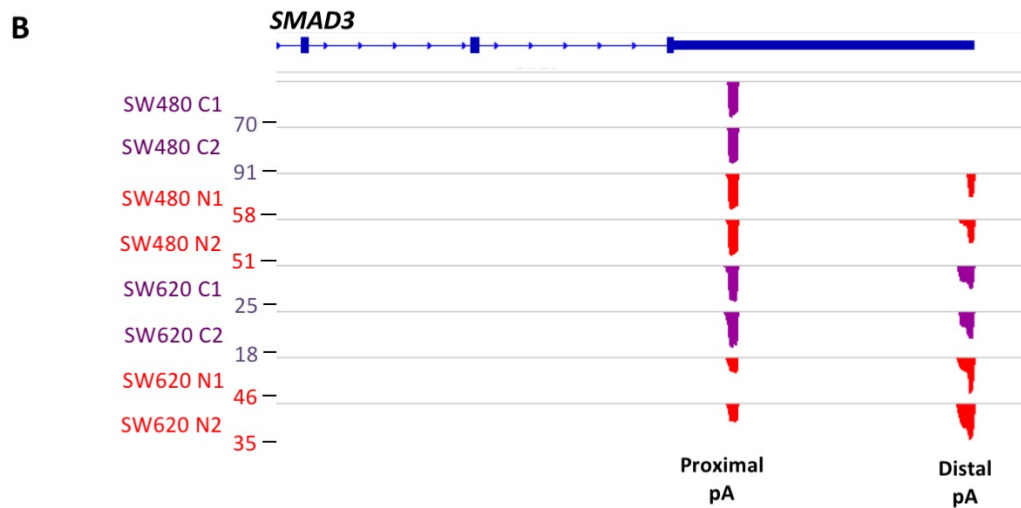
(Figure on the following page)

Figure 52: Changes in gene expression levels upon shifts in nuclear APA profiles between SW620s and SW480s

Panel A - Bar chart showing the changes in gene expression levels in the subset of genes that were isolated as shifting nuclear APA profiles when comparing SW620s to SW480s, and also a differential representation of those APA isoforms between the cytoplasmic and nuclear fractions of SW620s. These genes were separated into those that moved to more proximal pA sites in the nuclear fraction of SW620s, relative to SW480s (the three bars on the left), and those that moved to more distal sites (those three bars on the right). Changes in gene expression levels were defined as a significant change in the steady-state cytoplasmic RNA levels, as measured by 3' end sequencing ($p < 0.01$, Fisher's exact test). * Indicates a significant bias in direction to changes in gene expression level $p < 0.05$ (2-tailed binomial test)

Panel B - Screenshot of *SMAD3* as an example of a gene that shows shifts to the distal APA isoform when comparing the nuclear profiles of SW620s to that of SW480s, as well as differential representation of APA isoforms in the cytoplasmic and nuclear fractions of SW620s. This is accompanied by a significant decrease in cytoplasmic steady-state RNA levels in SW620s as shown by the read numbers to the left of each track.

Panel C - Screenshot of *CCT6A* as an example of a gene that shows shifts to the proximal APA isoform when comparing the nuclear profiles of SW620s to that of SW480s, as well as differential representation of APA isoforms in the cytoplasmic and nuclear fractions of SW620s. This is accompanied by a significant increase in cytoplasmic steady-state RNA levels in SW620s as shown by the read numbers to the left of each track.



6.2.7. *Gene ontology analysis of APA genes in colorectal cell lines*

To identify if the genes which have been isolated in Section 6.2.3 as differentially regulated, belong to any specific group of genes, PANTHER Gene Ontology analysis was performed on these subsets of genes. The genes that showed an overrepresentation of the proximal APA isoform in the cytoplasm were isolated for each cell line and run through the PANTHER analysis separately. The “biological processes” that showed significant enrichment ($p < 0.05$) are shown in Figure 53. The most highly enriched category was found in genes which showed an overrepresentation of the proximal APA isoform in the cytoplasm of SW620s, where there was a highly significant enrichment of genes involved in mRNA splicing and mRNA processing. This Gene Ontology analysis was also repeated for genes that showed an overrepresentation of the distal isoform in the cytoplasm, but these analyses did not identify significant enrichment of any “biological process” categories.

Subsequently, the genes that were found to shift their APA profiles when comparing the nuclear profiles of the different cell lines, and also identified as being differentially regulated (highlighted in Figure 50), were run through the PANTHER Gene Ontology Analysis. Once again ontology categories involving mRNA processing were among the most significantly enriched biological process categories. In addition, genes that shifted their nuclear APA profile between SW480s and SW620s, and were also isolated as showing differential APA isoform representation in the cytoplasmic and nuclear fractions of SW620s, were found to be enriched for genes involved in macromolecule localisation and cellular metabolic processes. Despite these enriched biological process categories, these still only make up a small subset of the genes that fall into each of these categories.

1CT Cytoplasmic vs Nuclear				
Shortening	Number found	Expected	Enrichment	P-value
aging	16	3.8	4.21	1.68E-02
protein localization to membrane	19	5.12	3.71	1.16E-02
apoptotic process	34	14.28	2.38	2.55E-02
negative regulation of macromolecule metabolic process	65	32.9	1.98	4.56E-04
negative regulation of cellular metabolic process	62	32.32	1.92	2.85E-03
response to stress	87	53.21	1.64	7.08E-03
single-organism metabolic process	105	70.47	1.49	2.59E-02
regulation of primary metabolic process	116	79.92	1.45	1.97E-02
cellular macromolecule metabolic process	137	97.3	1.41	5.22E-03
primary metabolic process	162	123.81	1.31	2.08E-02
SW480 Cytoplasmic vs Nuclear				
Shortening				
Positive regulation of cellular protein localization	21	5.13	4.09	6.92E-04
single-organism membrane organization	30	12	2.5	4.28E-02
single-organism intracellular transport	44	19.93	2.21	6.67E-03
SW620 Cytoplasmic vs Nuclear				
Shortening				
mRNA splicing, via spliceosome	19	2.25	8.43	1.93E-08
mRNA processing	20	3.77	5.31	1.54E-05
cellular response to decreased oxygen levels	9	1.2	7.49	3.53E-02
RNA catabolic process	12	2.09	5.75	1.35E-02
posttranscriptional regulation of gene expression	16	4.07	3.93	3.40E-02
intracellular protein transport	21	6.37	3.29	1.67E-02
cytoplasmic transport	24	7.93	3.03	1.29E-02
single-organism intracellular transport	31	11.91	2.6	8.24E-03
organelle organization	60	29.7	2.02	2.39E-04
negative regulation of biological process	69	41.23	1.67	2.22E-02
Shifts between SW480 Nuclear vs 1CT Nuclear and shows differential regulation in SW480s				
mRNA metabolic process	12	1.71	7.04	8.76E-04
RNA processing	12	2.32	5.16	2.26E-02
cytoplasmic transport	12	2.5	4.8	4.75E-02
Shifts between SW620 Nuclear vs 1CT Nuclear and shows differential regulation in SW620s				
mRNA splicing, via spliceosome	6	0.29	20.41	3.12E-03
Shifts between SW620 Nuclear vs SW480 Nuclear and shows differential regulation in SW620s				
macromolecule localization	18	5.66	3.18	4.61E-02
cellular metabolic process	42	24.44	1.72	1.21E-02

Figure 53: Gene Ontology analysis of Differentially Regulated Genes in Colorectal Cell lines

All gene ontology “biological process” categories with significant enrichment ($p < 0.05$) of the genes that showed an overrepresentation of proximal APA isoforms in the cytoplasm are shown for each colorectal cell line analysed. Those groups that showed an overrepresentation of the distal APA isoform in the cytoplasm were also analysed but showed no significant enrichment for any categories. Those genes, highlighted in Figure 51, which showed shifts in nuclear APA profiles throughout colorectal cancer progression, and displayed a differential representation of the APA isoforms when comparing the cytoplasmic and nuclear APA profiles, were also analysed for enrichment for specific “biological process” categories and are shown in the lower panels. Gene ontology analysis was performed using PANTHER Gene Ontology tools with the Statistical Overrepresentation Test using default settings.

6.2.8. *Genes that are differentially regulated in SW480s and SW620s but not 1CTs*

Another potentially interesting aspect to consider is using this model of colorectal cancer progression to isolate those genes that have differentially regulated APA isoforms in one cell line but not the other. For example, even if there was no shift in pA site usage, but there was a shift between the cytoplasmic and nuclear APA profiles of one cell line and not the other, this would indicate that there was differential regulation of the APA isoforms in the former cell line, but not the latter. This in itself is interesting as, even though it does not involve a regulated shift between pA sites, it provides a mechanism for identifying post-transcriptional regulatory pathways that are active in one cell line but not in others. For example, this could be through the expression of specific miRNAs or activation of regulatory RNA-binding proteins that are present in one cell line but not in the other.

Of the 375 APA events that show differential regulation in 1CTs, only 102 of these are showing the same differential regulation in SW480s. This means that the remaining 273 genes are being differentially regulated specifically in SW480s. One good example of this is *RAC1* where there is a common preference for the distal pA site in all of the 1CT, SW480 and SW620 nuclear fractions. However, there is no obvious shift in the cytoplasmic fraction of 1CTs, but a clear shift to the proximal pA site in the cytoplasmic fraction of in both SW480s and SW620s (Figure 54A). This indicates that the distal APA isoform is being either nuclear-retained or destabilised in the cytoplasm of only SW480s and SW620s, thereby indicating that there may be a post-transcriptional regulatory pathway which is present in SW480s and SW620s, but not in 1CTs.

Conversely, *CCND1* is an example where differential regulation is reduced through colorectal cancer progression. Specifically, in SW620s compared to both SW480s and 1CTs, where the distal APA isoform no longer appears to be being differentially regulated as the cytoplasmic and nuclear APA profiles of SW620s are not significantly different (Figure 54B). However, when considering the steady-state cytoplasmic levels of *CCND1*, there is not

the increase in RNA that would be expected with this stabilisation. This is likely because it is acting in conjunction with many other factors, such as transcriptional rates, which also contribute to this steady-state level. However, if the nuclear *CCND1* levels are also taken into account to crudely estimate transcription rates, it can be seen that SW620s display the highest ratio of cytoplasmic to nuclear *CCND1* RNA levels (0.47) compared to SW480s (0.12) and 1CTs (0.33). This indicates that, while differential regulation of *CCND1* APA isoforms may not be the sole determinant of steady-state cytoplasmic *CCND1* RNA levels, it is indeed a contributing factor.

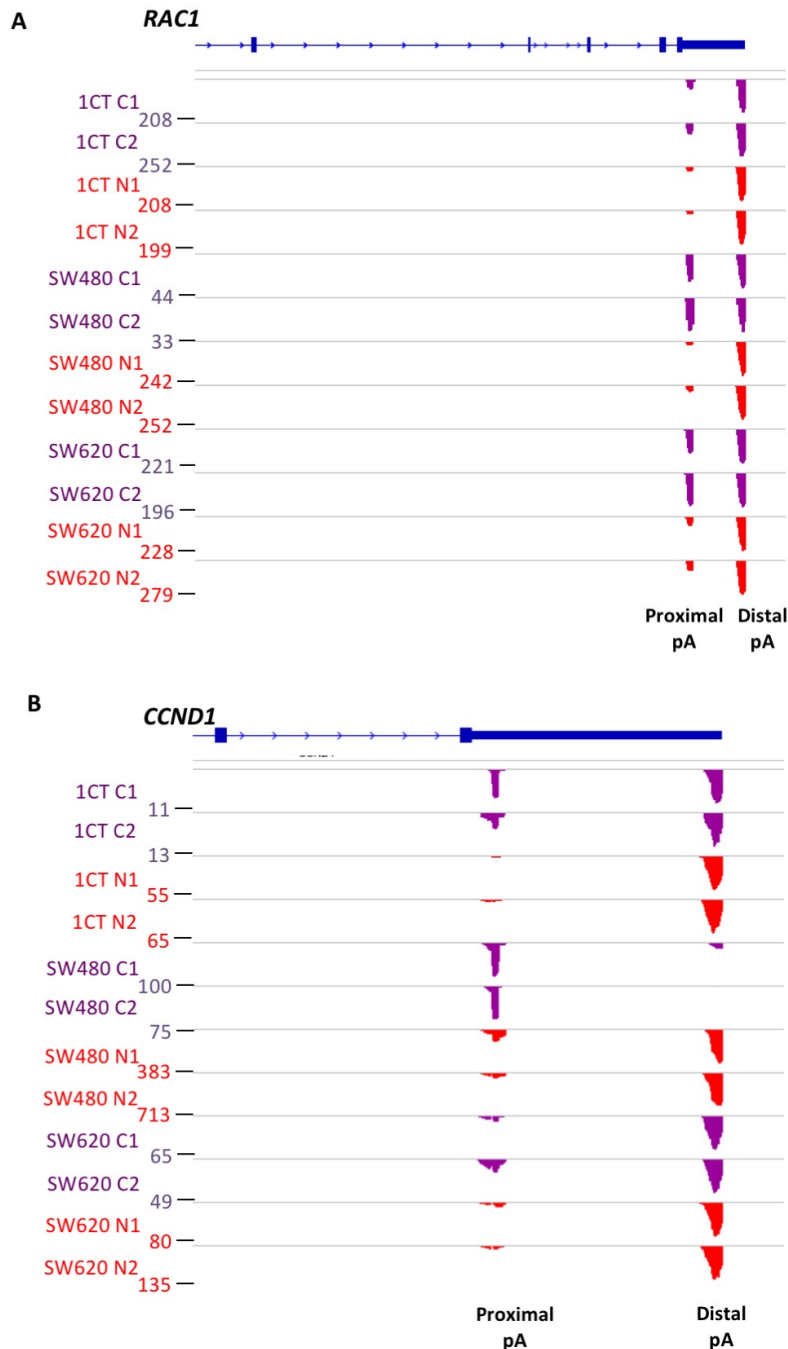


Figure 54: Examples of genes which display varying levels of differential regulation

Panel A - Screenshot of *RAC1* as an example of a gene that shows shifts to the proximal APA isoform when comparing the cytoplasmic and nuclear APA profiles of SW480s and SW620s, but not in 1CTs. This indicates differential regulation of *RAC1* APA isoforms through a post-transcriptional mechanism that is active in SW480s and SW620s, but not 1CTs.

Panel B - Screenshot of *CCND1* as an example of a gene that shows shifts to the proximal APA isoform when comparing the cytoplasmic and nuclear APA profiles of 1CTs and SW480s, but not in SW620s. This indicates differential regulation of *CCND1* APA isoforms through a post-transcriptional mechanism that is active in 1CTs and SW480s, but not SW620s.

6.2.9. *Motif analysis of aUTRs in genes that show differential regulation of APA isoforms*

In an attempt to identify the different post-transcriptional pathways that are active in these different colorectal cell lines, motif analysis and miRNA target site enrichments was performed as described in Sections 4.2.8 and 4.2.9 respectively. Firstly, the genes that showed an overrepresentation of the proximal APA isoform in the cytoplasm compared to the nucleus were isolated for each cell line. The aUTR locations and sequences of these genes were then extracted and used for miRNA and RNA-binding protein target site analysis.

For miRNA-target site enrichment analysis, the z-scores were calculated, and the top 4 most enriched miRNAs are shown in Figure 55A. This analysis identified a large number of miRNAs whose target sites were significantly enriched in the aUTRs of genes that produce APA isoforms that are differentially regulated in these three cell lines. Interestingly, target sites for miR-142 were found to be in the top 4 miRNAs with enriched target sites in the aUTRs of differentially regulated APA isoforms in all three cell lines. The other miRNAs appeared to be more specific to the individual cell lines, which would be expected, due to changes in miRNA expression profiles through colorectal cancer progression (Piepoli et al., 2012).

Subsequently, these same aUTR regions were run through DREME motif analysis using aUTRs from genes that showed no differential representation in any of the three cell lines, as the control set of sequences. DREME analysis was performed, searching only the given strand, and the identified motifs were compared to previously known motifs for RNA-binding proteins (Ray et al., 2013) using the TOMTOM motif comparison tool (Gupta et al., 2007). The top 3 most enriched motifs discovered for each cell line are shown in Figure 55B alongside the RNA-binding protein with the most similar motif ($p < 0.05$) shown next to each discovered motif. This analysis did isolate a distinct set of motifs, with corresponding RNA-

binding proteins, for each cell line thereby identifying different post-transcriptional regulatory pathways which may be active to varying degrees in these different cell lines.

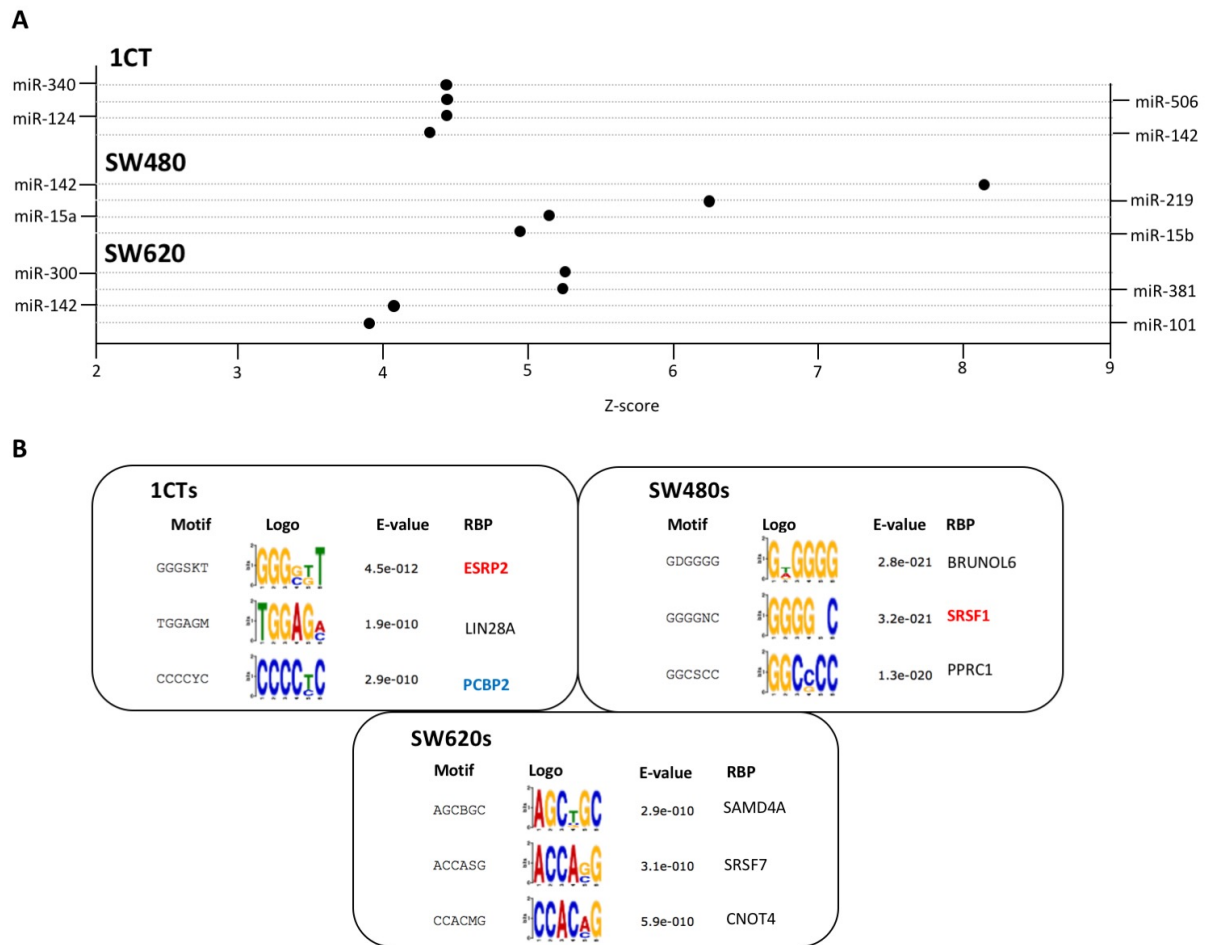


Figure 55: Motif analysis and miRNA-target site enrichment analysis in the aUTRs of differentially regulated genes in colorectal cell line

Panel A - Enrichment of miRNA target sites in the alternative UTRs of differentially regulated APA isoforms in colorectal cell lines. The x-axis represents the z-score of enrichment of miRNA target sites in the aUTRs of the genes that show an overrepresentation of the proximal APA isoform in the cytoplasm of each colorectal cell line analysed. This was calculated by comparing the observed frequency of target sites with what would be expected from random permutations of target sites across the set of alternative 3'UTRs. A positive z-score indicates that the frequency of miRNA target sites is greater than if sites were placed randomly in alternative regions of 3'UTRs. Only the top 4 miRNAs that have enriched target sites in these aUTRs for each cell line are shown.

Panel B - The aUTRs of all genes that showed an overrepresentation of the shorter APA isoform in the cytoplasm in each colorectal cell line were extracted and subjected to motif analysis using the DREME tool. The top 3 motifs for each cell line are shown as well as corresponding RNA-binding proteins that potentially bind these motifs (Ray et al. 2013). RNA-binding proteins are highlighted in red or blue if they have previously been classified as exhibiting a destabilising or stabilising effect to their target RNAs, respectively (Ray et al. 2013).

6.2.10. Coding region APA in colorectal cancer progression

Interpreting shifts in CR-APA in colorectal cancer progression are somewhat less complicated than UTR-APA because a shift in pA site choice in CR-APA will result in a different protein being produced. Therefore, the emphasis on whether the resulting APA isoforms are differentially regulated is reduced. To isolate events that shift between CR-APA isoforms between the different colorectal cell lines the equivalent subcellular fractions were compared throughout 1CTs, SW480s and SW620s. As in Section 4.2.2, this was done by combining all reads upstream of the terminal exon and comparing these to the combined reads within the 3'UTR. The number of events that differed significantly ($p < 0.01$, Fisher's Exact Test) between the equivalent subcellular fractions of the colorectal cell are shown in Figure 56A.

When comparing both the cytoplasmic and nuclear fractions of SW480s to the equivalent fractions in 1CTs, there is a bias towards shifting to pA sites that are located in the coding region of genes. However, from the progression from SW480s to SW620s, there is a bias in shifts towards APA isoforms that have pA sites located back in the 3'UTRs. Despite this, in both these cases, there are a noteworthy number of events moving in each direction. *RBCK1* is shown in Figure 56B and an individual example of this shift between CR-APA isoforms when comparing the three colorectal cell lines. In normal colorectal epithelial cells (1CTs), the full-length APA isoform is expressed at around the same levels as the shorter APA isoform that would correspond to the truncated protein. However, in the SW480 cell line this full-length APA isoform, is almost completely lost in both the cytoplasmic and nuclear fractions. Furthermore, in the SW620 cell line the full-length APA isoform is seen in the nuclear fraction to a larger extent than in SW480s, but as in SW480s, this full-length distal APA isoform is almost completely absent in the cytoplasmic fraction, relative to the shorter APA isoform.

Together, through this switch between APA isoforms, this would be expected to result in a much higher expression of the shorter protein isoform relative to the full-length protein, in both the SW480s and SW620s when compared to the 1CTs. From the Ensembl database, this shorter protein isoform would correspond to the *RBCK1-001* transcript and encode a 105 amino acid protein, and the transcripts utilising the pA site located in the 3'UTR would correspond to either *RBCK-009* or *RBCK-010*, which encode 468 and 510 amino acid proteins respectively.

To identify whether shifts in CR-APA profiles were occurring in genes belonging to specific biological processes, PANTHER Gene Ontology analysis was performed on the genes that were isolated in each of the comparisons shown in Figure 56A. Genes that showed shortening or lengthening were analysed separately. The only set of genes that showed significant enrichment for any biological processes category were those genes which showed a preference for the shorter CR-APA isoforms in the nuclear fraction of SW480s when compared to the equivalent fraction in 1CTs (Figure 56C). The biological processes found here included a number of categories, the most significant of which is interestingly, once again, RNA processing.

(Figure on the following page)

Figure 56: CR-APA analysis of the subcellular fractions of colorectal cancer cell lines

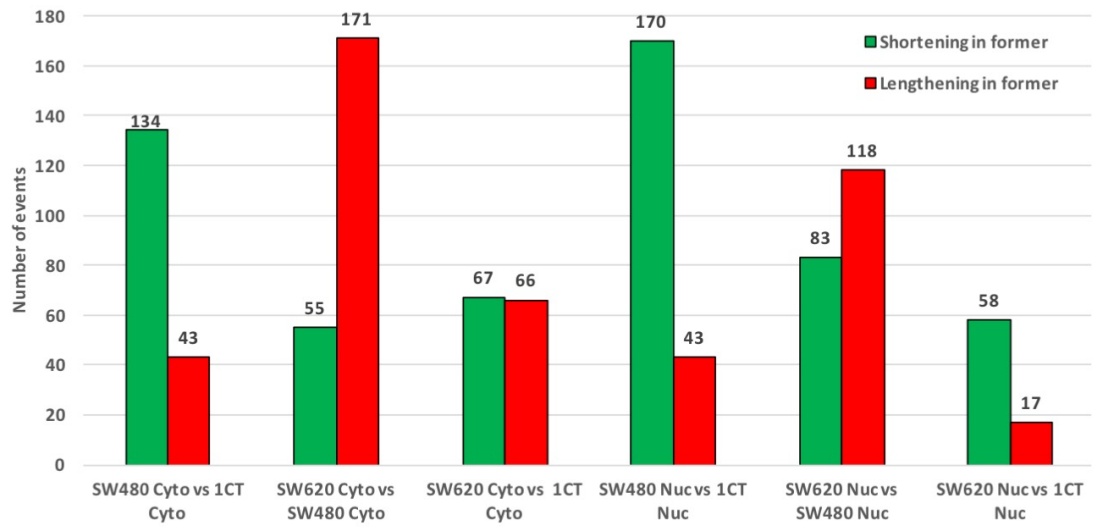
Panel A – Bar chart showing the number of events showing differential APA isoform representation when comparing CR-APA isoform abundances. The identity of the samples being compared is shown under each comparison with each bar representing events where the shorter APA isoforms (green) or longer APA isoforms (red) have a significantly higher relative representation in the former sample. Significant differentially represented APA isoforms are those in which $P \leq 0.01$ (Fisher's exact test) and the abundance change is $>5\%$.

Panel B - Screenshot of *RBCK1* as an example of a gene that shows a shift to shorter APA isoform, corresponding to the truncated protein, in the SW480 and SW620 cell lines when compared to 1CTs.

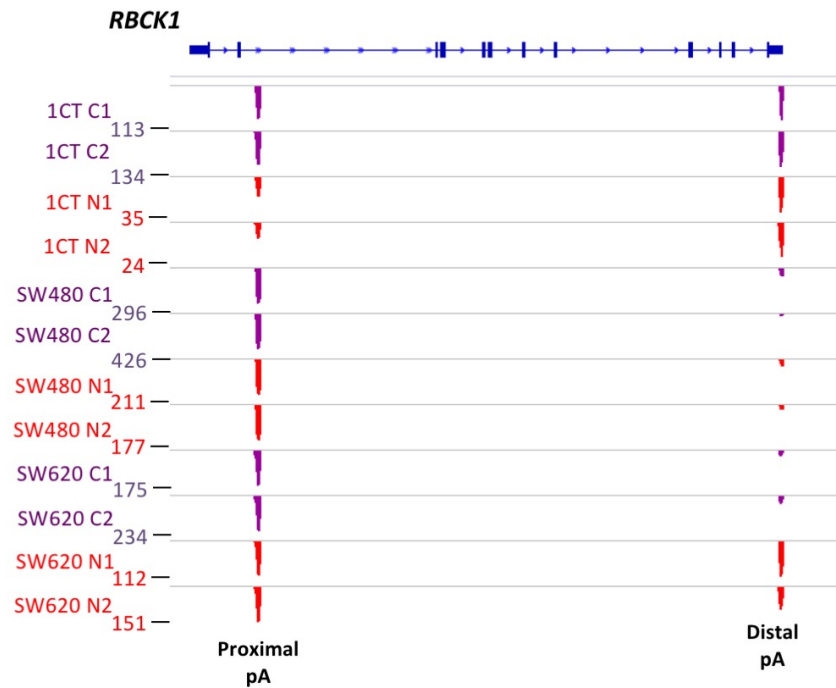
Panel C - All gene ontology “biological process” categories with significant enrichment ($p < 0.05$) of the genes that showed differential representation of CR-APA isoforms in the comparisons shown in Panel A. Those groups that are not shown did not display significant enrichment for any categories.

A

Coding region APA through colorectal cell lines



B



C

SW480 Nuclear vs 1CT Nuclear				
Shortening	Number found	Expected	Enrichment	P-value
DNA conformation change	12	2.36	5.08	4.36E-02
chromatin organization	18	4.85	3.71	1.62E-02
RNA processing	20	5.69	3.52	9.48E-03
gene expression	54	28.92	1.87	8.29E-03
cellular nitrogen compound biosynthetic process	50	26.18	1.91	1.36E-02
cellular macromolecule biosynthetic process	52	27.7	1.88	1.27E-02
regulation of macromolecule metabolic process	69	41.46	1.66	7.34E-03

6.3. DISCUSSION

6.3.1. *Introduction*

As discussed in Section 1.3.4, several studies have found large-scale shifts in APA profiles, when comparing cancerous to non-cancerous samples (Masamha et al., 2014; Mayr and Bartel, 2009; Xia et al., 2014). However, the identification that these APA events are occurring is only the beginning of identifying whether these events are truly contributing the observed phenotype or if they are merely a result of changing cellular conditions. To address this, this chapter applies the subcellular fractionation approach outlined in Section 1.10.2. This is applied to a cell line-based model of colorectal cancer to identify shifts in pA site choice throughout colorectal cancer progression and, using the comparison of cytoplasmic and nuclear APA profiles, isolate which of those events are resulting in the production of differentially regulated APA isoforms.

6.3.2. *Use of 1CTs, SW480s and SW620s as a model for colorectal cancer*

As discussed in Section 6.1.1, the use of 1CTs, SW480s and SW620s has been used extensively as a cell line-based model for colorectal cancer progression (Hewitt et al., 2000; Mouradov et al., 2014; Uzozie et al., 2014). 3'end sequencing of subcellular RNA fractions from these three cell lines yielded highly reproducible results (Figure 46A). Using this 3'end sequencing data to calculate gene expression levels, the top 50 genes that showed the highest degree of variability between the cytoplasmic samples were isolated (Figure 47). These 50 genes included a number of key genes that have previously been highlighted as altering gene expression levels upon the onset of colorectal cancer, as summarised in Table 12. Together, combined with previous studies, this gave confidence that this model is a valid one for studying colorectal cancer progression.

6.3.3. *General shifts in APA profiles through colorectal cancer progression*

Previous studies on colorectal cancer clinical samples found that during the progression from normal mucosa to adenomas there were shifts in APA profiles, with roughly equal numbers moving to either more proximal pA sites (105 of 216) or distal pA sites (111 of 216). However, from the progression of colorectal cancer from adenoma to carcinoma, there was a significant bias for shifts towards proximal pA sites (158/240) (Morris et al., 2012). This is in line with studies of other cancers types that see 3'UTR shortening upon cancer progression as discussed in Section 1.3.4.

The model for colorectal cancer progression seen here uses the 1CT cell line as the equivalent of a “normal” epithelial sample, while SW480s are derived from a primary adenocarcinoma of the colon and SW620s are derived from a secondary tumour after the primary tumour (from which SW480s were derived) had metastasized to the lymph nodes. Since this model does not contain an intermediary adenoma sample, it would be expected that the shift to proximal pA sites would be seen in both SW480s and SW620s when compared to 1CTs. When considering the nuclear fractions, which are taken as a more representative example of pA site choice (Li et al., 2015; Neve et al., 2016), a significant bias towards shifts to proximal APA isoforms is also seen in adenocarcinomas (SW480s and SW620s) when compared to the “normal” sample (1CTs) (Figure 48A). Although this bias is significant, there is also a large number of genes that shift in the opposite direction.

Furthermore, within each cell line the cytoplasmic and nuclear APA profiles were compared to identify the APA isoforms which differ in either their cytoplasmic stability or nuclear export rates. This analysis shows that 1CTs, SW480s and SW620s all follow the general trend that was discovered in Chapter 4, where proximal APA isoforms are generally overrepresented in the cytoplasm. Having the cytoplasmic and nuclear APA profiles for each of these colorectal cancer cell lines allows us to isolate those genes in where APA is likely to

be contributing to altered gene expression levels between these cell line. As discussed in Section 6.1, these genes fall into two categories:

1. Those genes which shift in pA site choice, thereby switching between differentially regulated APA isoforms.
2. Genes in which the APA isoforms are differentially regulated in one cell but not the other.

In regards to the genes that fit into the former category, the small minority of genes that displayed shifts in pA site choice were found to be occurring in genes which differed in either their cytoplasmic stability or nuclear export rates (Figure 51). In the case of UTR-APA, this indicates that the majority of shifts in APA profiles isolated here upon colorectal cancer progression are not occurring in genes where the corresponding APA isoforms are differing in cytoplasmic stability or nuclear export rates. This is not to say that these events are not being differentially regulated by some other means that would not be highlighted when comparing the APA profiles of the cytoplasm and the nucleus, such as different translational rates or specific localisation within the cytoplasm. The important aspect of this is that this intersection highlights the genes where APA is likely to be contributing to the regulation of gene expression and allows us to focus on these genes for further individual gene analysis.

The second category of regulation by APA is a more passive version, where, rather than changes in pA site choice, there is a change in the post-transcriptional mechanisms which govern the differential regulation. These, therefore, correspond to those genes that display differential representation of APA isoforms between the subcellular fractions of one cell line, but not the other. In line with the high degree of cell type specificity seen in Figure 22, the three colorectal cancer cell lines used in this study also exhibit a defined subset of genes where this differential regulation of APA isoforms is seen.

6.3.4. *Gene expression changes upon changes in pA site choice*

It has widely been proposed that APA provides a mechanism to switch between differentially regulated APA isoforms as a mechanism to control RNA levels, and thus the level of gene expression (Lutz and Moreira, 2010; Mayr and Bartel, 2009; Rehfeld et al., 2013). However, in the model used here, only a minority of APA events that are found to shift in their nuclear APA profiles between cell lines are also found to be differentially regulated by either cytoplasmic stability or nuclear export rates. However, in the small group of genes found where this is the case, these would be expected to be accompanied by a change in RNA levels in the cytoplasm. Analysis to see if this was indeed the case did show that in two out of the four gene cohorts, the changes in expression levels did display a significant bias in the expected direction (Figures 52A and 53A). The reason that all genes did not shift in the expected direction is likely because these APA events, and the differential regulation of the resulting APA isoforms, are only part of the processes which come together to determine the steady-state cytoplasmic RNA levels.

One of the key determinants of steady-state cytoplasmic RNA levels are the transcription rates of the gene themselves. Since we are comparing different cell lines with varying origins and phenotypes, it is unlikely that the transcriptional rates will be identical throughout these cell lines. Furthermore, transcription rates have been identified as the dominant determinant of steady-state RNA levels (Rabani et al., 2011). However, since we do see a significant bias in the expected direction in two out of the four gene cohorts suggests that these APA events are indeed contributing to the steady-state cytoplasmic RNA levels.

6.3.5. *Motif analysis of genes that show differential regulation of APA isoforms*

Motif analysis of the aUTRs of genes that showed differential regulation in each of 1CTs, SW480s and SW620s isolated a number of post-transcriptional regulatory pathways

which are potentially contributing to the observed differences in subcellular APA profiles. This analysis was simply applied to those genes that showed differential representation of APA isoforms in the cell line concerned, independent of their movement in other cell lines or between nuclear APA profiles. Therefore, this was purely aiming to isolate potential post-transcriptional regulatory mechanisms that were active in each cell line.

Interestingly, three of the miRNAs with the highest level of enrichment of target sites in 1CTs were miR-340, miR-506 and miR-124, which have previously been directly associated with inhibiting colorectal cancer progression (Chen et al., 2015; Sun et al., 2012) (Figure 54A). The z-scores for miR-340 in SW480s and SW620s are -0.12 and 0.57 respectively indicating that miR-340 target sites are not enriched in the aUTRs of genes that produce differentially regulated APA isoforms in SW480s or SW620s. For miR-124, the z-scores in SW480s and SW620s are 4.54 and 0.45 respectively, meaning that miR-124 target sites are enriched in the aUTRs of differentially regulated APA isoforms in SW480s, but not SW620s. This pattern of miR-124 target site enrichment is mirrored by that of miR-506.

This progressive loss of these target sites in the aUTRs of differentially regulated genes indicates that during colorectal cancer progression, these aUTRs are no longer being targeted by these miRNAs, and thus no longer being subject to differential regulation. Studies have shown that miR-124 and miR-506 expression are both reduced through colorectal cancer progression (Chen et al., 2015) which would explain the pattern observed here. Considering the previous association between these miRNAs and colorectal cancer, this is a very interesting result. To investigate this further, relative levels of these miRNAs will have to be determined in the cell lines used here to establish whether this observed pattern is indeed due to a drop in these miRNA levels when progressing from 1CTs to SW480s, and onto SW620s.

DREME motif analysis also highlighted sequences that were enriched in the aUTRs of differentially regulated APA isoforms when compared to the aUTRs of genes that were not

found to be differentially regulated, as in Section 6.2.9. Comparing these isolated motifs to known RNA-binding protein motifs identified several RNA-binding proteins that could be contributing towards the observed differences between cytoplasmic and nuclear APA profiles (Figure 54B). These included ESPR2 and SRSF1 that are known to result in the destabilisation of target RNAs (Ray et al., 2013). This analysis provides a good starting point to identify the post-transcriptional regulatory pathways that are active in these different cell lines. However, these pathways will have to be confirmed experimentally to elucidate the true extent to which these RNA-binding proteins are contributing to the observed differences between the cytoplasmic and nuclear APA profiles.

6.3.6. *APA events of interest isolated by the subcellular fraction approach*

The subcellular fraction approach presented here is ideal for identifying genes of interest where APA is likely to have an impact on gene expression. As discussed above, this has identified a number of small cohorts of genes. The next step is to consider the functions of the genes in these cohorts and identify, based on previous knowledge of these gene functions, whether these changes are likely contributing to the progression of colorectal cancer and the phenotypic differences seen between 1CTs, SW480s and SW620s.

Gene ontology analysis did highlight some enrichment of some biological processes categories when looking at the gene cohorts. When considering solely genes that were differentially regulated in the different cell lines, the most significantly enriched category is in SW620 in those genes involved in RNA processing, specifically, those genes that are involved in the spliceosome. This is in line with the gene ontology analysis performed in Section 4.2.7, Figure 24, wherein the only three cell lines where significant enrichment in biological process categories was found, each of those included RNA processing.

Furthermore, genes involved in RNA processing were enriched in those genes identified as showing both a change in nuclear APA profiles between SW480s and 1CTs as well as showing differential regulation in SW480s, as well as the equivalent genes when comparing SW620s to 1CTs. This constant enrichment for genes involved in RNA processing may reflect a mechanism in which RNA-binding proteins are autoregulatory, as discussed in Section 4.3.2.

In addition to the enriched Gene Ontology categories, some individual genes identified by this analysis provide an interesting insight into APA events that may be potentially contributing to the progression of colorectal cancer. *CCND1* was identified as a gene which shows differential regulation in both 1CTs and SW480s, but not SW620s (Figure 53B). *CCND1* is one of the most studied examples of where APA is thought to be contributing to carcinogenesis. This stems from the differential regulation of the two *CCND1* APA isoforms, whereby the distal isoform is destabilised as the aUTR contains miRNA target sites for miR-15/16 and AU-rich elements (Mayr and Bartel, 2009). The proposed mechanism for promoting carcinogenesis is a shift in pA site choice to the more stable proximal APA isoform, thus resulting in an increased protein production (Masamha et al., 2014; Mayr and Bartel, 2009; Wiestner et al., 2007). As cyclin D1 is a key regulator of cell cycle progression, specifically at the G₀/G₁ transition (Sherr, 1993), overexpression can lead to increased proliferation and reduced survival rates in cancer patients (Wiestner et al., 2007).

In contrast to this previously described mechanism of APA in *CCND1* contributing to carcinogenesis, a change in pA site choice in colorectal cancer progression is not seen here. However, from the APA profiles shown in Figure 53B, it appears that in 1CTs and SW480s the distal APA isoform of *CCND1* is either being degraded at a greater rate, or being exported from the nucleus at a slower rate, relative to the proximal APA isoform. In contrast to this, the cytoplasmic and nuclear APA profiles for *CCND1* in SW620s look almost identical indicating that the post-transcriptional regulatory mechanism that was causing the

difference in APA profiles between the cytoplasmic and nuclear fractions in 1CTs and SW480s is no longer active in SW620s. Despite this not being the same mechanism of pA site shifting as previously proposed (Mayr and Bartel, 2009), this relief of the differential regulation on *CCND1* in SW620s would have the same effect of increasing the levels of stable steady-state cytoplasmic *CCND1* RNA. Subsequent individual gene analysis, including stability experiments, in these different cell lines, would be required to confirm the exact mechanism of differential regulation in these cells. Furthermore, the reintroduction of the post-transcription regulatory pathway that is reduced in SW620s, such as overexpression of a specific miRNA, and observing any difference in phenotypes, will be key in elucidating whether this reduction in differential regulation of *CCND1* APA isoforms is indeed contributing to colorectal cancer progression.

SMAD3 is another gene isolated by the present analysis that is of interest in the context of colorectal cancer. When considering the progression from SW480s to the metastatic SW620s, *SMAD3* shows a shift in nuclear APA profile to favour the distal APA isoform in SW620s (Figure 52B). The comparison between the cytoplasmic and nuclear profiles of SW620s highlights that the APA isoforms of *SMAD3* are differentially regulated, with the distal APA isoform having either reduced cytoplasmic stability or nuclear export rate. This shift would be expected to result in an overall reduction in stable cytoplasmic levels of *SMAD3*, which is seen when comparing SW620 cytoplasmic levels compared to SW480 levels (Figure 52B). *SMAD3* protein is a transforming growth factor- β (TGF- β) receptor, and disrupting of the *Smad3* gene in mouse results in the onset of metastatic colorectal cancer after 4-6 months (Zhu et al., 1998). In normal epithelial tissue, TGF- β signalling provides a growth inhibitory signals, so downregulation or disruption of receptors on this pathway reduce or remove these inhibitory signals, thus increasing the rate of proliferation in these cells (Fleming et al., 2013). Once again, whether this APA event is responsible for the observed reduction in *SMAD3* cytoplasmic RNA levels in SW620s, and

whether this has any effect on SMAD3 protein levels and carcinogenesis, will have to be confirmed through individual gene analysis. However, this analysis presented here provides a solid starting point for establishing the true extent of the contribution of APA to colorectal cancer progression.

7. CHAPTER 7 – DISCUSSION

7.1. INTRODUCTION

This thesis has developed a pipeline that can be used to identify differentially regulated APA isoforms in any cell-based system where a clean subcellular fractionation can be obtained. Chapter 3 primarily focuses on the mapping of pA sites in the human genome and the sequence composition surrounding pA sites, considering the potential evolutionary origins and biological functions for having multiple pA sites within a single gene. Using both generic RNA sequencing and 3'-end-targeted sequencing, Chapter 4 then applies this subcellular fractionation pipeline to 9 human cell lines and identifies a large number of differentially regulated APA isoforms with a significant degree of cell type specificity. Subsequently, in Chapter 5, this pipeline is applied to a DICER1-depleted system in an attempt to identify miRNA-mediated differential regulation of APA isoforms. Unexpectedly, this resulted in the implication of DICER1 in affecting APA at the point of pA site choice, as well as through the miRNA-biogenesis pathway. Chapter 6 then goes on to apply this pipeline to observe the differentially regulated APA isoforms, alongside the changes pA site choice, in colorectal cancer progression.

In summary, the outcomes of this thesis are:

- A pipeline, involving subcellular fractionation followed by deep sequencing, has been developed
- This pipeline has been proven to be effective at isolating APA isoforms that differ in either their cytoplasmic stability or nuclear export rates
- In general, proximal APA isoforms are found to be overrepresented in the cytoplasmic APA profile, relative to the nuclear profile, across 11 of the 12 cell lines analysed

- Applying this subcellular fractionation to a cell line-based model of colorectal cancer, a small cohort of APA events that are likely to contribute to the regulation of gene expression have been isolated
- Nuclear DICER1 has been found to influence APA profiles at the point of pA site choice

This final discussion chapter will put these findings into the context of the current state of the APA field and put forward the outstanding questions that remain as well as the future directions of the present work.

7.2. THE PREVALENCE OF APA AS A METHOD OF REGULATING GENE EXPRESSION

When considering the number of genes in *C. elegans* versus humans, the numbers are relatively similar. However, when comparing the average length of UTRs, there is a significant increase, from the average UTR in *C. elegans* being 140 nt to 1200nt in humans (Jan et al., 2011). It is generally accepted that organisms with increased complexity have longer UTRs due to a higher degree of post-transcriptional regulation. Indeed, as discussed in Section 1.1, 3'UTRs are well known to influence mRNA stability, translational efficiency, and localisation. The subsequent discovery that around 70% of human genes contain multiple pA sites led to the general proposed model that is seen today: a gene can contain multiple pA sites and switching between these pA sites either includes or excludes *cis*-regulatory elements in the 3'UTR, thereby regulating gene expression via APA. This led to a flurry of activity of mapping pA site usage on a transcriptome scale in different cellular contexts, and claiming that the observed large-scale changes in APA profiles are potentially contributing to the observed phenotype (Section 1.3).

However, looking from a different perspective, Section 1.4.2 highlights that the sequence requirements to direct cleavage and polyadenylation are relatively low and the

probability of random sites being capable of directing cleavage and polyadenylation is quite high. It is therefore somewhat unsurprising that we do see multiple pA sites per gene. Furthermore, if we consider the gene *FOXP3*, we can see why it may be beneficial to have multiple pA sites in a single gene. *FOXP3* contains a single pA site and mutation in the PAS of this pA site from AAUAAA to AAUGAA results in inefficient transcription termination, reduced *FOXP3* expression and ultimately IPEX syndrome (Bennett et al., 2001). Therefore, it is easy to see how the introduction of a second pA site, simply as a fail-safe mechanism, would be evolutionarily favourable.

Evolutionary conservation of pA sites can tell us a lot about their importance. Looking at the conservation of tandem sites relative to single pA sites shows that the single pA sites are generally more highly conserved indicating that most of the novel pA sites are either neutral or deleterious (Ara et al., 2006). Furthermore, Section 3.2.7 shows that where tandem pA sites are present, the distal pA sites are generally stronger. The presence of a strong downstream pA site would reduce the selection pressure on the proximal pA site, resulting in these weaker proximal sites being less conserved, which is indeed the case (Ara et al., 2006). Furthermore, the strength of a pA site may itself define it as the most distal pA site, as through directing efficient cleavage and polyadenylation there would be little chance for the transcript to use any downstream would-be distal transcripts. Together, this would tend towards a layout as seen here, and elsewhere (Beaudoing, 2000), where many genes contain multiple pA sites with the distal pA site generally directing more efficient cleavage and polyadenylation than the proximal pA site.

Therefore, even without the introduction of regulating gene expression via the differential regulation of APA isoforms, a large amount of the APA landscape can be explained simply by considering the evolutionary requirements to ensure efficient cleavage and polyadenylation, and thus transcriptional termination. However, this fail-safe principle by no means negates that APA plays an important role in the regulation gene expression as there

are numerous studies confirming its importance at the single gene level (Section 1.5). It is, however, important to acknowledge that even though APA is highly prolific and occurs in around 70% of genes, it may be too simplistic to conclude that all or the majority of these events are having any functional and physiological consequence.

As discussed in Chapter 6, there are two distinct mechanisms by which APA could potentially regulate gene expression levels:

1. By changing pA site choice to switch between two differentially regulated APA isoforms.
2. By changing the post-transcriptional regulatory mechanism while maintaining the pA site choice. This would, therefore, lead to a change in the differential regulation of the corresponding APA isoforms.

To date, studies have somewhat blurred these, as, using whole-cell RNA, changes in APA profiles caused by these two mechanisms cannot be distinguished. The former mechanism is the classical mechanism by which gene expression can be regulated via APA, as gene expression levels can be regulated through simply switching between differentially regulated APA isoforms. The latter is a combination of gene regulation via APA and changes in the accompanying post-transcriptional regulatory pathways.

To examine both of these mechanisms, the key is isolating those APA events that are producing differentially regulated APA isoforms. This thesis employs a novel subcellular fractionation approach as described in Section 1.10.2. Chapter 4 uses 3'READS to map and isolate regulated APA events in HEK293 and HBL melanocyte cell lines and further exploited publically available RNA-sequencing data from the subcellular fractions of 9 different cell lines. Consistently, across these 11 different cell lines, it was generally found that there was a significant overrepresentation of the proximal APA isoforms in the cytoplasm (Sections 4.2.1, 4.2.4 and 6.2.3). This would equate to either distal APA isoforms that have lower cytoplasmic stability or lower nuclear export rates. These findings correlate

well with previous stability studies performed in mice (Gruber et al., 2014; Spies et al., 2013) and together indicate that the pervasiveness of APA isoforms differing in either their cytoplasmic stability or nuclear export rates is restricted to a small subset genes.

However, as seen in Section 4.2.6, with the comparison of differentially regulated APA isoforms in the nine different cell lines, there is a large degree of cell type specificity. This means that even if a single gene is undergoing APA in one cell line, but the resulting isoforms are not found to be differentially regulated, it by no means indicates that that APA event does not have some regulatory function in either a different cell type or different cell state. With the introduction and removal of different post-transcriptional regulatory pathways, e.g. due to changes in miRNA profiles between cells, this is exactly what would be expected.

In Chapter 6, the contribution of APA to gene expression was examined in a cell-line-based model for colorectal cancer progression. As this is a dynamic system, with multiple states of cancer progression it allowed us to examine gene regulation of APA through the former mechanism described above: changes in pA site choice between differentially regulated APA isoforms. Figure 50 isolated those genes that both shifted in their nuclear APA profile and the APA isoforms were differentially regulated. However, this was a very restricted set of genes, and the majority of nuclear APA shifts between the different colorectal cell lines were occurring in genes where the APA isoforms are not differentially regulated.

Similar to that seen when comparing the different cell lines from the ENCODE database (Section 4.2.6), when comparing the three colorectal cell lines (Section 6.2.6), it was found that there is a large degree of cell type specificity of those genes that display differential regulation of APA isoforms, as exemplified by *RAC1* and *CCND1* (Figure 53). The important point here is that the APA events that do result in differential regulation need

to be isolated in each experimental model to identify those that it will be interesting to focus on. The subcellular fractionation pipeline developed here enables exactly that.

7.3. THE SUBCELLULAR FRACTIONATION APPROACH

The correlation of the pervasiveness of APA producing differentially regulated APA isoforms between this subcellular fractionation approach and previous approaches (Spies et al., 2013) was a promising finding in regards to the effectiveness our approach. However, one key restriction with this approach is that, unlike transcriptional arrest-based approaches, it cannot give definitive values to the relative stabilities of the APA isoforms. The subcellular fractionation approach can simply tell if there is a differential representation of the APA isoforms in the different subcellular fractions. As the level of differential representation is dependent on multiple factors, including the absolute cytoplasmic stability of each isoform and the rate in which these isoforms are being transcribed and exported from the nucleus, it is impossible to calculate the absolute values or even relative values of differential regulation of APA isoforms. Additionally, the level of differential regulation of these APA isoforms must themselves be significant enough to manifest as being differentially represented in the different subcellular fractions. This is an important aspect to consider, as many mechanisms of post-transcriptional regulation have been classified as “fine-tuning” (P. Yao et al., 2012). In the cases where a low level of differential regulation is seen, it would be unlikely that these events would result in a high enough level of differential representation to be isolated using this subcellular fractionation approach.

One other caveat of previous studies that has been highlighted by this work (Section 1.7), and previous studies (Elkon et al., 2013; Li et al., 2015; Lianoglou et al., 2013) is the study of APA using RNA isolated from the whole-cell. RNA isolated from the whole-cell essentially reflects the steady-state APA landscape rather than the actual profile of pA site

choice. This study and others (Li et al., 2015) have proposed the use of the nuclear APA profile as a method of obtaining an APA profile that is more reflective of the actual pA site choice. The premise of this is that the majority of post-transcriptional regulation that occurs in the cell takes place in the cytoplasm, so using the nuclear APA profiles will mitigate the effect of this cytoplasmic differential regulation. However, there are still post-transcriptional regulatory mechanisms, such as nuclear retention of specific APA isoforms, which can result in the nuclear APA profiles being skewed away from the true pA site choice. For example, this would be the case with *CIAOI* as found here (Section 4.2.10), where the distal APA isoform does not appear to be efficiently exported from the nucleus. This would, therefore, result in an accumulation of solely the distal APA isoform in the nucleus and alter the nuclear APA profile away from the true pA site choice profile. Therefore, nuclear APA profiles are still not a perfect measure of pA site choice, but they do mitigate a significant portion of the post-transcriptional regulatory mechanisms that alter the APA profiles of RNA isolated from the whole-cell.

Finally, there are a couple of other limitations of the subcellular fractionation approach that should be highlighted in its use in a dynamic system. Chapter 6 demonstrated the effectiveness of the approach in isolating events that shift in nuclear APA profiles during colorectal cancer progression, and also show a differential representation of APA isoforms between the cytoplasmic and nuclear fractions. This corresponds to genes thought to shift in between APA isoforms that are differentially regulated. However, using this approach, there is a chance that some genes which should fall into this category will be overlooked. For example, if considering a complete switch to a more stable proximal pA site when comparing one nuclear fraction to another, there will be no remaining distal APA isoform being produced to highlight any differences between nuclear and cytoplasmic APA isoform ratios, and thus not be isolated as being differentially regulated. However, when visually inspecting genes that shift in their nuclear APA profiles, these complete switches which result in the

complete cease in production of the distal isoform are rare. It is, therefore, thought that only a small number of genes, if any, will be overlooked due to this complication.

Secondly, in a situation where the pA site choice remains constant between the two cell lines being compared, but the rate of nuclear export of one of the APA isoforms changes, this will result in a change in APA profiles when comparing the nuclear fractions as well as a change in the representation of APA isoforms in the cytoplasm versus the nucleus. This would, therefore, include genes in those highlighted in Figure 51, even though they do not alter their pA site choice. This comes back to the fact that even the nuclear APA profile is a steady-state profile. Despite these events not changing in pA site choice, they would still be of interest as there is a change in the levels of differential regulation, which itself is would have an impact on gene expression.

7.4. MECHANISMS INVOLVED IN DIFFERENTIAL REGULATION OF APA ISOFORMS

The differential regulation of APA isoforms can occur through a range of mechanisms as discussed in Section 1.5. The standard method used to attempt to identify the mechanisms of differential regulation is to look for the enrichment of specific motifs that may be being targeted by either RNA binding proteins or miRNAs, as has been done in Sections 4.2.8, 4.2.9 and 6.2.9. In Section 4.2.10, a putative mechanism of differentially regulating APA isoforms was also identified with intron retention appearing to be specifically coupled to the distal APA isoform in *CLAO1*. The mechanism by which this occurs is unknown but similar RNA-seq profiles that show both 3'UTR extension and intron retention have been seen in multiple other genes (Figure 30).

This brings us back to one of the limitations of all 3'end sequencing techniques: the identity of the upstream transcript is largely unknown. This technical limitation of the short read length associated with the majority of deep sequencing technologies makes reassembling

transcripts a major obstacle. Therefore, identifying common events, such as the intron retention and distal pA site selection seen here, that occur in the same transcript is a major obstacle. The only way of truly identifying the exact transcript structure on a transcriptome scale will be to sequence the full-length RNA as a single molecule. The prospect of doing this is becoming more feasible (Sharon et al., 2013) and the application of these approaches to study APA will be key in the identification of the pervasiveness of coupling intron retention to specific pA site choice.

7.5. OBTAINING SPECIFICITY IN REGULATING PA SITE CHOICE

To effectively regulate a gene through UTR-APA, there are several requirements that need to be fulfilled. Firstly, there needs to be a stimulus or change in cellular condition that results in a regulated switch between pA sites in a specific gene or subset of genes. The aUTR, that is either included or excluded by the shift in pA sites, must then contain one or more *cis*-elements that confer differential regulation of the resulting APA isoforms. This differential regulation of APA isoforms would then result in a change in the level of corresponding protein expressed.

This project has primarily focussed on the identification of differentially regulated APA isoforms and the potential mechanisms by which this differential regulation occurs. However, the regulated switch between pA sites is the other key aspect required to regulate gene expression via APA. The majority of studies that have looked at this do so by comparing the global APA profiles before and after depleting a factor thought to be involved in regulating pA site choice. Using this approach, an extensive list of factors thought to be involved in regulating APA has been generated (Tables 2-4). Furthermore, Chapter 5, unexpectedly adds DICER1 to this list of factors influencing APA profiles at the point of pA

site choice, from which the data generated here, alongside previous studies (Skourti-Stathaki et al., 2014), appears to act through the modification of chromatin conformation (Figure 42).

This leads to one key point that is often neglected. There is a significant difference between influencing pA site choice and regulating pA site choice (Section 5.3.3), and when considering this difference, it is important to contemplate how specificity is achieved. When comparing APA with a different mechanism of controlling gene expression, such as transcriptional initiation, the achievement of specificity is quite clear: it does so through a combination of sequence-specific transcription factors (Hargreaves et al., 2009). These transcription factors bind to *cis*-elements around the promoters of a very defined subset of genes and stimulate transcription initiation at those promoters. These genes are usually involved in a specific cellular process or pathway.

In an analogous situation, pA sites would be surrounded by specific *cis*-elements that are targeted by RNA-binding proteins, which themselves either promote or inhibit cleavage and polyadenylation at specific sites. These precise *cis*-elements would then be only present at a subset of pA sites that are to be regulated together e.g. in response to some external stimuli. This is indeed found in some cases, such as the regulation of APA through CPEB1 binding CPE sites and regulating cleavage and polyadenylation at surrounding pA sites (Bava et al., 2013).

The issue with some of the pA factors that have been put forward as **regulators** of APA is the ubiquity in which they act. Factors such as CF_{Im}25 (Masamha et al., 2014), CstF64 (Hwang et al., 2016; C. Yao et al., 2012) and CstF77 (Luo et al., 2013) are key to the efficiency of the cleavage and polyadenylation event as a whole (Mandel et al., 2008). Therefore, any depletion in the levels to a point where they become the limiting factor would be expected to alter both the pA site usage as a whole, reducing pA at the weaker pA sites, as well as potentially increasing the amount of read-through transcripts which are never fully processed. The large fluctuations in pA site choice are indeed seen in each of the studies

depleting these central pA factors, with nearly 1500 genes switching pA site choice in some cases (Masamha et al., 2014). There is no doubt that fluctuations in these central pA factors influence APA profiles, but it is hard to imagine how the cell would actively regulate APA at specific genes through these factors.

One possibility is that, as shown throughout this thesis, the majority of UTR-APA events do not actually result in differentially regulated APA isoforms, so switching between these APA isoforms would not affect the expression of these genes. Therefore, even though large scale APA shifts are triggered by these regulatory APA factors, only a small subset of genes are going to be shifting between differentially regulated APA isoforms. The majority of events will be shifting between APA isoforms that display no difference in their regulation, and can, therefore, be considered collateral to the differentially regulated APA isoforms which do display a shift in pA site choice.

This is largely what is seen when looking at colorectal cancer progression, where the large majority of genes that shift in the nuclear APA profiles of the colorectal cell lines are not the genes where differential regulation of APA isoforms is seen. This indicates that these APA events are unlikely to be contributing to any phenotypic changes that are seen between the cell lines. The shift between these pA sites may simply be a consequence of the changing levels of pA factors required to regulate the pA site choice in genes which do show differential regulation of APA isoforms, such as *SMAD3* identified here (Figure 52B).

This is similar to what would be expected with the CR-APA of IgM in pre-B cells and mature B cells. In pre-B cells CstF-64 is low resulting in the upstream intronic pA site being missed, where in mature B-cells CstF-64 levels are much higher, resulting in the intronic pA site outcompeting the 5'splice site ultimately producing an IgM protein which is C-terminally truncated and secreted from the cell (Takagaki et al., 1996). As CstF64 levels have been shown to alter APA at a significant number of genes, and not just IgM (C. Yao et al., 2012), it would be expected that the other events that occur as a result of the changes in CstF64

levels would just be a secondary consequence of regulating IgM protein isoforms via APA, and potentially have no physiological impact.

7.6. REGULATION OF GENE EXPRESSION VIA POST-TRANSCRIPTIONAL MECHANISMS

Taking a step back from APA and looking at post-transcriptional regulation as a whole, there have been some interesting transcriptome-wide studies that put post-transcriptional regulation into perspective when compared to other methods of regulating gene expression. Firstly, it has been well established that in almost every organism that has been examined, that there is a large degree of discrepancy between steady-state mRNA levels and protein levels (de Sousa Abreu et al., 2009). Looking at the contribution of the different RNA features to protein abundance variation, one study estimated that ~8% was due to characteristics of the 3'UTR, including length, nucleotide composition and secondary structures (Vogel et al., 2010). However, it is important to realise that this ~8% is an average across all genes, and in some genes the 3'UTR can be a major determinant in protein abundance.

A subsequent study simultaneously measured protein and RNA abundances in mouse fibroblast cells, as well as determining mRNA half-lives on a transcriptome scale (Schwanhäusser et al., 2011). Here they conclude that of the variance in protein levels that is determined by mRNA levels, most is due to changes in transcriptional rates whereas mRNA stability has a much smaller role (Schwanhäusser et al., 2011). Once again, this is not to say that APA is not a key regulator of gene expression, but it simply highlights that the pervasiveness may not be as extensive as previously thought, and a more focussed approach that aims to isolate the genes where APA is regulating expression will be key in future studies.

7.7. CONCLUSIONS AND FUTURE DIRECTIONS

The advent of high-throughput sequencing technologies has no doubt made a huge contribution to our knowledge of APA, as well as molecular genetics as a whole. It has allowed us to map pA sites on a transcriptome scale as well as their relative usage in different cellular states. However, the study of all genes at once can lead to the focus on identifying general trends that apply to large subsets of genes rather than considering that each gene will likely be unique in its function and regulation. APA is a classic example of this, where there are endless papers that describe “global” 3’UTR lengthening and shortening in different cellular contexts, while often never relating these genes back to a potentially physiological function. This focus on bioinformatic correlation rather than direct functional experimentation has previously been previously highlighted (Proudfoot, 2011), emphasising the fact that the true value of these transcriptome approaches can only be realised when combined with direct experimentation at an individual gene level.

The pipeline generated here allows the use of transcriptomic studies to isolate the genes in which APA is likely to play a major role in their regulation, which can then be investigated on an individual gene level. Therefore, applying this pipeline to the numerous examples of where a change in cell state or cell type results in large-scale changes in APA profiles, will highlight those events in which APA shifts also result in a change in RNA metabolism. It will be these events that will be most likely to have the key physiological impact on the cell, and only by isolating these events will the true pervasiveness of APA regulating gene expression be elucidated.

8. CHAPTER 8 – APPENDICES

Appendix Table 1 - Oligonucleotides used throughout the study

Oligo name	Sequence
RT-PCR primers	
<i>MALAT1</i> forward	CCTGCAAATTGTTAACAGAA
<i>MALAT1</i> reverse	TCAGCTTCCGCTAAGATGCTAGCTT
<i>GAPDH</i> forward	GGTGAAGGTCGGAGTCAACG
<i>GAPDH</i> reverse	TGGTTCACACCCATGACGAA
<i>DIDO1</i> forward	TGGGAAATTGCCTACGGAGC
<i>DIDO1</i> reverse	AGGCGAAACTGGACCAACAA
<i>ETNK1</i> forward	TGTTGAATTGGGAAATTGTGGCAT
<i>ETNK1</i> reverse	TTTTTTTTTTTTTTGTAGTTAGAACTCAAC
Oligo(dT) (AP) RT primer	GCCACGCGTCGACTAGTAC(T) ₁₅
Primers for <i>CIAO1</i> northern probe	
cUTR forward	TCCCAACTCGGATCCACAGC
cUTR reverse	CAAAGGCGGGGAAGGAAAGC
aUTR forward	GGCCACAAGCCCTGTATTCT
aUTR reverse	GTACAGCCCCACTGTCCCTC
intronic forward	GAGTCAGAGCAAATGGGGGT
intronic reverse	ACTGGCGTTTCTCAAGCAGA
Primers for <i>GLUL</i> northern probe	
cUTR forward	ACTAAGCAAGCGGCACCAGT
cUTR reverse	ACTAGGAGGGGCTCAACAGC
aUTR forward	TTAATGGGGTGCACCTGTCC
aUTR reverse	AGGGCATTTAGTCCCAAGCC
Primers for <i>IQGAP1</i> northern probe	
forward	TCAGAATGGGTCAAGGCACC
reverse	CGTGTAACATAGGTTGGCTGG
Primers for <i>ETNK1</i> northern probe	
forward	CGTTCAGGATCAGGAGGAGC
reverse	GGGTTGCAGACATGCTTTGG
RT-qPCR primers	
Phased oligo d(T)	TTCGTATACCCGGGTACCAA(T) ₁₅ N
<i>ETNK1</i> a-forward	AAACACCATGGAGGATGTAG
<i>GAPDH</i> b-forward	AAGGAAACTGTCTCGCATAG
linker	TTCGTATACCCGGGTACCAA

Appendix Table 2 - ChIP primers for ETNK1

Amplicon number		Sequence (5' -3')
1	fwd.	TGGGGTTATTTTAGAACGTG
	rev	GAAGTGTCTAGTCTGGAAAGC
2	fwd.	GGTGGATCAATCCCTTAGTC
	rev	AATCTGCAGTTGCTTGAATC
3	fwd.	TCCCATTCTCAGAGACTTTC
	rev	CCTTCCTTGATCCAAAGTG
4	fwd.	AAACACCATGGAGGATGTAG
	rev	GATCCAGTGCTTCTCCTTG
5	fwd.	TTATACAAGGAGAAGCACTGG
	rev	ACAAAAACAAAAACCCTAGC
6	fwd.	AATTGTCTTTGTTTGAGCAAC
	rev	AAAATAAGCAGAAGCCTTG
7	fwd.	CAGAGGGTCAAGGCTTCTG
	rev	AATAGGCAAGAACTCTGAAAAC
8	fwd.	ACCAACTTCTTTTCATCACC
	rev	CTCCACTCTGCTTCTCCTC
9	fwd.	ACTGAAGGACATTTTGGTTG
	rev	AAAATAGCACAGCCATTCTG
10	fwd.	GGTTCAAAAAGAGGGATTTC
	rev	GCTGGCAAGAATTCTGATAC
11	fwd.	TGCAACTTTTGATGTAGGTG
	rev	ACTGCCAACCTCATTTTATC
12	fwd.	TGGTTCCTTTTGCTATGTTC
	rev	TATCTCAGGAGCCAATAAGC
13	fwd.	AGGGGTCTCTTATACCCTTG
	rev	AGTAGAGATGGGGTTTCACC
14	fwd.	CAGAGTCTCGCTCTATCACC
	rev	GACTATCCTGGCTAACATGG
15	fwd.	TGTCTCTCCTTGTTTCTTCC
	rev	TGAGAAAGTCGTAGGGAAAC
16	fwd.	GATGATCTCCATCTCCTGAC
	rev	CGTTTGAACAACCTCTTTC
17	fwd.	CTGCCATTGAGAGTTGATGC
	rev	GTCAGTGAACCTGGCACA
18	fwd.	GAATTCAAGGCTTCAGTGAG
	rev	ACAAAAGCCAAGATTCATTC
GAPDH	fwd.	AACCTGCCAAATATGATGAC
	rev	AGGAAATGAGCTTGACAAAG
HPRT1	fwd.	AGATGTGATGAAGGAGATGG
	rev	AATAGCTCTTCAGTCTGATAAAATC
18S	fwd.	GCAAAGCTGAACTTAAAGG
	rev	CGTTCGTTATCGGAATTAAC

Appendix Table 3 - 3'end-sequencing data generated for this study

Sample	Raw Sequencing Reads	PASS reads
HEK293 Cytoplasmic Replicate 1	6,116,451	1,017,560
HEK293 Cytoplasmic Replicate 2	5,458,357	1,928,557
HEK293 Nuclear Replicate 1	9,369,217	2,134,122
HEK293 Nuclear Replicate 2	5,458,357	1,364,131
HEK293 Whole-cell Replicate 1	6,554,956	6,554,956
HEK293 Whole-cell Replicate 2	5,458,357	1,740,946
DICER1-depleted HEK293 Cytoplasmic Replicate 1	5,739,705	929,237
DICER1-depleted HEK293 Cytoplasmic Replicate 2	22,790,037	9,503,738
DICER1-depleted HEK293 Nuclear Replicate 1	6,276,626	1,106,995
DICER1-depleted HEK293 Nuclear Replicate 2	6,645,798	2,677,314
DICER1-depleted HEK293 Whole-cell Replicate 1	8,333,313	2,665,862
DICER1-depleted HEK293 Whole-cell Replicate 2	21,658,762	9,836,479
SW480 Cytoplasmic Replicate 1	4,784,124	2,637,949
SW480 Cytoplasmic Replicate 2	4,239,572	2,592,559
SW480 Nuclear Replicate 1	7,377,833	1,693,540
SW480 Nuclear Replicate 2	6,839,043	1,988,556
SW620 Cytoplasmic Replicate 1	5,086,506	2,240,747
SW620 Cytoplasmic Replicate 2	4,384,813	1,943,469
SW620 Nuclear Replicate 1	4,409,151	964,010
SW620 Nuclear Replicate 2	5,906,318	1,473,637
1CT Cytoplasmic Replicate 1	5,617,415	3,038,097
1CT Cytoplasmic Replicate 2	2,814,979	1,427,461
1CT Nuclear Replicate 1	4,185,692	1,070,423
1CT Nuclear Replicate 2	4,708,162	1,058,590

Appendix Table 4 - Publicly available RNA-seq datasets used in this study

Sample	GEO Accession Number	Raw Sequencing Reads
cyto_0_rp2	GSM1419118	26,359,043
cyto_0_rp3	GSM1419119	28,444,972
nucleus_0_rp1	GSM1419114	29,126,599
nucleus_0_rp2	GSM1419115	35,611,580
nucleus_12_rp1	GSM1419132	30,241,963
nucleus_12_rp2	GSM1419133	35,999,565
cyto_12_rp1	GSM1419135	35,999,565
cyto_12_rp3	GSM1419137	28,077,089
wgEncodeCshlLongRnaSeqK562CytosolPapAlnRep2	GSM767849	82,063,781
wgEncodeCshlLongRnaSeqK562NucleusPapAlnRep2	GSM765387	35,675,086
wgEncodeCshlLongRnaSeqHela3CytosolPapAlnRep2	GSM765404	39,455,691
wgEncodeCshlLongRnaSeqHela3NucleusPapAlnRep2	GSM765403	24,669,404
wgEncodeCshlLongRnaSeqHepg2CytosolPapAlnRep1	GSM758576	37,946,986
wgEncodeCshlLongRnaSeqHepg2CytosolPapAlnRep2	GSM758576	37,394,670
wgEncodeCshlLongRnaSeqHepg2NucleusPapAlnRep1	GSM758568	34,183,090
wgEncodeCshlLongRnaSeqHepg2NucleusPapAlnRep2	GSM758568	28,534,567
wgEncodeCshlLongRnaSeqHuvecCytosolPapAlnRep3	GSM758569	40,586,973
wgEncodeCshlLongRnaSeqHuvecCytosolPapAlnRep4	GSM758569	37,081,549
wgEncodeCshlLongRnaSeqHuvecNucleusPapAlnRep3	GSM758565	38,830,393
wgEncodeCshlLongRnaSeqHuvecNucleusPapAlnRep4	GSM758565	37,693,376
wgEncodeCshlLongRnaSeqImr90CytosolPapAlnRep1	GSM981244	114,006,059
wgEncodeCshlLongRnaSeqImr90CytosolPapAlnRep2	GSM981244	144,283,810
wgEncodeCshlLongRnaSeqImr90NucleusPapAlnRep1	GSM981248	167,746,727
wgEncodeCshlLongRnaSeqImr90NucleusPapAlnRep2	GSM981248	171,005,868
wgEncodeCshlLongRnaSeqMcf7CytosolPapAlnRep3	GSM981252	164,845,832
wgEncodeCshlLongRnaSeqMcf7CytosolPapAlnRep4	GSM981252	157,417,872
wgEncodeCshlLongRnaSeqMcf7NucleusPapAlnRep3	GSM981245	148,437,796
wgEncodeCshlLongRnaSeqMcf7NucleusPapAlnRep4	GSM981245	141,727,117
wgEncodeCshlLongRnaSeqSknshCytosolPapAlnRep3	GSM981251	241,123,024
wgEncodeCshlLongRnaSeqSknshCytosolPapAlnRep4	GSM981251	230,972,569
wgEncodeCshlLongRnaSeqSknshNucleusPapAlnRep3	GSM981250	149,659,336
wgEncodeCshlLongRnaSeqSknshNucleusPapAlnRep4	GSM981250	243,453,503

9. REFERENCES

- Akman, B.H., Can, T., Erson-Bensan, a E., 2012. Estrogen-induced upregulation and 3'-UTR shortening of CDC6. *Nucleic Acids Res.* 40, 10679–88. doi:10.1093/nar/gks855
- Al-Ahmadi, W., Al-Ghamdi, M., Al-Haj, L., Al-Saif, M., Khabar, K.S. a, 2009. Alternative polyadenylation variants of the RNA binding protein, HuR: abundance, role of AU-rich elements and auto-Regulation. *Nucleic Acids Res.* 37, 3612–24. doi:10.1093/nar/gkp223
- Alkan, S. a, Martincic, K., Milcarek, C., 2006. The hnRNPs F and H2 bind to similar sequences to influence gene expression. *Biochem. J.* 393, 361–371. doi:10.1042/BJ20050538
- Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., Sharp, P. a, 2013. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* 499, 360–3. doi:10.1038/nature12349
- Alt, F.W., Bothwell, a L., Knapp, M., Siden, E., Mather, E., Koshland, M., 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.
- Ameres, S.L., Martinez, J., Schroeder, R., 2007. Molecular Basis for Target RNA Recognition and Cleavage by Human RISC. *Cell* 130, 101–112. doi:10.1016/j.cell.2007.04.037
- An, J., Zhu, X., Wang, H., Jin, X., 2013. A dynamic interplay between alternative polyadenylation and microRNA regulation: Implications for cancer (Review). *Int. J. Oncol.* 995–1001. doi:10.3892/ijo.2013.2047
- 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., Xu, B., 2008. Distinct role of long 3' UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. *Cell* 134, 175–87. doi:10.1016/j.cell.2008.05.045
- Andreassi, C., Riccio, A., 2009. To localize or not to localize: mRNA fate is in 3'UTR ends. *Trends Cell Biol.* 19, 465–74. doi:10.1016/j.tcb.2009.06.001
- Ara, T., Lopez, F., Ritchie, W., Benech, P., Gautheret, D., 2006. Conservation of alternative polyadenylation patterns in mammalian genes. *BMC Genomics* 7, 189. doi:10.1186/1471-2164-7-189
- Arhin, G.K., Boots, M., Bagga, P.S., Milcarek, C., Wilusz, J., 2002. Downstream sequence elements with different affinities for the hnRNP H/H' protein influence the processing efficiency of mammalian polyadenylation signals. *Nucleic Acids Res.* 30, 1842–50.
- Baek, D., Villén, J., Shin, C., Camargo, F.D., Gygi, S.P., Bartel, D.P., 2008. The impact of microRNAs on protein output. *Nature* 455, 64–71. doi:10.1038/nature07242
- Bagga, P.S., Arhin, G.K., 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 Res.* 26, 5343–5350. doi:gkb833 [pii]
- Bail, S., Swerdel, M., Liu, H., Jiao, X., Goff, L. a, Hart, R.P., Kiledjian, M., 2010. Differential regulation of microRNA stability. *RNA* 16, 1032–9. doi:10.1261/rna.1851510
- Bailey, T.L., 2011. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics.* doi:10.1093/bioinformatics/btr261

- Bakheet, T., Williams, B.R.G., Khabar, K.S. a, 2006. ARED 3.0: the large and diverse AU-rich transcriptome. *Nucleic Acids Res.* 34, D111–4. doi:10.1093/nar/gkj052
- Barabino, S.M., Hubner, W., Jenny, a, Minvielle-Sebastia, L., Keller, W., 1997. The 30-kD subunit of mammalian cleavage and polyadenylation specificity factor and its yeast homolog are RNA-binding zinc finger proteins. *Genes Dev.* 11, 1703–1716. doi:10.1101/gad.11.13.1703
- Barnhart, M.D., Moon, S.L., Emch, A.W., Wilusz, C.J., Wilusz, J., 2013. Changes in Cellular mRNA Stability, Splicing, and Polyadenylation through HuR Protein Sequestration by a Cytoplasmic RNA Virus. *Cell Rep.* 1–9. doi:10.1016/j.celrep.2013.10.012
- Bartel, D.P., Lee, R., Feinbaum, R., 2004. MicroRNAs : Genomics, Biogenesis, Mechanism, and Function. *Cell* 116, 281–297.
- Batra, R., Charizanis, K., Manchanda, M., Mohan, A., Li, M., Finn, D.J., Goodwin, M., Zhang, C., Sobczak, K., Thornton, C.A., Swanson, M.S., 2014. Loss of MBNL Leads to Disruption of Developmentally Regulated Alternative Polyadenylation in RNA-Mediated Disease. *Mol. Cell* 56, 311–322. doi:10.1016/j.molcel.2014.08.027
- Bava, F.-A., Eliscovich, C., Ferreira, P.G., Miñana, B., Ben-Dov, C., Guigó, R., Valcárcel, J., Méndez, R., 2013. CPEB1 coordinates alternative 3'-UTR formation with translational regulation. *Nature* 495, 121–125. doi:10.1038/nature11901
- Beaudoing, E., 2000. Patterns of Variant Polyadenylation Signal Usage in Human Genes. *Genome Res.* 10, 1001–1010. doi:10.1101/gr.10.7.1001
- Beaulieu, Y.B., Kleinman, C.L., Landry-Voyer, A.-M., Majewski, J., Bachand, F., 2012. Polyadenylation-dependent control of long noncoding RNA expression by the poly(a)-binding protein nuclear 1. *PLoS Genet.* 8, e1003078. doi:10.1371/journal.pgen.1003078
- Beck, A.H., Weng, Z., Witten, D.M., Zhu, S., Foley, J.W., Lacroute, P., Smith, C.L., Tibshirani, R., van de Rijn, M., Sidow, A., West, R.B., 2010. 3'-end sequencing for expression quantification (3SEQ) from archival tumor samples. *PLoS One* 5, e8768. doi:10.1371/journal.pone.0008768
- Beisang, D., Bohjanen, P.R., 2012. Perspectives on the ARE as it turns 25 years old. *Wiley Interdiscip. Rev. RNA* 3, 719–31. doi:10.1002/wrna.1125
- Bembich, S., Herzog, J.S., De Conti, L., Stuani, C., Avendaño-Vázquez, S.E., Buratti, E., Baralle, M., Baralle, F.E., 2014. Predominance of spliceosomal complex formation over polyadenylation site selection in TDP-43 autoregulation. *Nucleic Acids Res.* 42, 3362–71. doi:10.1093/nar/gkt1343
- Bennett, C.L., Brunkow, M.E., Ramsdell, F., O'Briant, K.C., Zhu, Q., Fuleihan, R.L., Shigeoka, a O., Ochs, H.D., Chance, P.F., 2001. A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA-->AAUGAA) leads to the IPEX syndrome. *Immunogenetics* 53, 435–9. doi:10.1007/s002510100358
- Berg, M.G., Singh, L.N., Younis, I., Liu, Q., Pinto, A.M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L., Dreyfuss, G., 2012. U1 snRNP determines mRNA length and regulates isoform expression. *Cell* 150, 53–64. doi:10.1016/j.cell.2012.05.029
- Berkovits, B.D., Mayr, C., 2015. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature* 8. doi:10.1038/nature14321
- Bhattacharyya, S.N., Habermacher, R., Martine, U., Closs, E.I., Filipowicz, W., 2006. Relief of microRNA-Mediated

- Translational Repression in Human Cells Subjected to Stress. *Cell* 125, 1111–1124. doi:10.1016/j.cell.2006.04.031
- Biedler, J.L., Helson, L., Spengler, B. a, 1973. Morphology and growth, tumorigenicity, and cytogenetics of human neuroblastoma cells in continuous culture. *Cancer Res* 33, 2643–2652.
- Boutet, S.C., Cheung, T.H., Quach, N.L., Liu, L., Prescott, S.L., Edalati, A., Iori, K., Rando, T. a, 2012. Alternative polyadenylation mediates microRNA regulation of muscle stem cell function. *Cell Stem Cell* 10, 327–36. doi:10.1016/j.stem.2012.01.017
- Boutz, P.L., Bhutkar, A., Sharp, P.A., 2014. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev.* 29, 63–80. doi:10.1101/gad.247361.114.Braunschweig
- Brais, B., Bouchard, J.P., Xie, Y.G., Rochefort, D.L., Chrétien, N., Tomé, F.M., Lafrenière, R.G., Rommens, J.M., Uyama, E., Nohira, O., Blumen, S., Korczyn, a D., Heutink, P., Mathieu, J., Duranceau, a, Codère, F., Fardeau, M., Rouleau, G. a, 1998. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nat. Genet.* 18, 164–167. doi:10.1038/ng0298-164
- Braunschweig, U., Barbosa-Morais, N.L., Pan, Q., Nachman, E.N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M., Blencowe, B.J., 2014. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.* doi:10.1101/gr.177790.114
- Brennan, C.M., Steitz, J. a, 2001. HuR and mRNA stability. *Cell. Mol. Life Sci.* 58, 266–277. doi:10.1007/PL00000854
- Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafrenière, R.G., Xing, Y., Lawrence, J., Willard, H.F., 1992. The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* 71, 527–542. doi:10.1016/0092-8674(92)90520-M
- C.-H., W., F.-Y., C., J.-Y., C., J.-Y., W., 2013. Rapid detection of gene expression by a colorectal cancer Enzymatic Gene Chip Detection Kit. *Biomarkers Genomic Med.* 5, 87–91. doi:10.1016/j.gmbhs.2013.05.004
- Calin, G. a, Croce, C.M., 2006. MicroRNA signatures in human cancers. *Nat. Rev. Cancer* 6, 857–66. doi:10.1038/nrc1997
- Cardoso, J., Boer, J., Morreau, H., Fodde, R., 2007. Expression and genomic profiling of colorectal cancer. *Biochim. Biophys. Acta - Rev. Cancer* 1775, 103–137. doi:10.1016/j.bbcan.2006.08.004
- Carter, K.C., Taneja, K.L., Lawrence, J.B., 1991. Discrete nuclear domains of Poly (A) RNA and their relationship to the functional organizations of the nucleus. *Jcb* 115, 1191–1202.
- Castelo-Branco, P., Furger, A., Wollerton, M., Smith, C.W.J., Moreira, A., Proudfoot, N.J., 2004. Polypyrimidine tract binding protein modulates efficiency of polyadenylation. *Mol. Cell. Biol.* 24, 4174–4183. doi:10.1128/MCB.24.10.4174
- Chen, L.-L., DeCervo, J.N., Carmichael, G.G., 2008. Alu element-mediated gene silencing. *EMBO J.* 27, 1694–705. doi:10.1038/emboj.2008.94
- Chen, Z., Liu, S., Tian, L., Wu, M., Ai, F., Tang, W., Zhao, L., Ding, J., Zhang, L., Tang, A., 2015. miR-124 and miR-506 inhibit colorectal cancer progression by targeting DNMT3B and DNMT1. *Oncotarget* 6. doi:10.18632/oncotarget.5709
- Cheng, Y.J., Zhu, Z.X., Zhou, J.S., Hu, Z.Q., Zhang, J.P., Cai, Q.P., Wang, L.H., 2015. Silencing profilin-1 inhibits gastric cancer progression via integrin β 1/focal adhesion kinase pathway modulation. *World J. Gastroenterol.* 21, 2323–2335. doi:10.3748/wjg.v21.i8.2323
- Colmenares, S.U., Buker, S.M., Buhler, M., Dlakić, M., Moazed, D., 2007. Coupling of Double-Stranded RNA Synthesis

- and siRNA Generation in Fission Yeast RNAi. *Mol. Cell* 27, 449–461. doi:10.1016/j.molcel.2007.07.007
- Cuk, K., Zucknick, M., Heil, J., Madhavan, D., Schott, S., Turchinovich, A., Arlt, D., Rath, M., Sohn, C., Benner, A., Junkermann, H., Schneeweiss, A., Burwinkel, B., 2013. Circulating microRNAs in plasma as early detection markers for breast cancer. *Int. J. Cancer* 132, 1602–1612. doi:10.1002/ijc.27799
- Dai, W., Zhang, G., Makeyev, E. V., 2011. RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage. *Nucleic Acids Res.* 40, 787–800. doi:10.1093/nar/gkr783
- Dalziel, M., Nunes, N.M., Furger, A., 2007. Two G-rich regulatory elements located adjacent to and 440 nucleotides downstream of the core poly(A) site of the intronless melanocortin receptor 1 gene are critical for efficient 3' end processing. *Mol. Cell. Biol.* 27, 1568–80. doi:10.1128/MCB.01821-06
- Dass, B., McDaniel, L., Schultz, R. a., Attaya, E., MacDonald, C.C., 2002. The gene CSTF2T, encoding the human variant CstF-64 polyadenylation protein TCstF-64, lacks introns and may be associated with male sterility. *Genomics* 80, 509–514. doi:10.1016/S0888-7543(02)96862-X
- de Klerk, E., Venema, A., 2012. Poly (A) binding protein nuclear 1 levels affect alternative polyadenylation. *Nucleic Acids Res.* 40, 9089–9101. doi:10.1093/nar/gks655
- de la Mata, M., Alonso, C.R., Kadener, S., Fededa, J.P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D., Kornblihtt, A.R., 2003. A Slow RNA Polymerase II Affects Alternative Splicing In Vivo. *Mol. Cell* 12, 525–532. doi:10.1016/j.molcel.2003.08.001
- de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M., Vogel, C., 2009. Global signatures of protein and mRNA expression levels. *Mol. Biosyst.* 5, 1512–1526. doi:10.1039/b908315d
- Derti, A., Garrett-Engle, P., Macisaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C. a, Johnson, J.M., Babak, T., 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res.* 22, 1173–83. doi:10.1101/gr.132563.111
- Deschodt-Lanckman, M., Vanneste, Y., Loir, B., Michel, A., Libert, A., Ghanem, G., Lejeune, F., 1990. Degradation of alpha-melanocyte stimulating hormone (alpha-MSH) by CALLA/endopeptidase 24.11 expressed by human melanoma cells in culture. *Int. J. Cancer* 46, 1124–1130.
- Di Giammartino, D.C., Li, W., Ogami, K., Yashinskies, J.J., Hoque, M., Tian, B., 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. doi:10.1101/gad.245787.114
- Di Giammartino, D.C., Nishida, K., Manley, J.L., 2011. Mechanisms and consequences of alternative polyadenylation. *Mol. Cell* 43, 853–66. doi:10.1016/j.molcel.2011.08.017
- Dickson, J.R., Kruse, C., Montagna, D.R., Finsen, B., Wolfe, M.S., 2013. Alternative Polyadenylation and miR-34 Family Members Regulate Tau Expression. *J. Neurochem.* 1–11. doi:10.1111/jnc.12437
- Djebali, S., Davis, C. a, Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B. a, Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakraborty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E.,

- Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T.R., 2012. Landscape of transcription in human cells. *Nature* 489, 101–8. doi:10.1038/nature11233
- Doller, A., Schlepckow, K., Schwalbe, H., Pfeilschifter, J., Eberhardt, W., 2010. Tandem phosphorylation of serines 221 and 318 by protein kinase Cdelta coordinates mRNA binding and nucleocytoplasmic shuttling of HuR. *Mol. Cell. Biol.* 30, 1397–1410. doi:10.1128/MCB.01373-09
- Duc, C., Sherstnev, A., Cole, C., Barton, G.J., Simpson, G.G., 2013. Transcription Termination and Chimeric RNA Formation Controlled by Arabidopsis thaliana FPA. *PLoS Genet.* 9, e1003867. doi:10.1371/journal.pgen.1003867
- Dye, M.J., Proudfoot, N.J., 2001. Multiple transcript cleavage precedes polymerase release in termination by RNA polymerase II. *Cell* 105, 669–681. doi:10.1016/S0092-8674(01)00372-5
- Edwards-Gilbert, G., Veraldi, K.L., Milcarek, C., 1997. Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res.* 25, 2547–61.
- Elkon, R., Drost, J., van Haaften, G., Jenal, M., Schrier, M., Oude Vrielink, J.A., Agami, R., 2012. E2F mediates enhanced alternative polyadenylation in proliferation. *Genome Biol.* 13, R59. doi:10.1186/gb-2012-13-7-r59
- Elkon, R., Ugalde, A.P., Agami, R., 2013. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.* 14, 496–506. doi:10.1038/nrg3482
- Eulalio, A., Tritschler, F., Izaurralde, E., 2009. The GW182 protein family in animal cells: new insights into domains required for miRNA-mediated gene silencing. *RNA* 15, 1433–42. doi:10.1261/rna.1703809
- Evsyukova, I., Bradrick, S.S., Gregory, S.G., Garcia-Blanco, M. a, 2013. Cleavage and polyadenylation specificity factor 1 (CPSF1) regulates alternative splicing of interleukin 7 receptor (IL7R) exon 6. *RNA* 19, 103–15. doi:10.1261/rna.035410.112
- Fan, H., Lu, Y., Qin, H., Zhou, Y., Gu, Y., Zhou, J., Wang, X., Fan, D., 2013. High Ran level is correlated with poor prognosis in patients with colorectal cancer. *Int. J. Clin. Oncol.* 18, 856–863. doi:10.1007/s10147-012-0465-x
- Fan, X.C., Steitz, J. a., 1998. Overexpression of HuR, a nuclear-cytoplasmic shuttling protein, increases the in vivo stability of ARE-containing mRNAs. *EMBO J.* 17, 3448–3460. doi:10.1093/emboj/17.12.3448
- Flavell, S.W., Kim, T.-K., Gray, J.M., Harmin, D. a, Hemberg, M., Hong, E.J., Markenscoff-Papadimitriou, E., Bear, D.M., Greenberg, M.E., 2008. Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. *Neuron* 60, 1022–38. doi:10.1016/j.neuron.2008.11.029
- Fleming, N.I., Jorissen, R.N., Mouradov, D., Christie, M., Sakthianandeswaren, A., Palmieri, M., Day, F., Li, S., Tsui, C., Lipton, L., Desai, J., Jones, I.T., McLaughlin, S., Ward, R.L., Hawkins, N.J., Ruskiewicz, A.R., Moore, J., Zhu, H.J., Mariadason, J.M., Burgess, A.W., Busam, D., Zhao, Q., Strausberg, R.L., Gibbs, P., Sieber, O.M., 2013. SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. *Cancer Res.* 73, 725–735. doi:10.1158/0008-5472.CAN-12-2706
- Fox-Walsh, K., Davis-Turak, J., Zhou, Y., Li, H., Fu, X.-D., 2011. A multiplex RNA-seq strategy to profile poly(A+) RNA: application to analysis of transcription response and 3' end formation. *Genomics* 98, 266–71. doi:10.1016/j.ygeno.2011.04.003

- Friedman, R.C., Farh, K.K.-H., Burge, C.B., Bartel, D.P., 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 19, 92–105. doi:10.1101/gr.082701.108
- Fu, Y., Sun, Y., Li, Y., Li, J., Rao, X., Chen, C., Xu, A., 2011. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res.* 21, 741–7. doi:10.1101/gr.115295.110
- Furger, A., Monks, J., Proudfoot, N.J., 2001. The retroviruses human immunodeficiency virus type 1 and Moloney murine leukemia virus adopt radically different strategies to regulate promoter-proximal polyadenylation. *J. Virol.* 75, 11735–11746. doi:10.1128/JVI.75.23.11735
- Fusby, B., Kim, S., Erickson, B., Kim, H., Peterson, M.L., Bentley, D.L., 2015. Coordination of RNA Polymerase II Pausing and 3' end processing factor recruitment with alternative polyadenylation. *Mol. Cell. Biol.* 36, MCB.00898-15. doi:10.1128/MCB.00898-15
- Gaidatzis, D., van Nimwegen, E., Hausser, J., Zavolan, M., 2007. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics* 8, 69. doi:10.1186/1471-2105-8-69
- Gantier, M.P., McCoy, C.E., Rusinova, I., Saulep, D., Wang, D., Xu, D., Irving, A.T., Behlke, M. a, Hertzog, P.J., Mackay, F., Williams, B.R.G., 2011. Analysis of microRNA turnover in mammalian cells following Dicer1 ablation. *Nucleic Acids Res.* 39, 5692–703. doi:10.1093/nar/gkr148
- Gautheret, D., Poirot, O., Lopez, F., 1998. Alternate Polyadenylation in Human mRNAs : A Large-Scale Analysis by EST Clustering. *Genome Res.* 8, 524–530. doi:10.1101/gr.8.5.524
- Geisberg, J.V., Moqtaderi, Z., Fan, X., Oszolak, F., Struhl, K., 2014. Global Analysis of mRNA Isoform Half-Lives Reveals Stabilizing and Destabilizing Elements in Yeast. *Cell* 156, 812–824. doi:10.1016/j.cell.2013.12.026
- Gil, A., Proudfoot, N.J., 1987. Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit B-globin mRNA 3' end formation. *Cell* 49, 399–406. doi:10.1016/0092-8674(87)90292-3
- Ginno, P. a, Lim, Y.W., Lott, P.L., Korf, I., Chédin, F., 2013. GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res.* 23, 1590–1600. doi:10.1101/gr.158436.113
- Ginno, P. a, Lott, P.L., Christensen, H.C., Korf, I., Chédin, F., 2012. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol. Cell* 45, 814–25. doi:10.1016/j.molcel.2012.01.017
- Grimson, A., Farh, K.K.-H., Johnston, W.K., Garrett-Engle, P., Lim, L.P., Bartel, D.P., 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell* 27, 91–105. doi:10.1016/j.molcel.2007.06.017
- Gruber, A.R., Martin, G., Müller, P., Schmidt, A., Gruber, A.J., Gumienny, R., Mittal, N., Jayachandran, R., Pieters, J., Keller, W., van Nimwegen, E., Zavolan, M., 2014. Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nat. Commun.* 5, 5465. doi:10.1038/ncomms6465
- Gunderson, S.I., Beyer, K., Martin, G., Keller, W., Boelens, W.C., Mattaj, I.W., 1994. The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase. *Cell* 76, 531–541. doi:10.1016/0092-8674(94)90116-3
- Guo, H., Ingolia, N.T., Weissman, J.S., Bartel, D.P., 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466, 835–40. doi:10.1038/nature09267
- Gupta, I., Clauder-Münster, S., Klaus, B., Järvelin, A.I., Aiyar, R.S., Benes, V., Wilkening, S., Huber, W., Pelechano, V.,

- Steinmetz, L.M., 2014. Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions. *Mol. Syst. Biol.* 10, 719. doi:10.1002/msb.135068
- Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., Noble, W., 2007. Quantifying similarity between motifs. *Genome Biol.* 8, R24. doi:10.1186/gb-2007-8-2-r24
- Haenni, S., Ji, Z., Hoque, M., Rust, N., Sharpe, H., Eberhard, R., Browne, C., Hengartner, M.O., Mellor, J., Tian, B., Furger, A., 2012. Analysis of *C. elegans* intestinal gene expression and polyadenylation by fluorescence-activated nuclei sorting and 3'-end-seq. *Nucleic Acids Res.* 40, 6304–6318. doi:10.1093/nar/gks282
- Hafner, M., Renwick, N., Brown, M., Mihailović, A., Holoch, D., Lin, C., Pena, J.T.G., Nusbaum, J.D., Morozov, P., Ludwig, J., Ojo, T., Luo, S., Schroth, G., Tuschl, T., 2011. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* 17, 1697–712. doi:10.1261/rna.2799511
- Hargreaves, D.C., Horng, T., Medzhitov, R., 2009. Control of inducible gene expression by signal-dependent transcriptional elongation. *Cell* 138, 129–45. doi:10.1016/j.cell.2009.05.047
- Hazzalin, C. a, Cuenda, A., Cano, E., Cohen, P., Mahadevan, L.C., 1997. Effects of the inhibition of p38/RK MAP kinase on induction of five fos and jun genes by diverse stimuli. *Oncogene* 15, 2321–31. doi:10.1038/sj.onc.1201403
- Herman, R.C., Williams, J.G., Penman, S., 1976. Message and non-message sequences adjacent to poly(A) in steady state heterogeneous nuclear RNA of HeLa cells. *Cell* 7, 429–437. doi:10.1016/0092-8674(76)90173-2
- Hewitt, R.E., McMarlin, A., Kleiner, D., Wersto, R., Martin, P., Tsokos, M., Stamp, G.W., Stetler-Stevenson, W.G., 2000. Validation of a model of colon cancer progression. *J Pathol* 192, 446–454. doi:10.1002/1096-9896(2000)9999:9999<::AID-PATH775>3.0.CO;2-K
- Hinske, L.C., Galante, P. a. F., Limbeck, E., Möhnle, P., Parmigiani, R.B., Ohno-Machado, L., Camargo, A. a., Kreth, S., 2015. Alternative Polyadenylation Allows Differential Negative Feedback of Human miRNA miR-579 on Its Host Gene ZFR. *PLoS One* 10, e0121507. doi:10.1371/journal.pone.0121507
- Holden, P., Horton, W.A., 2009. Crude subcellular fractionation of cultured mammalian cell lines. *BMC Res. Notes* 2, 243. doi:10.1186/1756-0500-2-243
- Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J.Y., Yehia, G., Tian, B., 2013. Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat. Methods* 10, 133–9. doi:10.1038/nmeth.2288
- Hoque, M., Li, W., Tian, B., 2014. Accurate Mapping of Cleavage and Polyadenylation Sites by 3' Region Extraction and Deep Sequencing, in: *Polyadenylation*. pp. 119–129. doi:10.1007/978-1-62703-971-0_10
- Hsin, J., Manley, J.L., 2012. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev.* 26, 2119–2137. doi:10.1101/gad.200303.112.Transcription
- Hu, X., Schwarz, J.K., Lewis, J.S., Huettner, P.C., Rader, J.S., Deasy, J.O., Grigsby, P.W., Wang, X., 2010. A microRNA expression signature for cervical cancer prognosis. *Cancer Res.* 70, 1441–1448. doi:10.1158/0008-5472.CAN-09-3289
- Huang, H., Chen, J., Liu, H., Sun, X., 2013. The nucleosome regulates the usage of polyadenylation sites in the human genome. *BMC Genomics* 14, 912. doi:10.1186/1471-2164-14-912
- Hung, L.-H., Heiner, M., Hui, J., Schreiner, S., Benes, V., Bindereif, A., 2008. Diverse roles of hnRNP L in mammalian mRNA processing: a combined microarray and RNAi analysis. *RNA* 14, 284–296. doi:10.1261/rna.725208

- Huynh, C., Segura, M.F., Gaziel-Sovran, A., Menendez, S., Darvishian, F., Chiriboga, L., Levin, B., Meruelo, D., Osman, I., Zavadi, J., Marcussen, E.G., Hernando, E., 2011. Efficient in vivo microRNA targeting of liver metastasis. *Oncogene* 30, 1481–1488. doi:10.1038/onc.2010.523
- Hwang, H.-W., Park, C.Y., Goodarzi, H., Fak, J.J., Mele, A., Moore, M.J., Saito, Y., Darnell, R.B., 2016. PAPERCLIP Identifies MicroRNA Targets and a Role of CstF64/64tau in Promoting Non-canonical poly(A) Site Usage. *Cell Rep.* 15, 423–435. doi:10.1016/j.celrep.2016.03.023
- Jaffe, E.A., Nachman, R.L., Becker, C.G., Minick, C.R., 1973. Culture of human endothelial cells derived from umbilical veins. Identification by morphologic and immunologic criteria. *J. Clin. Invest.* 52, 2745–2756. doi:10.1172/JCI107470
- Jambhekar, A., Derisi, J.L., 2007. Cis -acting determinants of asymmetric , cytoplasmic RNA transport Cis -acting determinants of asymmetric , cytoplasmic RNA transport 625–642. doi:10.1261/rna.262607.transcripts
- Jan, C.H., Friedman, R.C., Ruby, J.G., Bartel, D.P., 2011. Formation, regulation and evolution of *Caenorhabditis elegans* 3'UTRs. *Nature*. doi:10.1038/nature09616
- Jenal, M., Elkon, R., Loayza-Puch, F., van Haaften, G., Kühn, U., Menzies, F.M., Oude Vrielink, J. a F., Bos, A.J., Drost, J., Rooijers, K., Rubinsztein, D.C., Agami, R., 2012. The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell* 149, 538–53. doi:10.1016/j.cell.2012.03.022
- Ji, L., Wei, Y., Jiang, T., Wang, S., 2014. Correlation of Nrf2, NQO1, MRP1, cmyc and p53 in colorectal cancer and their relationships to clinicopathologic features and survival. *Int. J. Clin. Exp. Pathol.* 7, 1124–1131.
- Ji, X., Wan, J., Vishnu, M., Xing, Y., Liebhaber, S. a, 2013. α CP Poly(C) binding proteins act as global regulators of alternative polyadenylation. *Mol. Cell. Biol.* 33, 2560–73. doi:10.1128/MCB.01380-12
- Ji, Z., Lee, J.Y., Pan, Z., Jiang, B., Tian, B., 2009. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci. U. S. A.* 106, 7028–33. doi:10.1073/pnas.0900028106
- Ji, Z., Tian, B., 2009. Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PLoS One* 4, e8419. doi:10.1371/journal.pone.0008419
- Johnstone, R., Wang, J., Tommerup, N., Vissing, H., Roberts, T., Shi, Y., 1998. Cio 1 Is a Novel WD40 Protein That Interacts with the Tumor Suppressor Protein WT1. *J. Biol. ...* 273, 10880–10887.
- Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., Dreyfuss, G., 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* 468, 664–8. doi:10.1038/nature09479
- Kallehauge, T.B., Robert, M.-C., Bertrand, E., Jensen, T.H., 2012. Nuclear retention prevents premature cytoplasmic appearance of mRNA. *Mol. Cell* 48, 145–52. doi:10.1016/j.molcel.2012.07.022
- Katz, Y., Wang, E.T., Airolidi, E.M., Burge, C.B., 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* 7, 1009–15. doi:10.1038/nmeth.1528
- Kaufmann, I., Martin, G., Friedlein, A., Langen, H., 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, 616–26. doi:10.1038/sj.emboj.7600070
- Kim, V.N., Han, J., Siomi, M.C., 2009. Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.* 10, 126–39. doi:10.1038/nrm2632
- Knowles BB, Howe CC, A.D., 1980. Human hepatocellular carcinoma cell lines secrete the major plasma proteins and

- hepatitis B surface antigen. *Science* (80-.). 209, 497–499.
- Kreth, S., Limbeck, E., Hinske, L.C., Schütz, S. V., Thon, N., Hoefig, K., Egensperger, R., Kreth, F.W., 2013. In human glioblastomas transcript elongation by alternative polyadenylation and miRNA targeting is a potent mechanism of MGMT silencing. *Acta Neuropathol.* 125, 671–81. doi:10.1007/s00401-013-1081-1
- Krol, J., Loedige, I., Filipowicz, W., 2010. The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.* 11, 597–610. doi:10.1038/nrg2843
- Kubicek, S., O’Sullivan, R.J., August, E.M., Hickey, E.R., Zhang, Q., Teodoro, M.L., Rea, S., Mechtler, K., Kowalski, J. a., Homon, C.A., Kelly, T. a., Jenuwein, T., 2007. Reversal of H3K9me2 by a Small-Molecule Inhibitor for the G9a Histone Methyltransferase. *Mol. Cell* 25, 473–481. doi:10.1016/j.molcel.2007.01.017
- Kubo, T., Wada, T., Yamaguchi, Y., Shimizu, A., Handa, H., 2006. Knock-down of 25 kDa subunit of cleavage factor Im in Hela cells alters alternative polyadenylation within 3’-UTRs. *Nucleic Acids Res.* 34, 6264–6271. doi:10.1093/nar/gkl794
- Kyburz, A., Friedlein, A., Langen, H., 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. *Mol. Cell* 23, 195–205. doi:10.1016/j.molcel.2006.05.037
- Lackford, B., Yao, C., Charles, G.M., Weng, L., Zheng, X., Choi, E., Xie, X., Wan, J., Xing, Y., Freudenberg, J.M., Yang, P., Jothi, R., 2014. Fip 1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *EMBO J.* 1–12. doi:10.1002/embj.201386537
- Lafontaine, D.L.J., 2015. Noncoding RNAs in eukaryotic ribosome biogenesis and function. *Nat. Struct. & Mol. Biol.* 22, 11–19. doi:10.1038/nsmb.2939
- Lazarov, M.E., Martin, M.M., Willardson, B.M., Elton, T.S., 1999. Human phosducin-like protein (hPhLP) messenger RNA stability is regulated by cis-acting instability elements present in the 3’-untranslated region. *Biochim. Biophys. Acta* 1446, 253–64.
- Lee, D.-H., Lim, M.-H., Youn, D.-Y., Jung, S.E., Ahn, Y.S., Tsujimoto, Y., Lee, J.-H., 2009. hnRNP L binds to CA repeats in the 3’UTR of bcl-2 mRNA. *Biochem. Biophys. Res. Commun.* 382, 583–7. doi:10.1016/j.bbrc.2009.03.069
- Lee, E.K., Gorospe, M., 2011. Coding region: The neglected post-transcriptional code. *RNA Biol.* 8, 44–48. doi:10.4161/rna.8.1.13863
- Legendre, M., Gautheret, D., 2003. Sequence determinants in human polyadenylation site selection. *BMC Genomics* 4, 7.
- Legendre, M., Ritchie, W., Lopez, F., Gautheret, D., 2006. Differential repression of alternative transcripts: a screen for miRNA targets. *PLoS Comput. Biol.* 2, e43. doi:10.1371/journal.pcbi.0020043
- Lei, R., Tang, J., Zhuang, X., Deng, R., Li, G., Yu, J., Liang, Y., Xiao, J., Wang, H.-Y., Yang, Q., Hu, G., 2014. Suppression of MIM by microRNA-182 activates RhoA and promotes breast cancer metastasis. *Oncogene* 33, 1287–96. doi:10.1038/onc.2013.65
- Leibovitz, A., Stinson, J.C., McCombs III, W.B., McCoy, C.E., Mazur, K.C., Mabry, N.D., 1976. Classification of human colorectal adenocarcinoma cell lines. *Cancer Res.* 36, 4562–4569.
- Lejeune, F., Maquat, L.E., 2005. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Curr. Opin. Cell Biol.* 17, 309–315. doi:10.1016/j.ceb.2005.03.002
- Lembo, A., Di Cunto, F., Provero, P., 2012. Shortening of 3’UTRs correlates with poor prognosis in breast and lung cancer.

- Letzen, B.S., Liu, C., Thakor, N. V., Gearhart, J.D., All, A.H., Kerr, C.L., 2010. MicroRNA Expression Profiling of Oligodendrocyte Differentiation from Human Embryonic Stem Cells. *PLoS One* 5, e10480. doi:10.1371/journal.pone.0010480
- Li, W., Park, J.Y., Zheng, D., Hoque, M., Yehia, G., Tian, B., 2016. Alternative cleavage and polyadenylation in spermatogenesis connects chromatin regulation with post-transcriptional control. *BMC Biol.* 14, 6. doi:10.1186/s12915-016-0229-6
- Li, W., You, B., Hoque, M., Zheng, D., Luo, W., Ji, Z., Park, J.Y., Gunderson, S.I., Kalsotra, A., Manley, J.L., Tian, B., 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 Genet.* 11, e1005166. doi:10.1371/journal.pgen.1005166
- Liang, Y., Ridzon, D., Wong, L., Chen, C., 2007. Characterization of microRNA expression profiles in normal human tissues. *BMC Genomics* 8, 166. doi:10.1186/1471-2164-8-166
- Lianoglou, S., Garg, V., Yang, J.L., Leslie, C.S., Mayr, C., 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* 0–17. doi:10.1101/gad.229328.113
- Liaw, H.-H., Lin, C.-C., Juan, H.-F., Huang, H.-C., 2013. Differential MicroRNA Regulation Correlates with Alternative Polyadenylation Pattern between Breast Cancer and Normal Cells. *PLoS One* 8, e56958. doi:10.1371/journal.pone.0056958
- Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T. a, Schweitzer, A.C., Blume, J.E., Wang, X., Darnell, J.C., Darnell, R.B., 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456, 464–9. doi:10.1038/nature07488
- Lin, Y., Li, Z., Oszolak, F., 2012. An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res.* 40, 8460–8471. doi:10.1093/nar/gks637
- Liu, D., Brockman, J.M., Dass, B., Hutchins, L.N., Singh, P., McCarrey, J.R., MacDonald, C.C., Graber, J.H., 2007. Systematic variation in mRNA 3'-processing signals during mouse spermatogenesis. *Nucleic Acids Res.* 35, 234–246. doi:10.1093/nar/gkl919
- Liu, Y., Hu, W., Murakawa, Y., Yin, J., Wang, G., Landthaler, M., Yan, J., 2013. Cold-induced RNA-binding proteins regulate circadian gene expression by controlling alternative polyadenylation. *Sci. Rep.* 3, 1–11. doi:10.1038/srep02054
- Lou, H., Neugebauer, K., 1998. Regulation of Alternative Polyadenylation by U1 snRNPs and SRp20. *Mol. Cell. Biol.* 18, 4977.
- Lozzio, C.B., Lozzio, B.B., 1975. Human chronic myelogenous leukemia cell line with positive Philadelphia chromosome. *Blood* 45, 321–334.
- Luo, W., Ji, Z., Pan, Z., You, B., Hoque, M., Li, W., Gunderson, S.I., Tian, B., 2013. The Conserved Intronic Cleavage and Polyadenylation Site of CstF-77 Gene Imparts Control of 3' End Processing Activity through Feedback Autoregulation and by U1 snRNP. *PLoS Genet.* 9, e1003613. doi:10.1371/journal.pgen.1003613
- Lutz, C.S., Moreira, A., 2010. Alternative mRNA polyadenylation in eukaryotes: an effective regulator of gene expression. *Wiley Interdiscip. Rev. RNA* 2, 22–31. doi:10.1002/wrna.47

- MacDonald, C.C., McMahon, K.W., 2010. Tissue-specific mechanisms of alternative polyadenylation: testis, brain, and beyond. *Wiley Interdiscip. Rev. RNA* 1, 494–501. doi:10.1002/wrna.29
- Majoros, W.H., Ohler, U., 2007. Spatial preferences of microRNA targets in 3' untranslated regions. *BMC Genomics* 8, 152. doi:10.1186/1471-2164-8-152
- Mandel, C.R., Bai, Y., Tong, L., 2008. Protein factors in pre-mRNA 3'-end processing. *Cell. Mol. Life Sci.* 65, 1099–122. doi:10.1007/s00018-007-7474-3
- Mangone, M., Macmenamin, P., Zegar, C., Piano, F., Gunsalus, K.C., 2008. UTRome.org: a platform for 3'UTR biology in *C. elegans*. *Nucleic Acids Res.* 36, D57–62. doi:10.1093/nar/gkm946
- Maniatis, T., Reed, R., 2002. Expression Machines. *Nature* 416, 499–506. doi:10.1038/416499a
- Mao, Y.S., Zhang, B., Spector, D.L., 2011. Biogenesis and function of nuclear bodies. *Trends Genet.* 27, 295–306. doi:10.1016/j.tig.2011.05.006
- Mapendano, C.K., Lykke-Andersen, S., Kjems, J., Bertrand, E., Jensen, T.H., 2010. Crosstalk between mRNA 3' End Processing and Transcription Initiation. *Mol. Cell* 40, 410–422. doi:10.1016/j.molcel.2010.10.012
- Martin, G., Gruber, A.R., Keller, W., 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 Rep.* 1, 753–763. doi:10.1016/j.celrep.2012.05.003
- Martinson, H.G., 2011. An active role for splicing in 3'-end formation. *Wiley Interdiscip. Rev. RNA* 2, 459–470. doi:10.1002/wrna.68
- Masamha, C.P., Xia, Z., Yang, J., Albrecht, T.R., Li, M., Shyu, A.-B., Li, W., Wagner, E.J., 2014. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* 509, 412–416. doi:10.1038/nature13261
- Masuda, A., Takeda, J., Okuno, T., Okamoto, T., Ohkawara, B., Ito, M., Ishigaki, S., Sobue, G., Ohno, K., 2015. Position-specific binding of FUS to nascent RNA regulates mRNA length. *Genes Dev.* 1045–1057. doi:10.1101/gad.255737.114.itation
- Mayr, C., Bartel, D.P., 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* 138, 673–84. doi:10.1016/j.cell.2009.06.016
- Mccracken, S., Lambermon, M., Benjamin, J., Blencowe, B.J., 2002. SRm160 Splicing Coactivator Promotes Transcript 3' - End Cleavage. *Mol. Cell. Biol.* 22, 148–160. doi:10.1128/MCB.22.1.148
- Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T., Thomas, P.D., 2016. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res* 44, D336–42. doi:10.1093/nar/gkv1194
- Millevoi, S., Decorsière, A., Loulergue, C., Iacovoni, J., Bernat, S., Antoniou, M., Vagner, S., 2009. A physical and functional link between splicing factors promotes pre-mRNA 3' end processing. *Nucleic Acids Res.* 37, 4672–83. doi:10.1093/nar/gkp470
- Millevoi, S., Loulergue, C., Dettwiler, S., Karaa, S.Z., Keller, W., Antoniou, M., Vagner, S., 2006. An interaction between U2AF 65 and CF I(m) links the splicing and 3' end processing machineries. *EMBO J.* 25, 4854–4864. doi:10.1038/sj.emboj.7601331
- Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J.O., Lai, E.C., 2013. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Res.* 23, 812–25. doi:10.1101/gr.146886.112
- Moll, P., Ante, M., Seitz, A., Reda, T., 2014. QuantSeq 3' mRNA sequencing for RNA quantification. *Nat. Methods* 11, 25.

doi:10.1038/nmeth.f.376

- Moore, M.J., 2005. From birth to death: the complex lives of eukaryotic mRNAs. *Science* 309, 1514–8. doi:10.1126/science.1111443
- Moqtaderi, Z., Geisberg, J. V., Jin, Y., Fan, X., Struhl, K., 2013. Species-specific factors mediate extensive heterogeneity of mRNA 3' ends in yeasts. *Proc. Natl. Acad. Sci. U. S. A.* 110, 11073–8. doi:10.1073/pnas.1309384110
- Morris, A.R., Bos, A., Diosdado, B., Rooijers, K., Elkon, R., Bolijn, A.S., Carvalho, B., Meijer, G. a, Agami, R., 2012. Alternative cleavage and polyadenylation during colorectal cancer development. *Clin. Cancer Res.* 18, 5256–66. doi:10.1158/1078-0432.CCR-12-0543
- Mouradov, D., Sloggett, C., Jorissen, R.N., Love, C.G., Li, S., Burgess, A.W., Arango, D., Strausberg, R.L., Buchanan, D., Wormald, S., O'Connor, L., Wilding, J.L., Bicknell, D., Tomlinson, I.P.M., Bodmer, W.F., Mariadason, J.M., Sieber, O.M., 2014. Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. *Cancer Res.* 74, 3238–3247. doi:10.1158/0008-5472.CAN-14-0013
- Murthy, K.G., Manley, J.L., 1995. The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation. *Genes Dev.* 9, 2672–2683. doi:10.1101/gad.9.21.2672
- Nagaike, T., Logan, C., Hotta, I., Rozenblatt-Rosen, O., Meyerson, M., Manley, J.L., 2011. Transcriptional activators enhance polyadenylation of mRNA precursors. *Mol. Cell* 41, 409–18. doi:10.1016/j.molcel.2011.01.022
- Naganuma, T., Nakagawa, S., Tanigawa, A., Sasaki, Y.F., Goshima, N., Hirose, T., 2012. Alternative 3'-end processing of long noncoding RNA initiates construction of nuclear paraspeckles. *EMBO J.* 31, 4020–34. doi:10.1038/emboj.2012.251
- Nam, J.-W., Rissland, O.S., Koppstein, D., Abreu-Goodger, C., Jan, C.H., Agarwal, V., Yildirim, M. a, Rodriguez, A., Bartel, D.P., 2014. Global Analyses of the Effect of Different Cellular Contexts on MicroRNA Targeting. *Mol. Cell* 53, 1031–1043. doi:10.1016/j.molcel.2014.02.013
- Nesterova, T.B., Popova, B.C., Cobb, B.S., Norton, S., Senner, C.E., Tang, Y.A., Spruce, T., Rodriguez, T. a, Sado, T., Merckenschlager, M., Brockdorff, N., 2008. Dicer regulates Xist promoter methylation in ES cells indirectly through transcriptional control of Dnmt3a. *Epigenetics Chromatin* 1, 2. doi:10.1186/1756-8935-1-2
- Neve, J., Burger, K., Li, W., Hoque, M., Patel, R., Tian, B., Gullerova, M., Furger, A., 2016. Subcellular RNA profiling links splicing and nuclear DICER1 to alternative cleavage and polyadenylation. *Genome Res.* 26, 24–35. doi:10.1101/gr.193995.115
- Neve, J., Furger, A., 2014. Alternative polyadenylation: less than meets the eye? *Biochem. Soc. Trans.* 42, 1190–5. doi:10.1042/BST20140054
- Nichols, W.W., Murphy, D.G., Cristofalo, V.J., Toji, L.H., Greene, A.E., Dwight, S.A., 1977. Characterization of a new human diploid cell strain, IMR-90. *Science* 196, 60–3. doi:10.1126/science.841339
- Niwa, M., MacDonald, C.C., Berget, S.M., 1992. Are vertebrate exons scanned during splice-site selection? *Nature* 360, 277–280. doi:10.1038/360277a0
- Nunes, N.M., Li, W., Tian, B., Furger, A., 2010. A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence. *EMBO J.* 29, 1523–1536. doi:10.1038/emboj.2010.42
- Oktaba, K., Zhang, W., Lotz, T.S., Jun, D.J., Lemke, S.B., Ng, S.P., Esposito, E., Levine, M., Hilgers, V., 2014. ELAV Links Paused Pol II to Alternative Polyadenylation in the Drosophila Nervous System. *Mol. Cell* 1–8.

doi:10.1016/j.molcel.2014.11.024

- Orkin, S.H., Cheng, T.C., Antonarakis, S.E., Kazazian, H.H., 1985. Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene. *EMBO J.* 4, 453–456.
- Ozsolak, F., Kapranov, P., Foissac, S., Kim, S.W., Fishilevich, E., Monaghan, A.P., John, B., Milos, P.M., 2010. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* 143, 1018–29. doi:10.1016/j.cell.2010.11.020
- Ozsolak, F., Milos, P.M., 2011. Transcriptome Profiling Using Single-Molecule Direct RNA Sequencing, in: *High-Throughput Next Generation Sequencing: Methods and Applications*. pp. 195–205. doi:10.1007/978-1-61779-089-8
- Pandya-Jones, A., 2011. Pre-mRNA splicing during transcription in the mammalian system. *Wiley Interdiscip. Rev.* 2, 700–717. doi:http://dx.doi.org/10.1002/wrna.86
- Pelechano, V., Wei, W., Steinmetz, L.M., 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* 497, 127–31. doi:10.1038/nature12121
- Pérez Cañadillas, J.M., Varani, G., 2003. Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. *EMBO J.* 22, 2821–30. doi:10.1093/emboj/cdg259
- Peterson, M.L., Bryman, M.B., Peiter, M., Cowan, C., 1994. Exon size affects competition between splicing and cleavage-polyadenylation in the immunoglobulin mu gene. *Mol. Cell. Biol.* 14, 77–86. doi:10.1128/MCB.14.1.77.Updated
- Peterson, M.L., Perry, R.P., 1989. The regulated production of mu m and mu s mRNA is dependent on the relative efficiencies of mu s poly(A) site usage and the c mu 4-to-M1 splice. *Mol. Cell. Biol.* 9, 726–738.
- Phillips, C., Pachikara, N., Gunderson, S.I., 2004. U1A inhibits cleavage at the immunoglobulin M heavy-chain secretory poly(A) site by binding between the two downstream GU-rich regions. *Mol. Cell. Biol.* 24, 6162–6171. doi:10.1128/MCB.24.14.6162-6171.2004
- Piepoli, A., Tavano, F., Copetti, M., Mazza, T., Palumbo, O., Panza, A., Di Mola, F.F., Pazienza, V., Mazzocchi, G., Biscaglia, G., Gentile, A., Mastrodonato, N., Carella, M., Pellegrini, F., Di Sebastiano, P., Andriulli, A., 2012. Mirna expression profiles identify drivers in colorectal and pancreatic cancers. *PLoS One* 7. doi:10.1371/journal.pone.0033663
- Pinto, P. a B., Henriques, T., Freitas, M.O., Martins, T., Domingues, R.G., Wyrzykowska, P.S., Coelho, P. a, Carmo, A.M., Sunkel, C.E., Proudfoot, N.J., Moreira, A., 2011. RNA polymerase II kinetics in polo polyadenylation signal selection. *EMBO J.* 30, 2431–44. doi:10.1038/emboj.2011.156
- Proudfoot, N.J., 2011. Ending the message: poly(A) signals then and now. *Genes Dev.* 25, 1770–82. doi:10.1101/gad.17268411
- Proudfoot, N.J., Furger, A., Dye, M.J., 2002. Integrating mRNA processing with transcription. *Cell* 108, 501–12.
- Puck, T.T., Marcus, P.I., Cieciura, S.J., 1956. Clonal growth of mammalian cells in vitro; growth characteristics of colonies from single HeLa cells with and without a feeder layer. *J. Exp. Med.* 103, 273–283. doi:10.1084/jem.103.2.273
- Rabani, M., Levin, J.Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., Amit, I., Regev, A., 2011. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* 29, 436–442. doi:10.1038/nbt.1861
- Rattenbacher, B., Beisang, D., Wiesner, D.L., Jeschke, J.C., von Hohenberg, M., St Louis-Vlasova, I. a, Bohjanen, P.R., 2010. Analysis of CUGBP1 targets identifies GU-repeat sequences that mediate rapid mRNA decay. *Mol. Cell. Biol.*

30, 3970–3980. doi:10.1128/MCB.00624-10

- Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L.H., Dale, R.K., Smith, S. a, Yarosh, C. a, Kelly, S.M., Nabet, B., Mecnas, D., Li, W., Laishram, R.S., Qiao, M., Lipshitz, H.D., Piano, F., Corbett, A.H., Carstens, R.P., Frey, B.J., Anderson, R. a, Lynch, K.W., Penalva, L.O.F., Lei, E.P., Fraser, A.G., Blencowe, B.J., Morris, Q.D., Hughes, T.R., 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499, 172–7. doi:10.1038/nature12311
- Rehfeld, A., Plass, M., Døssing, K., Knigge, U., Kjær, A., Krogh, A., Friis-Hansen, L., 2014. Alternative polyadenylation of tumor suppressor genes in small intestinal neuroendocrine tumors. *Front. Endocrinol. (Lausanne)*. 5, 46. doi:10.3389/fendo.2014.00046
- Rehfeld, A., Plass, M., Krogh, A., Friis-Hansen, L., 2013. Alterations in polyadenylation and its implications for endocrine disease. *Front. Endocrinol. (Lausanne)*. 4, 53. doi:10.3389/fendo.2013.00053
- Rigo, F., Kazerounia, A., Nag, A., Martinson, H.G., 2005. The RNA tether from the poly(A) signal to the polymerase mediates coupling of transcription to cleavage and polyadenylation. *Mol. Cell* 20, 733–45. doi:10.1016/j.molcel.2005.09.026
- Roig, A.I., Eskiocak, U., Hight, S.K., Kim, S.B., Delgado, O., Souza, R.F., Spechler, S.J., Wright, W.E., Shay, J.W., 2010. Immortalized epithelial cells derived from human colon biopsies express stem cell markers and differentiate in vitro. *Gastroenterology* 138, 1012–21–5. doi:10.1053/j.gastro.2009.11.052
- Roy, P.G., Pratt, N., Purdie, C. a, Baker, L., Ashfield, A., Quinlan, P., Thompson, A.M., 2010. High CCND1 amplification identifies a group of poor prognosis women with estrogen receptor positive breast cancer. *Int. J. Cancer* 127, 355–60. doi:10.1002/ijc.25034
- Sampath, P., Mazumder, B., Seshadri, V., Gerber, C. a, Chavatte, L., Kinter, M., Ting, S.M., Dignam, J.D., Kim, S., Driscoll, D.M., Fox, P.L., 2004. Noncanonical function of glutamyl-prolyl-tRNA synthetase: Gene-specific silencing of translation. *Cell* 119, 195–208. doi:10.1016/j.cell.2004.09.030
- Sánchez-Laorden, B.L., Herraiz, C., Valencia, J.C., Hearing, V.J., Jiménez-Cervantes, C., García-Borrón, J.C., 2009. Aberrant trafficking of human melanocortin 1 receptor variants associated with red hair and skin cancer: Steady-state retention of mutant forms in the proximal golgi. *J. Cell. Physiol.* 220, 640–54. doi:10.1002/jcp.21804
- Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P. a, Burge, C.B., 2008. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* 320, 1643–7. doi:10.1126/science.1155390
- Sawicka, K., Bushell, M., Spriggs, K.A., Willis, A.E., 2008. Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem Soc Trans* 36, 641–647.
- Schmidt, M.J., Norbury, C.J., 2010. Polyadenylation and beyond: Emerging roles for noncanonical poly(A) polymerases. *Wiley Interdiscip. Rev. RNA* 1, 142–151. doi:10.1002/wrna.16
- Schmitter, D., Filkowski, J., Sewer, A., Pillai, R.S., Oakeley, E.J., Zavolan, M., Svoboda, P., Filipowicz, W., 2006. Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells. *Nucleic Acids Res.* 34, 4801–15. doi:10.1093/nar/gkl646
- Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., Selbach, M., 2011. Global quantification of mammalian gene expression control. *Nature* 473, 337–42. doi:10.1038/nature10098
- Scotto-Lavino, E., Du, G., Frohman, M. a, 2006. 3' end cDNA amplification using classic RACE. *Nat. Protoc.* 1, 2742–5.

doi:10.1038/nprot.2006.481

- Shankarling, G.S., Macdonald, C.C., 2013. Polyadenylation site-specific differences in the activity of the neuronal β CstF-64 protein in PC-12 cells. *Gene* 64, 6–13. doi:10.1016/j.gene.2013.08.007
- Sharon, D., Tilgner, H., Grubert, F., Snyder, M., 2013. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31, 1009–14. doi:10.1038/nbt.2705
- Sheets, M.D., Ogg, S.C., Wickens, M.P., 1990. Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res.* 18, 5799–805.
- Shell, S. a, Hesse, C., Morris, S.M., Milcarek, C., 2005. Elevated levels of the 64-kDa cleavage stimulatory factor (CstF-64) in lipopolysaccharide-stimulated macrophages influence gene expression and induce alternative poly(A) site selection. *J. Biol. Chem.* 280, 39950–61. doi:10.1074/jbc.M508848200
- Shepard, P.J., Choi, E.-A., Lu, J., Flanagan, L. a, Hertel, K.J., Shi, Y., 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* 17, 761–72. doi:10.1261/rna.2581711
- Sherr, C.J., 1993. Mammalian G1 cyclins. *Cell* 73, 1059–1065. doi:10.1016/0955-0674(94)90054-X
- Sherstnev, A., Duc, C., Cole, C., Zacharaki, V., Hornyik, C., Ozsolak, F., Milos, P.M., Barton, G.J., Simpson, G.G., 2012. Direct sequencing of *Arabidopsis thaliana* RNA reveals patterns of cleavage and polyadenylation. *Nat. Struct. Mol. Biol.* 19, 845–852. doi:10.1038/nsmb.2345
- Shi, Y., 2012. Alternative polyadenylation: New insights from global analyses. *RNA* 18, 2105–2117. doi:10.1261/rna.035899.112
- Siegel, T.N., Hekstra, D.R., Wang, X., Dewell, S., Cross, G. a M., 2010. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. *Nucleic Acids Res.* 38, 4946–4957. doi:10.1093/nar/gkq237
- Simonelig, M., 2012. PABPN1 shuts down alternative poly(A) sites. *Cell Res.* 22, 1419–21. doi:10.1038/cr.2012.86
- Singh, P., Alley, T.L., Wright, S.M., Kamdar, S., Schott, W., Wilpan, R.Y., Mills, K.D., Graber, J.H., 2009. Global changes in processing of mRNA 3' untranslated regions characterize clinically distinct cancer subtypes. *Cancer Res.* 69, 9422–30. doi:10.1158/0008-5472.CAN-09-2236
- Skourti-Stathaki, K., Kamieniarz-gdula, K., Proudfoot, N.J., 2014. R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature* 516, 436–439. doi:10.1038/nature13787
- Slobodin, B., Agami, R., 2015. Transcription Initiation Determines Its End. *Mol. Cell* 57, 205–206. doi:10.1016/j.molcel.2015.01.006
- Smibert, P., Miura, P., Westholm, J.O., Shenker, S., May, G., Duff, M.O., Zhang, D., Eads, B.D., Carlson, J., Brown, J.B., Eisman, R.C., Andrews, J., Kaufman, T., Cherbas, P., Celniker, S.E., Graveley, B.R., Lai, E.C., 2012. Global patterns of tissue-specific alternative polyadenylation in *Drosophila*. *Cell Rep.* 1, 277–89. doi:10.1016/j.celrep.2012.01.001
- Soule, H.D., Vazquez, J., Long, A., Albert, S., Brennan, M., 1973. A Human Cell Line From a Pleural Effusion Derived From a Breast Carcinoma. *J Natl Cancer Inst* 51, 1409–1416. doi:10.1093/jnci/51.5.1409
- Spangenberg, L., Shigunov, P., Abud, A.P.R., Cofré, A.R., Stimamiglio, M. a, Kuligovski, C., Zych, J., Schittini, A. V, Costa, A.D.T., Rebelatto, C.K., Brofman, P.R.S., Goldenberg, S., Correa, A., Naya, H., Dallagiovanna, B., 2013. Polysome profiling shows extensive posttranscriptional regulation during human adipocyte stem cell differentiation into adipocytes. *Stem Cell Res.* 11, 902–912. doi:10.1016/j.scr.2013.06.002

- Spies, N., Burge, C., Bartel, D., 2013. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.* 23, 2078–2090. doi:10.1101/gr.156919.113
- Spies, N., Nielsen, C.B., Padgett, R. a, Burge, C.B., 2009. Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell* 36, 245–54. doi:10.1016/j.molcel.2009.10.008
- Sterne-Weiler, T., Martinez-Nunez, R., 2013. Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res.* 23, 1615–1623. doi:10.1101/gr.148585.112
- Sun, Y., Zhao, X., Zhou, Y., Hu, Y., 2012. MiR-124, miR-137 and miR-340 regulate colorectal cancer growth via inhibition of the Warburg effect. *Oncol. Rep.* 28, 1346–1352. doi:10.3892/or.2012.1958
- Tabas-Madrid, D., Muniategui, A., Sánchez-Caballero, I., Martínez-Herrera, D.J., S Sorzano, C.O., Rubio, A., Pascual-Montano, A., 2014. Improving miRNA-mRNA interaction predictions. *BMC Genomics* 15, S2. doi:10.1186/1471-2164-15-S10-S2
- Takagaki, Y., Seipelt, R.L., Peterson, M.L., Manley, J.L., 1996. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell* 87, 941–52.
- Takenaga, K., Nakanishi, H., Wada, K., Suzuki, M., Matsuzaki, O., Matsuura, A., Endo, H., 1997. Increased expression of S100A4, a metastasis-associated gene, in human colorectal adenocarcinomas. *Clin. Cancer Res.* 3, 2309–2316.
- Taliaferro, J.M., Vidaki, M., Oliveira, R., Gertler, F.B., Swanson, M.S., Burge, C.B., Taliaferro, J.M., Vidaki, M., Oliveira, R., Olson, S., Zhan, L., Saxena, T., Wang, E.T., 2016. Distal Alternative Last Exons Localize mRNAs to Neural Projections. *Mol. Cell* 1–13. doi:10.1016/j.molcel.2016.01.020
- Tan, J.Y., Sirey, T., Honti, F., Graham, B., Piovesan, A., Merkschlager, M., Webber, C., Ponting, C.P., Marques, A.C., 2015. Extensive microRNA-mediated crosstalk between lncRNAs and mRNAs in mouse embryonic stem cells. *Genome Res.* 1–12. doi:10.1101/gr.181974.114.
- Tian, B., Graber, J.H., 2012. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip. Rev. RNA* 3, 385–96. doi:10.1002/wrna.116
- Tian, B., Hu, J., Zhang, H., Lutz, C.S., 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* 33, 201–12. doi:10.1093/nar/gki158
- Tian, B., Pan, Z., Lee, J.Y., 2007. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res.* 17, 156–65. doi:10.1101/gr.5532707
- Tian, P., Li, J., Liu, X., Li, Y., Chen, M., Ma, Y., Zheng, Y.Q., Fu, Y., Zou, H., 2014. Tandem alternative polyadenylation events of genes in non-eosinophilic nasal polyp tissue identified by high-throughput sequencing analysis. *Int. J. Mol. Med.* 33, 1423–30. doi:10.3892/ijmm.2014.1734
- Tranter, M., Helsley, R.N., Paulding, W.R., McGuinness, M., Brokamp, C., Haar, L., Liu, Y., Ren, X., Jones, W.K., 2011. Coordinated post-transcriptional regulation of Hsp70.3 gene expression by microRNA and alternative polyadenylation. *J. Biol. Chem.* 286, 29828–37. doi:10.1074/jbc.M111.221796
- Ulitsky, I., Shkumatava, A., Jan, C.H., Subtelny, A.O., Koppstein, D., Bell, G.W., Sive, H., Bartel, D.P., 2012. Extensive alternative polyadenylation during zebrafish development. *Genome Res.* 22, 2054–2066. doi:10.1101/gr.139733.112
- Uzozie, A., Nanni, P., Staiano, T., Grossmann, J., Barkow-Oesterreicher, S., Shay, J.W., Tiwari, A., Buffoli, F., Laczko, E., Marra, G., 2014. Sorbitol dehydrogenase overexpression and other aspects of dysregulated protein expression in human precancerous colorectal neoplasms: a quantitative proteomics study. *Mol. Cell. Proteomics* 13, 1198–218.

doi:10.1074/mcp.M113.035105

- Van Nguyen, T., Giannoni, F., Dubois, M.F., Seo, S.J., Vigneron, M., Kédinger, C., Bensaude, O., 1996. In vivo degradation of RNA polymerase II largest subunit triggered by α -amanitin. *Nucleic Acids Res.* 24, 2924–2929. doi:10.1093/nar/24.15.2924
- Velten, L., Anders, S., Pekowska, A., Järvelin, A.I., Huber, W., Pelechano, V., Steinmetz, L.M., 2015. Single-cell polyadenylation site mapping reveals 3' isoform choice variability. *Mol. Syst. Biol.* 11, 1–14.
- Veraldi, K., Arhin, G., 2001. hnRNP F influences binding of a 64-kilodalton subunit of cleavage stimulation factor to mRNA precursors in mouse B cells. ... *Cell. Biol.* 21, 1228–1238. doi:10.1128/MCB.21.4.1228
- Verdel, A., 2004. RNAi-Mediated Targeting of Heterochromatin by the RITS Complex. *Science* (80-.). 303, 672–676. doi:10.1126/science.1093686
- Vogel, C., Abreu, R. de S., Ko, D., Le, S.-Y.Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R., Marcotte, E.M., Penalva, L.O., 2010. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *TL - 6. Mol. Syst. Biol.* 6 VN-re, 400. doi:10.1038/msb.2010.59
- Volpe, T. a, Kidner, C., Hall, I.M., Teng, G., Grewal, S.I.S., Martienssen, R. a, 2002. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* 297, 1833–1837. doi:10.1126/science.1074973
- Wang, D., Mann, J.R., DuBois, R.N., 2005. The role of prostaglandins and other eicosanoids in the gastrointestinal tract. *Gastroenterology* 128, 1445–61. doi:S0016508504017494 [pii]
- Wang, E.T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., Burge, C.B., 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–6. doi:10.1038/nature07509
- Wang, L., Dowell, R.D., Yi, R., 2013. Genome-wide maps of polyadenylation reveal dynamic mRNA 3'-end formation in mammalian cell lineages. *RNA* 19, 413–25. doi:10.1261/rna.035360.112
- Wang, Y., Loomis, P. a, Zinkowski, R.P., Binder, L.I., 1993. A novel tau transcript in cultured human neuroblastoma cells expressing nuclear tau. *J. Cell Biol.* 121, 257–267. doi:10.1083/jcb.121.2.257
- Wehrspau, C.C., Ponting, C.P., Marques, A.C., 2014. Brain-expressed 3'UTR extensions strengthen miRNA cross-talk between ion channel/transporter encoding mRNAs. *Front. Genet.* 5, 41. doi:10.3389/fgene.2014.00041
- West, S., Proudfoot, N.J., 2009. Transcriptional Termination Enhances Protein Expression in Human Cells. *Mol. Cell* 33, 354–364. doi:10.1016/j.molcel.2009.01.008
- West, S., Proudfoot, N.J., 2008. Human Pcf11 enhances degradation of RNA polymerase II-associated nascent RNA and transcriptional termination. *Nucleic Acids Res.* 36, 905–914. doi:10.1093/nar/gkm1112
- White, E., Schlackow, M., Kamieniarz-Gdula, K., Proudfoot, N.J., Gullerova, M., 2014. Human nuclear Dicer restricts the deleterious accumulation of endogenous double-stranded RNA. *Nat. Struct. Mol. Biol.* 21, 552–9. doi:10.1038/nsmb.2827
- Whitelaw, E., 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.
- Wiestner, A., Tehrani, M., Chiorazzi, M., Wright, G., Gibellini, F., Nakayama, K., Liu, H., Rosenwald, A., Muller-Hermelink, H.K., Ott, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Vose, J., Armitage, J.O., Gascoyne, R.D., Connors, J.M., Campo, E., Montserrat, E., Bosch, F., Smeland, E.B., Kvaloy, S., Holte, H., Delabie, J., Fisher, R.I., Grogan, T.M., Miller, T.P., Wilson, W.H., Jaffe, E.S., Staudt, L.M., 2007. Point mutations and genomic deletions in

- CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. *Blood* 109, 4599–606. doi:10.1182/blood-2006-08-039859
- Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., Bähler, J., 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453, 1239–1243. doi:10.1038/nature07002
- Wilkening, S., Pelechano, V., Järvelin, A.I., Tekkedil, M.M., Anders, S., Benes, V., Steinmetz, L.M., 2013. An efficient method for genome-wide polyadenylation site mapping and RNA quantification. *Nucleic Acids Res.* 41, 1–8. doi:10.1093/nar/gks1249
- Wilting, S.M., Snijders, P.J.F., Verlaet, W., Jaspers, a, van de Wiel, M. a, van Wieringen, W.N., Meijer, G. a, Kenter, G.G., Yi, Y., le Sage, C., Agami, R., Meijer, C.J.L.M., Steenbergen, R.D.M., 2013. Altered microRNA expression associated with chromosomal changes contributes to cervical carcinogenesis. *Oncogene* 32, 106–16. doi:10.1038/onc.2012.20
- Wilusz, J.E., Freier, S.M., Spector, D.L., 2008. 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* 135, 919–32. doi:10.1016/j.cell.2008.10.012
- Wood, A.J., Schulz, R., Woodfine, K., Koltowska, K., Beechey, C. V, Peters, J., Bourc'his, D., Oakey, R.J., 2008. Regulation of alternative polyadenylation by genomic imprinting. *Genes Dev.* 22, 1141–6. doi:10.1101/gad.473408
- Woolcock, K.J., Gaidatzis, D., Punga, T., Bühler, M., 2011. Dicer associates with chromatin to repress genome activity in *Schizosaccharomyces pombe*. *Nat. Struct. Mol. Biol.* 18, 94–99. doi:10.1038/nsmb.1935
- Woolcock, K.J., Stunnenberg, R., Gaidatzis, D., Hotz, H.R., Emmerth, S., Barraud, P., Bühler, M., 2012. RNAi keeps Atf1-bound stress response genes in check at nuclear pores. *Genes Dev.* 26, 683–692. doi:10.1101/gad.186866.112
- Xia, Z., Donehower, L. a, Cooper, T. a, Neilson, J.R., Wheeler, D. a, Wagner, E.J., Li, W., 2014. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat. Commun.* 5, 5274. doi:10.1038/ncomms6274
- Xiang, K., Tong, L., Manley, J.L., 2014. Delineating the Structural Blueprint of the Pre-mRNA 3'-End Processing Machinery. *Mol. Cell. Biol.* 34, 1894–910. doi:10.1128/MCB.00084-14
- Yao, C., Biesinger, J., Wan, J., Weng, L., Xing, Y., Xie, X., 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. doi:10.1073/pnas.1211101109
- Yao, P., Potdar, A. a, Arif, A., Ray, P.S., Mukhopadhyay, R., Willard, B., Xu, Y., Yan, J., Saidel, G.M., Fox, P.L., 2012. Coding region polyadenylation generates a truncated tRNA synthetase that counters translation repression. *Cell* 149, 88–100. doi:10.1016/j.cell.2012.02.018
- Yasuda, M., Shabbeer, J., Osawa, M., Desnick, R.J., 2003. Fabry disease: novel alpha-galactosidase A 3'-terminal mutations result in multiple transcripts due to aberrant 3'-end formation. *Am. J. Hum. Genet.* 73, 162–173. doi:10.1086/376608
- Yekta, S., Shih, I.-H., Bartel, D.P., 2004. MicroRNA-directed cleavage of HOXB8 mRNA. *Science* 304, 594–596. doi:10.1126/science.1097434
- Yoon, O.K., Brem, R.B., 2010. Noncanonical transcript forms in yeast and their regulation during environmental stress. *RNA* 16, 1256–67. doi:10.1261/rna.2038810
- Yoon, O.K., Hsu, T.Y., Im, J.H., Brem, R.B., 2012. Genetics and regulatory impact of alternative polyadenylation in human

- B-lymphoblastoid cells. *PLoS Genet.* 8, e1002882. doi:10.1371/journal.pgen.1002882
- Yoon, Y., McKenna, M.C., Rollins, D. a, Song, M., Nuriel, T., Gross, S.S., Xu, G., Glatt, C.E., 2013. Anxiety-associated alternative polyadenylation of the serotonin transporter mRNA confers translational regulation by hnRNPK. *Proc. Natl. Acad. Sci. U. S. A.* 110, 11624–11629. doi:10.1073/pnas.1301485110
- You, L., Wu, J., Feng, Y., Fu, Y., Guo, Y., Long, L., Zhang, H., Luan, Y., Tian, P., Chen, L., Huang, G., Huang, S., Li, Y., Li, J., Chen, C., Zhang, Y., Chen, S., Xu, a., 2014. APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. *Nucleic Acids Res.* 43, D59–D67. doi:10.1093/nar/gku1076
- Young, L.E., Dixon, D. a, 2010. Posttranscriptional Regulation of Cyclooxygenase 2 Expression in Colorectal Cancer. *Curr. Colorectal Cancer Rep.* 6, 60–67. doi:10.1007/s11888-010-0044-3
- Yu, L., Volkert, M.R., 2013. UV damage regulates alternative polyadenylation of the RPB2 gene in yeast. *Nucleic Acids Res.* 41, 3104–14. doi:10.1093/nar/gkt020
- Zhang, H., Lee, J.Y., Tian, B., 2005. Biased alternative polyadenylation in human tissues. *Genome Biol.* 6, R100. doi:10.1186/gb-2005-6-12-r100
- Zhu, Y., Richardson, J.A., Parada, L.F., Graff, J.M., 1998. Smad3 Mutant Mice Develop Metastatic Colorectal Cancer. *Cell* 94, 703–714. doi:10.1016/S0092-8674(00)81730-4