Human cleavage factor I (CFIm) and its role in alternative polyadenylation of pre-mRNA



Jessica Grace Hardy

Sir William Dunn School of Pathology

Lincoln College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity Term 2017

Abstract

Jessica Grace Hardy
Lincoln College

Submitted for the degree of Doctor of Philosophy - Trinity Term 2017

Human cleavage factor I (CFIm) and its role in alternative polyadenylation of pre-mRNA

For many human protein-coding genes, alternative cleavage and polyadenylation (APA) of pre-mRNA generates distinct 3' untranslated regions (3'UTRs) with differing regulatory potential. Widespread 3'UTR shortening via APA occurs in proliferative cell states, including cancer, where it can lead to oncogene overexpression. There has therefore been significant interest in identifying factors which influence poly(A) site choice in different physiological states. The multi-subunit human cleavage factor I complex (CFIm), a core component of the mammalian pre-mRNA cleavage machinery, has been identified as a potential master regulator of APA, as its depletion leads to widespread 3'UTR shortening. However, mechanistic understanding of how CFIm influences poly(A) site selection, and how its activity is regulated, is lacking. In this work, gene editing was used to generate cell lines with substantial, permanent depletion of the 25 kDa or 68 kDa subunits of CFIm (CFIm25 and CFIm68), which exhibited the expected 3'UTR shortening for representative transcripts. Reversal of this 3'UTR shortening by CFIm25 or CFIm68 re-expression provided the basis for a complementation assay, which allowed various aspects of CFIm25 and CFIm68 function to be investigated in vivo. The capacity of CFIm25 to recognise UGUA RNA sequences was shown to make an important contribution to poly(A) site selection transcriptome-wide, and a novel function for the C-terminal arginine/serine-rich (RS domain) of CFIm68 in poly(A) site selection was identified. The potential contribution of CFIm post-translational modification (PTM) to APA regulation was also explored. Novel acetylation sites on CFIm25 and CFIm68 were identified, as well as extensive serine phosphorylation in the CFIm68 RS domain. Complementation analysis revealed that phosphomimetic mutations in this RS domain inhibited distal poly(A) site selection, suggesting a potential role for CFIm68 phosphorylation in APA regulation. Taken together, the findings presented here provide insights into several important determinants of CFIm function, and the complementation assay developed provides a useful tool for future investigations.

Acknowledgements

I am very grateful to the many people who have helped and supported me over the course of the past four years. My sincere thanks first of all go to my supervisor Chris, who has been a fantastic source of advice, help, encouragement and optimism throughout my project. I would also like to thank previous members of the Norbury lab, who made me feel very welcome in my first year. Particular thanks must go to Sophie Fleurdépine, who as well as patiently teaching me a wide range of lab techniques, including radiation work, has been a great friend and source of advice – I will miss our regular lunch meetings! Thanks also go to members of the Sattentau and James labs on OMPI2, who took me under their wing as an 'honorary lab member' when my own lab got a little small, and were great company at lunch and tea breaks!

I have been lucky to have been part of some very productive collaborations during the course of my project, and would like to thank my collaborators for their extensive help and advice. Prof. Andre Furger provided much enthusiastic support and guidance on the 3' end sequencing aspect of my project, as well as kindly allowing me to work in his lab and use the sequencing technology they had available. Members of the Furger lab, in particular Jonathan Neve and Radhika Patel, deserve great thanks for their patience in helping me to troubleshoot sequencing library preparation and for their invaluable assistance with the data analysis. Prof. Shona Murphy has also been a great source of enthusiastic support during my collaborations with her lab, and I would like to express particular thanks to Dr Michael Tellier for his patience in teaching me the ChIP technique, and for the great deal of time and effort he put into analysing the sequencing data.

In addition, I gratefully acknowledge the financial support I have received from a CRUK Oxford Cancer Centre Studentship, which has enabled me to carry out my DPhil studies.

Last, but certainly not least, I would like to express my utmost thanks and dedicate my thesis to my fantastic family and friends, whose support and love has been invaluable. To my wonderful husband Sam, who has shared this adventure of moving to Oxford with me and never failed to cheer me up after a tough day, and my ever-supportive mum and dad, who have always encouraged me and had faith in me – thank you so much. I couldn't have done it without you!

Authorship statement

Work presented in this thesis was performed by me, with the following exceptions:

Chapter 3:

- Design of sgRNAs for *NUDT21* and *CPSF6* targeting, cloning into the pX459 plasmid and initial efficiency testing by HRMA: performed by Dr Andrew Bassett, Genome Engineering Oxford.

- Construction of pcDNA3.1-CFIm25 WT plasmid: performed by Rupen Hargreaves.

- 3' end sequencing analysis - library preparation (for one of the runs), preparation of sequencing chips, sequencing and bioinformatic analysis to generate the scatter plots of pairwise comparisons: performed by members of Prof. Andre Furger's lab, Department of Biochemistry (Dr Jonathan Neve, Radhika Patel, Alastair Louey, Dr Zhiqiao Wang).

Chapter 4:

- 3' end sequencing analysis (as above).

- Live cell microscopy - set-up of microscope and capturing of images as directed by me: performed by Dr Alan Wainman.

- Pol II ChIP-seq – ChIP sample preparation (performed jointly – I processed half the samples alongside), sequencing analysis, bioinformatic analysis of data and generation of resulting metaprofiles etc.: performed by Dr Michael Tellier (sequencing performed by Wellcome Trust Centre for Human Genetics, Oxford).

Chapter 5:

- Mass spectrometry analysis – processing/trypsin digest of gel bands (in two of three cases), running of samples through spectrometer, initial data processing to upload to Mascot server: performed by members of University of Oxford Proteomics Facility (Dr Ben Thomas, Svenja Hester, Prof. Shabaz Mohammed).

Appendix:

- 3' end sequencing and mass spectrometry analysis (as above).

Abbreviations

Abbreviation	Definition
3'READS	3' region extraction and deep sequencing
Ap_4A	diadenosine tetraphosphate
APA	alternative polyadenylation
ARE	AU-rich element
ATP	adenosine triphosphate
BSA	bovine serum albumin
cDNA	complementary DNA
CDS	coding sequence
CFIm	cleavage factor I (mammalian)
CFIIm	cleavage factor II (mammalian)
ChIP	chromatin immunoprecipitation
CLIP	crosslinking and immunoprecipitation
CLK	CDC-like kinase
Co-TC	co-transcriptional cleavage
CPA	cleavage and polyadenylation
CPSF	cleavage and polyadenylation specificity factor
CR-APA	coding region APA
CRISPR	clustered regularly interspaced short palindromic repeats
CSF	cleavage specificity factor
CstF	cleavage stimulation factor
CTD	C-terminal domain (of pol II)
DaPars	dynamic analysis of APA from RNA-seq
DMSO	dimethyl sulphoxide
EMSA	electrophoretic mobility shift assay
EST	expressed sequence tag
FLP	flippase
Flp-In	HEK293 Flp-In T-REx cells
FRT	flippase recognition target
GAR motif	glycine/arginine-rich motif
HEK	human embryonic kidney
HIV	human immunodeficiency virus
HRMA	high-resolution melt analysis
IgM	immunoglobulin M
IP	immunoprecipitation
kb	kilobase
KD	knockdown
LDS	lithium dodecyl sulphate

MCL	mantle cell lymphoma
miRNA	microRNA
mRNA	messenger RNA
MS	mass spectrometry
NHEJ	non-homologous end joining
NLS	nuclear localisation signal
nt	nucleotide
P/D ratio	proximal/distal ratio
pA	poly(A)
PAP	poly(A) polymerase
PAS	polyadenylation signal
PCR	polymerase chain reaction
PEG	polyethylene glycol
pol II	RNA polymerase II
PRR	proline-rich region
PSP	PhosphoSite Plus
PTM	post-translational modification
qPCR	quantitative PCR
qRT-PCR	quantitative reverse transcriptase PCR
RBP	RNA binding protein
RIP	RNA immunoprecipitation
RNAi	RNA interference
RPM	Reads per million
RRM	RNA recognition motif
RS domain	arginine/serine-rich domain
SDS	sodium dodecyl sulphate
SELEX	systematic evolution of ligands by exponential enrichment
SEM	standard error of the mean
seq	sequencing
SF	specificity factor
sgRNA	single-guide RNA
shRNA	short hairpin RNA
siRNA	small interfering RNA
SRPK	serine-arginine protein kinase
TCEP	tris(2-carboxyethyl)phosphine
TEAB	tetraethylammonium bromide
TREX complex	transcription/export complex
TSS	transcription start site
UTR	untranslated region
UTR-APA	untranslated region APA
UV	ultraviolet
WT	wild type

Table of Contents

Abstract i

Acknowledgements ii

Authorship statement iii

Abbreviations iv

Table of Contents vi

Chapter 1 Introduction 1

1.1 mRNA 3' end processing in human cells 1
1.1.1 Endonucleolytic cleavage of the 3' end 1
1.1.2 Polyadenylation of the cleaved 3' end 2
1.1.3 The cleavage and polyadenylation machinery 4
1.1.4 Transcription termination 15
1.2 Alternative polyadenylation (APA) 17
1.2.1 Classes of APA 18
1.2.2 Biological consequences of APA 18
1.2.3 Methods to identify APA sites 20
1.2.4 Modulation of APA under different physiological conditions 22
1.2.5 Factors regulating APA 26
1.3 Biological consequences of UTR-APA 33
1.3.1 Regulatory elements within 3'UTRs 33
1.3.2 Effects of UTR-APA at the single gene level 35
1.3.3 Global effects of UTR-APA 37
1.4 CFIm and its contribution to APA 39
1.4.1 Structure and function of CFIm25 39
1.4.2 Structure and function of CFIm68 42
1.4.3 Structure and function of CFIm59 44
1.4.4 Post-translational modification of CFIm 45
1.4.5 Effect of CFIm depletion on APA 46
1.4.6 Models for regulation of APA by CFIm 48
1.4.7 Physiological regulation and influence of CFIm 49
1.5 Arginine/serine-rich (RS) domains in SR and SR-like proteins 53
1.5.1 Domain organisation and structure of SR proteins 53
1.5.2 Functions and localisation of SR proteins 54
1.5.3 RS domain phosphorylation 57
1.5.4 CFIm68 and CFIm59 as SR-like proteins 59
1.6 Overview of the thesis 60

Chapter 2 Materials and Methods 62

2.1 DNA methods 62
2.1.1 Determination of concentration 62
2.1.2 Genomic DNA extraction 62
2.1.3 Taq PCR 62

2.1.4 Phusion PCR 63
2.1.5 PCR-based site-directed mutagenesis 63
2.1.6 Quantitative PCR (qPCR) 63
2.1.7 Generation of labelled northern probes 64
2.1.8 Agarose gel electrophoresis 64
2.1.9 DNA purification from PCR or gel extraction 65
2.1.10 Restriction enzyme digests 65
2.1.11 Ligation (including TA cloning) 65
2.1.12 Generation of CFIm25/CFIm68 expression plasmids 65
2.1.13 Sanger sequencing 67
2.2 Bacterial methods 67
2.2.1 Preparation of competent cells 67
2.2.2 Transformation of competent cells 68
2.2.3 Plasmid minipreparation ('miniprep') 68
2.2.4 Plasmid midipreparation ('midiprep') 69
2.3 RNA methods 69
2.3.1 RNA extraction and quantification 69
2.3.2 RNA agarose gel electrophoresis 69
2.3.3 Northern blotting 70
2.4 Protein methods 71
2.4.1 Preparation of protein lysates 71
2.4.2 Immunoprecipitation 72
2.4.3 SDS-polyacrylamide gel electrophoresis (PAGE) 73
2.4.4 Coomassie staining 74
2.4.5 Western blotting 74
2.4.6 pIMAGO detection 76
2.4.7 Mass spectrometry (MS) analysis of PTMs 76
2.4.8 Chromatin immunoprecipitation (ChIP) 77
2.5 Mammalian cell culture and analysis methods 78
2.5.1 Cell lines and routine maintenance 78
2.5.2 Cell harvesting and cryopreservation 79
2.5.3 Transfection 79
2.5.4 CRISPR/Cas9-mediated gene disruption 79
2.5.5 CFIm25 complementation experiments 80
2.5.6 Stable cell line generation for CFIm68 complementation experiments 80
2.5.7 Live cell confocal microscopy 80
2.5.8 Cell cycle analysis by flow cytometry 81
2.6 Whole genome/transcriptome methods 81
2.6.1 3' mRNA sequencing analysis 81
2.6.2 Pol II ChIP-sequencing analysis 82

Chapter 3 Establishment of a CFIm complementation assay 84

3.1 Introduction 84
3.2 Generation and identification of CFIm25-depleted cell lines using CRISPR/Cas9 gene editing 85
3.3 Sequence analysis of CFIm25 knockdown cell lines 89
3.4 Analysis of APA in selected CFIm25 knockdown cell lines 90
3.5 Complementation of APA in the 25KD cell line by re-expression of wild type CFIm25 93
3.6 Generation and identification of CFIm68-depleted cell lines using CRISPR/Cas9 gene editing 97
3.7 Sequence analysis of CFIm68 knockdown cell lines 99
3.8 Analysis of APA in selected CFIm68 knockdown cell lines 101
3.9 Complementation of APA in the 68KD cell line by re-expression of wild type CFIm68 103
3.10 Phenotypic analysis of 25KD and 68KD cell lines 105

Chapter 4 Mechanistic determinants of CFIm activity 110

4.1 Introduction 110
4.2 The CFIm25 Y158A Y160A mutant is unable to interact with CFIm68 in vivo 112
4.3 The CFIm68 Δ116-122 deletion mutant is unable to interact with CFIm25 in vivo 115
4.4 The CFIm25-CFIm68 interaction is necessary for normal regulation of APA by both subunits ... 117
4.5 Recognition of UGUA sequences by CFIm25 plays a key role in regulation of APA transcriptome-wide ... 121
4.6 The CFIm68 RS domain is necessary for normal regulation of APA 136
4.7 CFIm68 partial RS domain deletion mutants are able to localise to the nucleus and interact with CFIm25 .. 138
4.8 Transcriptional consequences of CFIm depletion 146

Chapter 5 Post-translational modification of CFIm 155

5.1 Introduction 155
5.2 Selection of known PTM sites for analysis 156
5.3 Identification of new CFIm PTM sites by mass spectrometry 159
5.4 Identification of serine phosphorylation sites in the CFIm68 RS domain 163
5.5 Analysis of CFIm25 PTM site mutants in the APA complementation assay 172
5.6 Analysis of CFIm68 PTM site mutants in the APA complementation assay 178

Chapter 6 Discussion 184

6.1 Roles of individual CFIm subunits and the interplay between them 184
6.2 Mechanistic insights into regulation of poly(A) site selection by CFIm 187
6.3 A role for the CFIm68 RS domain in poly(A) site selection 190
6.4 Identification and functional analysis of CFIm PTMs 193
6.5 Concluding remarks 194

References 196

Appendix 1 Studying proliferation-associated regulation of CFIm and APA ... 214

A1.1 Introduction 214
A1.2 Serum starvation and re-feeding to induce synchronous cell cycle progression in BJ-5ta fibroblasts ... 214
A1.3 Analysis of APA patterns across a cell cycle time course 215
A1.4 Analysis of CFIm expression and PTM profiles across a cell cycle time course 217

Appendix 2 Table of oligonucleotides 221

Appendix 3 CFIm25/CFIm68 vertebrate protein sequence alignments 229

Chapter 1
Introduction

1.1 mRNA 3' end processing in human cells

To generate a mature, translatable eukaryotic messenger RNA (mRNA) from the primary pre-mRNA transcript, several RNA processing reactions need to be completed. Most processing occurs co-transcriptionally, following which the mature transcript can be exported into the cytoplasm for translation. Along with capping and splicing, the other major pre-mRNA processing reactions are cleavage and polyadenylation (CPA), which together generate the mature 3' end of the mRNA. As this thesis explores alternative cleavage and polyadenylation (APA), this section will describe these 3' end processing reactions in detail. Unless otherwise mentioned, discussion will be focused on these processes in mammals.

1.1.1 Endonucleolytic cleavage of the 3' end

Following transcription of the coding region and 3' untranslated region (3'UTR), the pre-mRNA needs to be released from the transcription unit for export. This occurs via endonucleolytic cleavage at the poly(A) site, which is stimulated by recognition of a polyadenylation signal (PAS) and other sequence elements on the transcript.

The consensus PAS was first identified as an AAUAAA sequence around 20-30 nucleotides (nt) upstream of the poly(A) tail in six different mammalian transcripts (Proudfoot and Brownlee, 1976). This finding has since been generalised, with the demonstration that most metazoan protein-coding transcripts possess a PAS 10-30 nt upstream of the cleavage site (Beaudoing et al., 2000; Colgan and

Manley, 1997; Tabaska and Zhang, 1999). The critical contribution of the PAS to transcript cleavage has been confirmed by mutagenesis studies, whereby mutation of this sequence greatly decreases or abolishes cleavage efficiency (Sheets et al., 1990; Wickens and Stephenson, 1984). Although the majority of PAS sequences conform to the canonical AAUAAA or closely related AUUAAA sequence, other similar sequences, such as AGUAAA, have also been shown to act as functional PASs (albeit less efficiently), while ~10% of poly(A) sites have no obvious PAS hexamer (Beaudoing et al., 2000; Neve et al., 2016). The cleavage and polyadenylation specificity factor (CPSF) complex (discussed in Section 1.1.3) recognises the PAS.

Whilst PAS recognition makes a key contribution to poly(A) site definition, this alone is often insufficient to stimulate cleavage. Other less-conserved sequence elements surrounding the poly(A) site, both upstream and downstream, also significantly contribute to cleavage efficiency. In particular, the presence of a GU-rich downstream element recognised by CstF-64 is strongly stimulatory, and upstream UGUA sequences recognised by the CFIm complex also play an important role (Gil and Proudfoot, 1987; MacDonald et al., 1994; Venkataraman et al., 2005). Recognition of these elements is particularly important in aiding CPSF recruitment in the absence of a canonical PAS hexamer. These cleavage factors will be discussed in more detail in Section 1.1.3.

Together, binding of these proteins to sequence elements surrounding the poly(A) site stimulates endonucleolytic cleavage – a reaction performed by the CPSF-73 endonuclease (Mandel et al., 2006). There is some heterogeneity in the exact position of cleavage, although it generally occurs around 10-30 nt downstream of the PAS, with a bias towards CA dinucleotides (Martin et al., 2012). Cleavage defines the end of the final exon and releases the transcript from the polymerase.

1.1.2 Polyadenylation of the cleaved 3' end

Following cleavage, the newly generated 3' end is subject to the addition of a stretch of non-templated adenosine monophosphates - the poly(A) tail. This polyadenylation step is catalysed by poly(A)

polymerase (PAP), and is closely coupled to cleavage via interaction of PAP with the CPSF complex (Bienroth et al., 1993).

Once ~12 adenosines have been added by PAP, the poly(A) tail can be bound by polyadenylate-binding protein nuclear 1 (PABPN1, also known as PABP-2). PABPN1 interacts with PAP and improves its processivity by increasing its affinity for RNA, thus stimulating efficient further elongation of the poly(A) tail (Kuhn et al., 2009). Further extension of the poly(A) tail then allows further PABPN1 molecules to bind. Oligomerised PABPN1 forms a spherical structure and folds the tail back to allow maintenance of the CPSF-PAP interaction, which strongly stimulates PAP activity (Kuhn et al., 2009). In this way, PABPN1 acts as a 'molecular ruler' to limit poly(A) tail length to around 250 nt. Once this length is exceeded, the PABPN1 spherical structure is disrupted, thus decreasing PAP processivity and terminating polyadenylation. The length of the mature poly(A) tail is, however, less than 250 nt due to trimming by deadenylases, with a recent estimate suggesting a median poly(A) tail length of around 50-100 nt in human HeLa cells (Chang et al., 2014).

The poly(A) tail regulates several aspects of mRNA dynamics – in particular, it promotes mRNA export (Fuke and Ohno, 2008) and increases transcript stability and translational efficiency (Sonenberg and Hinnebusch, 2009). Shortening and eventual loss of the poly(A) tail by deadenylation is the first step in the two major mRNA decay pathways in mammals, either promoting decapping by the DCP1-DCP2 complex and 5' to 3' decay, or leaving the 3' end vulnerable to 3' to 5' decay by the exosome complex (Garneau et al., 2007).

In summary, cleavage and polyadenylation are crucial steps in the regulation of gene expression, which together promote the release of the transcript from RNA polymerase II (pol II) and, ultimately, its successful translation in the cytoplasm. The next section will discuss the proteins and complexes involved in these processing reactions in more detail.

1.1.3 The cleavage and polyadenylation machinery

Although the CPA reactions are conceptually simple processing steps, the number of proteins and complexes involved in their regulation and catalysis is surprisingly large. The first attempts to purify and characterise the proteins necessary for CPA in human cells were made in the late 1980s using in vitro cleavage assays with fractionated HeLa cell nuclear extracts. Initial experiments showed that a 'cleavage specificity factor' (CSF) that efficiently cleaves SV40 pre-mRNA could be separated by fractionation from a PAP (Takagaki et al., 1988). Further fractionation experiments led to separation of the CSF complex into four factors with distinct activities – so-called specificity factor (SF), required for specific CPA, cleavage factors I and II (CFI and CFII), which reconstituted active cleavage when mixed with SF, and cleavage stimulation factor (CstF), which significantly enhanced cleavage efficiency (Takagaki et al., 1989). These four multi-subunit factors were shown to be necessary and sufficient for cleavage of the SV40 late transcript, although addition of PAP was necessary for cleavage of other transcripts such as the adenovirus L3 pre-mRNA. Only PAP was needed for the polyadenylation step, although the AAUAAA-linked specificity was only gained in the presence of SF.

In the decades following these initial studies, the CPA machinery has been characterised in greater depth. The five initially identified factors – CPSF (formerly SF), CstF, CFIm, CFIIm (the added 'm' denoting mammalian) and PAP - are still considered the 'core' CPA factors, with the additional protein symplekin acting as an important scaffold for these factors (Takagaki and Manley, 2000). However, more recently, a comprehensive proteomic analysis of an intact cleavage complex isolated using an MS2 tagged pre-mRNA revealed it to be larger and more complex than first thought, containing around 85 proteins (Shi et al., 2009). This approach identified, for example, the serine/threonine phosphatase PP1 as a polyadenylation stimulator, as well as new bona fide cleavage factors such as the WDR33 subunit of CPSF. Proteins from a range of other cellular pathways were also co-purified, including DNA repair proteins and components of the splicing and translation machinery, highlighting the intimate links between 3' end processing and other aspects of nucleic acid

metabolism. It is likely that there is considerable heterogeneity in cleavage complex composition, and that these 'non-core' factors may interact transiently, or only at specific times, with the core complex. This idea is supported by the heterogeneous appearance of purified cleavage complex particles as visualised by electron microscopy (Shi et al., 2009).

There is substantial evidence that the CPA machinery is recruited to the transcription unit by interaction with pol II, which ultimately enhances co-transcriptional poly(A) site recognition and CPA efficiency. Specifically, the pol II C-terminal domain (CTD) has been shown to interact directly with CPSF and CstF and to be necessary for efficient CPA (and other aspects of mRNA processing) in vivo (Hirose and Manley, 1998; McCracken et al., 1997). Phosphorylation of serine 2 within the CTD heptad repeats appears to be important for this coupling (Ahn et al., 2004; Ni et al., 2004), suggesting that CPA factors are recruited by elongating pol II, but additional evidence suggests they can also be recruited at the transcription start site (Glover-Cutter et al., 2008; Nojima et al., 2015). There is also significant interdependency between 3' end processing and splicing – recognition of the 3' splice site in the last intron strongly stimulates 3' end processing, via interactions between the splicing and CPA machinery, and this also reciprocally enhances splicing (Dye and Proudfoot, 1999; Kyburz et al., 2006; Niwa et al., 1990).

The core cleavage factors, illustrated in the context of a poly(A) site in Figure 1.1, will now be discussed in further detail.

Cleavage and polyadenylation specificity factor (CPSF)

CPSF is perhaps the complex most clearly implicated in CPA, as it is involved in both PAS recognition and catalysis of endonucleolytic cleavage. The complex is now considered to contain six subunits – CPSF30, CPSF73 CPSF100, CPSF160 and the two more recently characterised members hFip1 and WDR33.

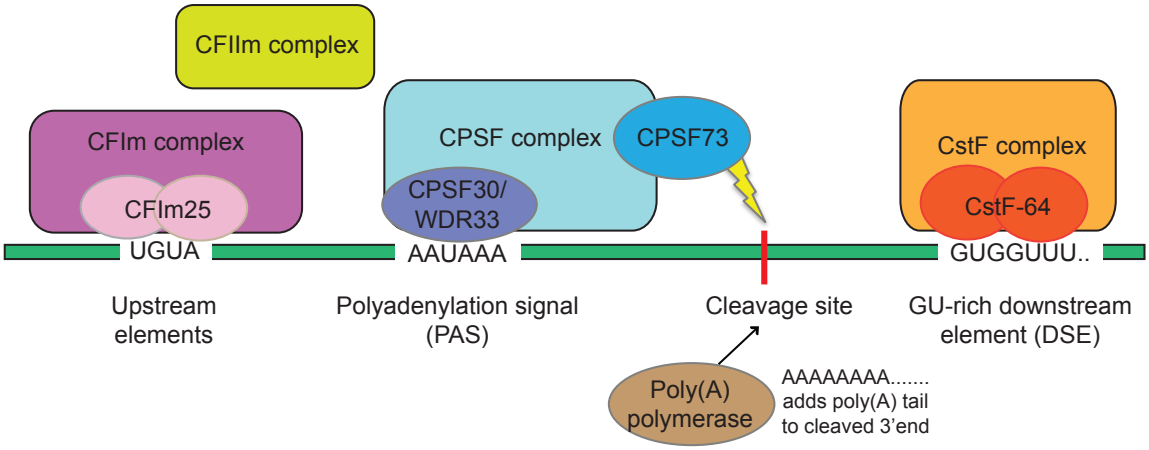


Figure 1.1: The core cleavage factors in the context of a mammalian poly(A) site. Typical, 'strong' poly(A) sites are preceded by a hexameric polyadenylation signal (PAS) around 20 nt upstream, of consensus sequence A(A/U)UAAA, which is recognised by cleavage and polyadenylation specificity factor (CPSF). The downstream GU-rich element is bound by cleavage stimulation factor (CstF), and upstream UGUA sequences by cleavage factor I (CFIm). The pre-mRNA is shown in green with key sequence elements highlighted; boxes represent complexes, and individual protein subunits are shown as ovals. This is a simplified illustration, with only the RNA-binding subunits of CPSF, CstF and CFIm and the CPSF73 endonuclease shown, and many other non-core cleavage factors omitted for simplicity. Cleavage factor II (CFIIm) is not shown bound to mRNA as its binding position/sequence has not been defined.

CPSF has long been known to bind the PAS, but the components responsible for this recognition have only recently been identified. It was long assumed that the CPSF160 subunit recognised the PAS, as a band of the predicted protein size was crosslinked by ultraviolet (UV) to an AAUAAA substrate (Keller et al., 1991), and CPSF160 showed a preference for AAUAAA over point mutants in a binding assay (Murthy and Manley, 1995). However, more recent analysis of cleavage factor binding sites by crosslinking and immunoprecipitation (CLIP) did not show any binding of CPSF160 at PASs (Martin et al., 2012).

Recent work has in fact demonstrated that WDR33, a WD40 repeat-containing protein, and CPSF30, an iron-sulphur cluster protein, both specifically bind the PAS (Chan et al., 2014; Schonemann et al., 2014; Shimberg et al., 2016; Sun et al., 2017). The similar molecular weight of WDR33 and CPSF160 (160 kDa) may explain the previous mistaken identification. Both WDR33 and CPSF30 could be specifically UV cross-linked to AAUAAA-containing RNAs, and transcriptome-wide CLIP analysis identified WDR33 and CPSF30 binding sites in and very close to PAS sequences in vivo (Chan et al., 2014; Schonemann et al., 2014). Analysis of CPSF30 mutants demonstrated that its zinc finger domains are needed for specific PAS recognition, and that this RNA binding is essential for 3' end processing (Chan et al., 2014; Shimberg et al., 2016). Recently, a cryo-electron microscopy structure of a quaternary complex of human CPSF-160, WDR33, CPSF-30 and an AAUAAA RNA molecule was determined at 3.4-Å resolution (Sun et al., 2017). This gave key insights into the molecular basis of specific PAS recognition – the CPSF-30 zinc fingers recognise the A1-A2 and A4-A5 bases, while the U3 and A6 bases form an intramolecular Hoogsteen base pair and are contacted by WDR33.

The findings that WDR33 and CPSF30 recognise the PAS initially called into question the role of CPSF160. The crosslinking pattern observed by CLIP suggests it bind relatively diffusely along the transcription unit, with a small peak around 70 nt upstream of the PAS (Martin et al., 2012), and it is essential for polyadenylation, at least in vitro (Schonemann et al., 2014). However, the recent cryo-electron microscopy study suggested that CPSF160 does also make a crucial indirect contribution to PAS recognition, acting as a scaffold for positioning of WDR33 and CPSF-30 (Sun et al., 2017).

hFip1 is another component of CPSF with RNA binding activity. This member was identified based on amino acid sequence homology to *S. cerevisiae* Fip1, an essential component of the yeast polyadenylation machinery (Kaufmann et al., 2004). Analysis of hFip1 confirmed that it was a bona fide constituent of mammalian CPSF, as western blotting identified its presence within the previously purified HeLa CPSF fraction and co-immunoprecipitation (co-IP) revealed strong interactions with CPSF160, CPSF100 and CPSF73 (Kaufmann et al., 2004). Further in vitro analysis showed that hFip1 stimulates both cleavage and polyadenylation, with the latter mediated through interaction with PAP. Competition assays revealed U-rich sequences as the preferred binding motif, and further analysis found hFip1 binding to U-rich sequences both in the upstream elements and the region between the PAS and the poly(A) site (Kaufmann et al., 2004; Martin et al., 2012). hFip1 therefore appears to be a key member of CPSF, helping to define the poly(A) site by binding to surrounding sequence elements and stimulating 3' end processing through interactions with other members of the complex.

The remaining two CPSF components, CPSF73 and CPSF100, appear to be involved with catalysis of the cleavage reaction itself, and are dispensable for polyadenylation (Schonemann et al., 2014). CPSF73 was identified as the likely pre-mRNA endonuclease through analysis of its domain composition and determination of its crystal structure (Mandel et al., 2006). It is a member of the β-CASP subfamily of the metallo-beta-lactamase (MBL) family, many members of which have been shown to act as nucleic acid hydrolases (Callebaut et al., 2002). Confirmation of nuclease activity was obtained using an in vitro cleavage assay, which showed that CPSF73 could non-specifically cleave pre-mRNA substrates, and that this activity was abolished by mutagenesis of zinc-binding residues (Mandel et al., 2006).

A crystal structure of the yeast homologue of CPSF100 (Ydh1) shows significant similarity to CPSF73, but importantly lacks the zinc-binding residues, suggesting it is unlikely to have nuclease activity (Mandel et al., 2006). Human CPSF100, however, does contain some of the zinc binding residues, and so may have weak nuclease activity (Yang and Doublie, 2011). Mutation of residues in the MBL active site of CPSF100 impaired histone pre-mRNA cleavage in a complementation assay,

suggesting that even if CPSF100 is catalytically inactive, it is likely to stimulate CPSF73 activity (Kolev et al., 2008). It has been suggested that CPSF73 and CPSF100 may form a heterodimer necessary for active cleavage, and supporting this, co-IP experiments have demonstrated a stable CPSF73-CPSF100 interaction in *D. melanogaster* (Sullivan et al., 2009).

In summary, CPSF is a key 3' end-processing complex, which couples endonucleolytic cleavage with recognition of the PAS and other sequence elements surrounding the poly(A) site.

Cleavage stimulation factor (CstF)

CPSF binding alone, however, is generally not sufficient to ensure the efficiency and positional specificity of cleavage. CstF was first identified as a complex that stimulates cleavage in vitro (Takagaki et al., 1989). The complex contains three subunits – CstF-50, CstF-64 (or its homologue CstF-64τ) and CstF-77 – each of which is present in two copies within the assembled complex (Moreno-Morcillo et al., 2011).

CstF-64 is the RNA-binding subunit of CstF, and binds to GU-rich sequences found downstream of poly(A) sites. Such GU-rich sequences generally contain a combination of GU dinucleotides and poly(U) stretches, and a study using systematic evolution of ligands by exponential enrichment (SELEX) demonstrated that these sequences are the preferred binding sites of the CstF-64 N-terminal RNA binding domain (Takagaki and Manley, 1997). The propensity for CstF-64 to bind these downstream GU-rich sequences in vivo has been confirmed by transcriptome-wide CLIP studies (Martin et al., 2012; Yao et al., 2012).

CstF-64 has a closely related homologue, CstF-64τ. This was originally thought to be a testis-specific isoform, but has since been shown to be ubiquitously expressed (Yao et al., 2013). CstF-64τ displays similar RNA-binding specificities to CstF-64, and studies have suggested some degree of functional redundancy – depletion of either subunit led to compensatory increased expression of the other, and

co-depletion was necessary to see effects on poly(A) site usage (Yao et al., 2013; Yao et al., 2012).

Upon binding to GU-rich sequences via CstF-64, the CstF complex interacts with CPSF via CstF-77, aiding in formation and stabilisation of an active cleavage complex. CstF-77 forms a homodimer (Bai et al., 2007; Legrand et al., 2007), which interacts both with CPSF160 (Murthy and Manley, 1995) and with the hinge domain of CstF-64 (Takagaki and Manley, 2000). CstF-77 also provides the scaffold for the assembly of the CstF complex as a whole, linking CstF-64 with CstF-50 (Takagaki and Manley, 2000).

The CstF-50 subunit is a WD-repeat containing protein, which, like CstF-77, forms a homodimer (Moreno-Morcillo et al., 2011). This subunit is key to coupling 3' end processing with transcription, as it interacts directly with the pol II CTD (Fong and Bentley, 2001). Disruption of CstF function by BARD1/BRCA1 binding to CstF-50 contributes to inhibition of 3' end processing in the presence of DNA damage, highlighting the contribution of CstF-50 to efficient cleavage (Edwards et al., 2008).

In summary, the CstF complex binds downstream of poly(A) sites through recognition of GU-rich sequences and stimulates efficient cleavage through interaction with CPSF and pol II.

Cleavage factor I (CFIm)

CFIm was identified as one of the four separable components required to reconstitute efficient in vitro cleavage (Takagaki et al., 1989), and further purification identified three polypeptides of 25, 59 and 68 kDa in the fraction (Ruegsegger et al., 1996). These have since been named CFIm25 (or CPSF5, gene name *NUDT21*), CFIm59 (or CPSF7) and CFIm68 (or CPSF6) – see Figure 1.2. CFIm68 also has a 72 kDa splice isoform, CFIm72, which is generated through inclusion of an additional exon, though this is generally expressed at low levels and has not been extensively studied (Ruegsegger et al., 1998).

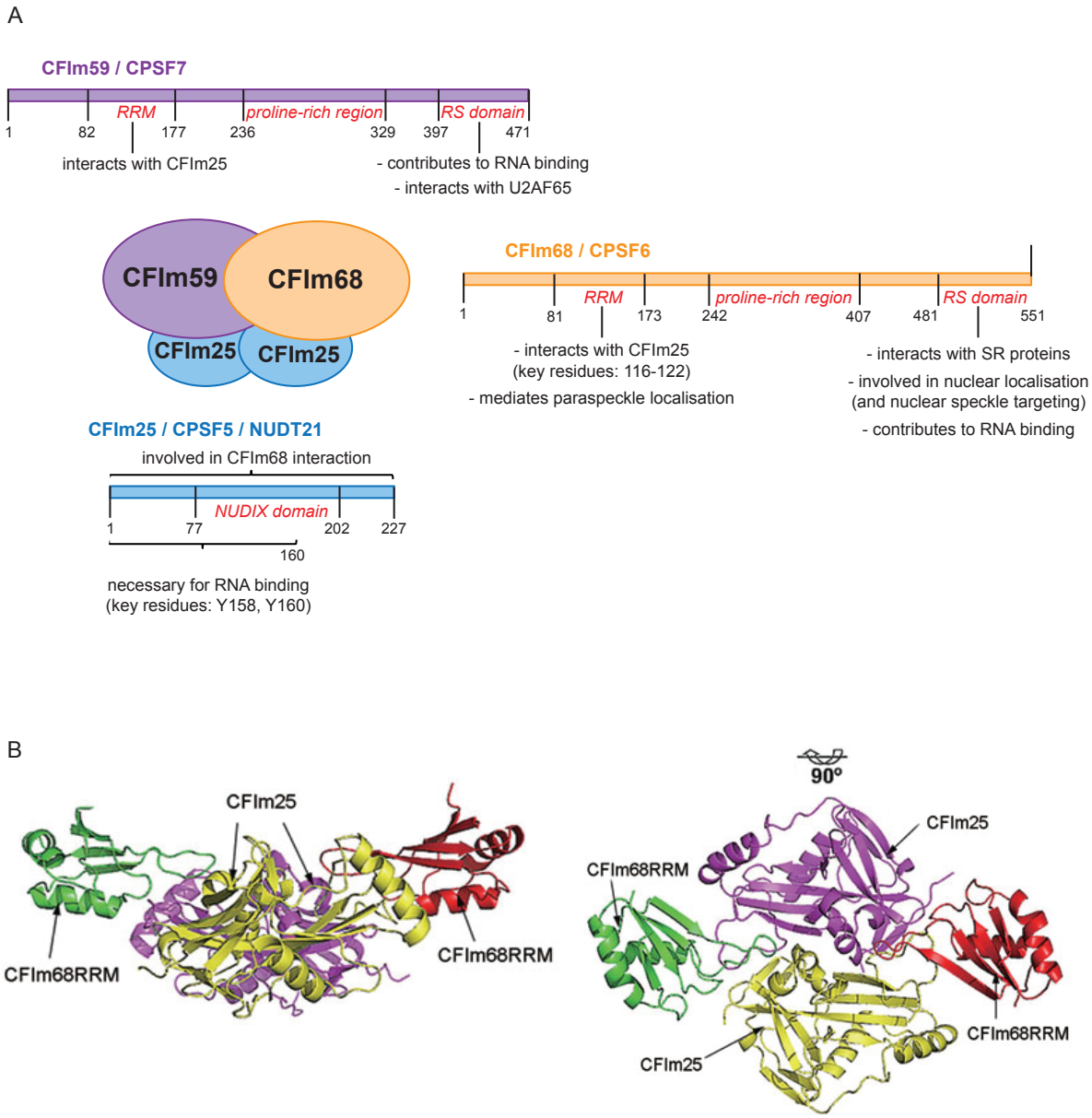


Figure 1.2: Domain organisation and structure of CFIm subunits. (A) Schematic of the domain organisation of the three CFIm subunits (not to scale) and their known roles, alongside the overall heterotetramer structure. A CFIm complex may contain two CFIm59 or two CFIm68 subunits as an alternative to one of each as shown here. RRM = RNA recognition motif, RS = arginine-serine rich. (B) A crystal structure of the CFIm25 homodimer in complex with two CFIm68 RRMs. The right-hand image is rotated 90° around the horizontal axis compared to the left-hand image. Image reproduced with permission from (Li et al., (2011)) under a RightsLink licence.

In vitro experiments suggested that CFIm acts at the earliest stages of poly(A) site recognition, as pre-incubation of CFIm with a pre-mRNA enhanced subsequent cleavage efficiency (Ruegsegger et al., 1998). In addition, chromatin immunoprecipitation (ChIP) experiments identified CFIm throughout the transcription unit in a model gene, supporting the idea that it is poised for early recognition of poly(A) sites (Venkataraman et al., 2005). The observation that CFIm stabilised binding of CPSF-CstF-PAP to RNA provides a possible mechanistic explanation for its ability to enhance cleavage efficiency (Ruegsegger et al., 1996), which is supported by the identification of an interaction between CFIm and the hFip1 subunit of CPSF (Venkataraman et al., 2005). CFIm25 also interacts with PAP and PABPN1, suggesting it may play a role in coupling cleavage with polyadenylation (Dettwiler et al., 2004; Kim and Lee, 2001).

While all CFIm subunits can be UV cross-linked individually to pre-mRNA, CFIm25 is the major RNA-binding subunit of the complex (Dettwiler et al., 2004). SELEX studies identified UGUAN (N = A > U ≥ G/C) as the preferred binding motif of CFIm25 (Brown and Gilmartin, 2003), and structural evidence helped to define the residues involved in this sequence-specific recognition (Yang et al., 2010). UGUA sites are often found around 20 nt upstream of the PAS, which co-localises well with the peak of CFIm binding as identified by CLIP (Martin et al., 2012). The functional importance of UGUA sites was confirmed in vitro, with the demonstration that mutation of these sites in the *PAPOLA/PAPOLG* transcripts significantly reduced CFIm recruitment and impaired cleavage (Venkataraman et al., 2005). The latter study also highlighted the potential for UGUA-bound CFIm to compensate for the lack of a functional PAS, by mediating PAS-independent recruitment of CPSF. Overall, these findings support the notion that CFIm stimulates cleavage via stabilising CPSF recruitment. However, there is also evidence that CFIm binding suppresses cleavage in certain circumstances – for example, binding of CFIm at a UGUA which overlaps a PAS in the *CPSF6* transcript sterically hinders CPSF binding, and thus CPA (Brown and Gilmartin, 2003).

The CFIm59 and CFIm68 subunits have less well-defined roles in the CPA reaction. Though encoded by separate genes, these two subunits show extensive similarity in their sequence, domain organisation

and RNA recognition motif (RRM) ternary structure (Ruegsegger et al., 1998; Yang et al., 2011b) – see Figure 1.2. CFIm68 has been more extensively studied than CFIm59. Although only weakly binding RNA itself, CFIm68 was able to significantly enhance the binding of CFIm25 to RNA in vitro, and the RRM makes clear contacts with the *PAPOLA* RNA in a CFIm25-CFIm68-RNA crystal structure (Dettwiler et al., 2004; Yang et al., 2011a). The arginine-serine rich (RS) domains of the large CFIm subunits are considered protein-protein interaction domains (Dettwiler et al., 2004) and may stimulate CPA via interactions with cleavage factors, though this has not been directly demonstrated.

Original indications from in vitro experiments were that functional CFIm could be reconstituted as a CFIm25-CFIm59 or CFIm25-CFIm68 heterodimer (Ruegsegger et al., 1998), but size exclusion chromatography and crystal structure evidence suggest that it in fact exists as a heterotetramer (Kim et al., 2010; Yang et al., 2011a). The heterotetramer contains a CFIm25 homodimer, which has been well characterised in several crystal structures (Coseno et al., 2008; Li et al., 2011; Yang et al., 2011a; Yang et al., 2010), and each CFIm25 subunit interacts with a CFIm59 or CFIm68 subunit (Figure 1.2). Only the RRM of CFIm59/CFIm68 has been crystallised in complex with CFIm25 (Li et al., 2011; Yang et al., 2011a), and it is thought that functional CFIm may contain two CFIm68/CFIm72 subunits, two CFIm59 subunits, or one of each (Kim et al., 2010).

In summary, CFIm binds upstream of cleavage sites, often through recognition of UGUA sequences, and in general appears to stimulate cleavage via interaction with other cleavage factors. As this thesis focuses on CFIm, a more detailed description of its structure and function is given in Section 1.4.

Cleavage factor II (CFIIm)

CFIIm, the final 'core' cleavage factor identified in the fractionation experiment (Takagaki et al., 1989), has since been shown to comprise two subunits - Pcf11 and Clp1 (de Vries et al., 2000). These subunits were named on the basis of their similarity to the *S. cerevisiae* Pcf11 and Clp1 proteins,

which are part of the *S. cerevisiae* CF IA complex. Both subunits were shown to be essential for cleavage in vitro, with immunodepletion of the complex leading to cleavage inhibition (de Vries et al., 2000). However, in contrast to their *S. cerevisiae* orthologues, human Pcf11 and Clp1 are not required for polyadenylation, suggesting they have differing functions in mammals and yeast (de Vries et al., 2000).

Immunoprecipitation (IP) experiments demonstrated that human Clp1 interacts directly with CFIm and CPSF, suggesting that it may bridge interactions between these two cleavage factors (de Vries et al., 2000). Aside from this observation, information about the role of Clp1 in cleavage is lacking, although its role as a polynucleotide kinase involved in tRNA splicing is well characterised (Hanada et al., 2013). It is not clear whether this kinase activity is also involved in mRNA 3' end cleavage.

Human Pcf11 has mainly been studied with respect to its role in transcription termination. Knockdown of Pcf11 leads to impaired transcription termination and read-through transcription, suggesting that it may be involved in promoting pol II dissociation from the template (West and Proudfoot, 2008). In addition, Pcf11 acts in a PAS-dependent manner to promote degradation of the 3' cleaved product, probably in co-operation with Xrn2 (West and Proudfoot, 2008). However, the additional contribution of Pcf11 to the pre-cleavage/cleavage step, as identified by in vitro studies, has not been characterised and requires further investigation.

In summary, CFIIm has a poorly defined, though essential, role in 3' end cleavage and additionally makes an important contribution to effective transcription termination.

Poly(A) polymerase (PAP)

The major nuclear poly(A) polymerase (PAP), PAPα (encoded by the *PAPOLA* gene), was first purified from calf thymus, on the basis of its ability to catalyse polyadenylation (Winters and Edmonds, 1973a, b). Although initial experiments suggested PAP may only be necessary for

polyadenylation, follow-up work demonstrated that it is also essential for the cleavage step (Christofori and Keller, 1989; Takagaki et al., 1989). Other PAP isoforms have since been identified, and PAPγ (encoded by the *PAPOLG* gene) has a similar structure and enzymatic activity to PAPα (Kyriakopoulou et al., 2001; Yang et al., 2014). It is not yet clear whether these different isoforms have specialist roles or whether they display functional redundancy.

PAP alone has low affinity for RNA and adds nucleotides inefficiently in a distributive manner, without sequence specificity (Wahle, 1991). However, direct interaction with PABPN1 and CPSF bound to the PAS stimulates specific and processive polyadenylation (Bienroth et al., 1993), as discussed in Section 1.1.2. The requirement of PABPN1 for efficient polyadenylation may warrant it also being considered a 'core' CPA factor, even if it was not identified as such in vitro. Elucidation of a bovine PAPα crystal structure identified two adenosine triphosphate (ATP) binding sites, only one of which appears to play a role in catalysis of nucleotide addition (Martin et al., 2004). The mechanism by which PAP stimulates the cleavage step remains elusive, but may be connected to its interactions with CPSF and CFIm.

1.1.4 Transcription termination

The CPA reaction is intrinsically linked with the subsequent process of transcription termination, whereby the RNA cleavage product downstream of the poly(A) site is degraded and pol II dissociates from the DNA template. Effective transcription termination is critical to prevent problems associated with read-through transcription, such as impaired initiation at downstream promoters, and also ensures timely recycling of pol II to allow further rounds of initiation.

Pol II transcription termination occurs at heterogeneous sites up to several kilobases (kb) downstream of the poly(A) site (Proudfoot, 1989) and transcription of the PAS is required to trigger the process, with PAS mutation leading to impaired termination (Connelly and Manley, 1988). There are two

prevailing models for how the PAS triggers termination. The first is the allosteric model, which suggests that pol II senses transcription of the PAS and undergoes a conformational change that promotes its release from the DNA. Some support for this was provided by the ability of a pol II CTD monoclonal antibody to disassemble an elongation complex, suggesting that changing interactions with CPA factors upon transcription of the PAS may have a similar effect (Zhang et al., 2004). Further work using an in vitro termination assay suggested that inhibition of cleavage did not impair termination and that alpha-amanitin could inhibit the PAS-dependent polymerase dissociation, again partially supporting the possibility of a conformational change-based mechanism (Zhang et al., 2015).

The second model suggests that the PAS-dependency of termination is explained by a causal role for transcript cleavage in the process. This ‘torpedo’ model proposes that the pol II-associated free 5' end is subject to degradation by a 5'-to-3' exonuclease, which ‘catches’ pol II and forces its release from the DNA (Connelly and Manley, 1988). Evidence for this mechanism was provided by the observation that depletion of the mammalian XRN2 exonuclease impaired termination at the beta-globin gene (West et al., 2004), and similar observations were made upon deletion of the homologous RAT1 gene in yeast (Kim et al., 2004). Although some genome-wide analyses failed to show significant effects of small interfering RNA (siRNA)-mediated XRN2 depletion on termination (Nojima et al., 2015), experiments including a dominant negative nuclease-dead XRN2 have demonstrated a clear, widespread termination defect (Fong et al., 2015).

The allosteric and torpedo models both have good supporting evidence, and need not be mutually exclusive. Indeed, it was found that XRN2 inhibition decreased the efficiency of conformational change-induced termination in vitro (Zhang et al., 2015), suggesting that a pol II conformational change and XRN2-mediated degradation of the 3' cleavage product may co-operate to stimulate termination.

An alternative termination mechanism found to apply to some genes, including beta-globin, involves cleavage and XRN2-mediated degradation at a site other than the poly(A) site. This co-transcriptional

cleavage (Co-TC) occurs when a self-cleaving Co-TC sequence some way downstream of the PAS is transcribed before CPA has occurred. Immediate cleavage at this site provides a 5' end for degradation by XRN2, triggering dissociation of the polymerase (West et al., 2004; White et al., 2013). Cleavage at the poly(A) site in such cases thus occurs post-transcriptionally.

A feature that is key to termination in general is pausing of pol II following PAS transcription. Depletion of CPSF-73 and CstF-64 diminishes the peak of paused pol II following the PAS (Nojima et al., 2015), suggesting that binding of CPA factors to pol II contributes to pausing. In addition, G-rich terminator sequences following the PAS can promote formation of R-loops and associated heterochromatin, which can have a key role in further promoting pol II pausing (Gromak et al., 2006; Skourti-Stathaki et al., 2014; Skourti-Stathaki et al., 2011). This pausing provides more time for XRN2 to degrade the pol II-associated RNA and stimulate dissociation, enhancing termination efficiency (Gromak et al., 2006).

In summary, the CPA site makes an important stimulatory contribution to subsequent transcription termination, and thus is important for helping to define separate gene units and maintain genome integrity.

1.2 Alternative polyadenylation (APA)

While CPA could be considered a constitutive, essential stage of pre-mRNA processing, it is in fact also subject to considerable regulation and able to contribute to transcriptomic and proteomic diversity. This is owing to the fact that most genes contain more than one site where CPA may occur. Use of different poly(A) sites is known as alternative polyadenylation (APA), and ultimately leads to the generation of distinct mRNA transcripts.

1.2.1 Classes of APA

APA events can be divided into two major classes depending on the location of alternative poly(A) sites (Figure 1.3). Often, alternative poly(A) sites are found in the 3'UTR, and in this situation ('UTR-APA'), use of different sites changes the length of the 3'UTR but not the protein encoded.

Less commonly, alternative poly(A) sites may be found within the coding region, and use of these alters the protein-coding capacity of the transcript as well as the 3'UTR. This is known as coding region APA (CR-APA). CR-APA events can be further sub-classified by the location of the APA site within the coding region (Figure 1.3). 'Intronic' or 'composite exon' APA involves use of an intronic poly(A) site, giving an extended CDS and alternative 3'UTR composed of intronic sequence, while 'alternative terminal exon' APA describes the use of sites at the 3' ends of different terminal exons, and is closely linked to alternative splicing decisions.

1.2.2 Biological consequences of APA

The first identified examples of APA were mostly 'CR-APA' events, which result in alternative protein isoforms being generated. A particularly well-studied example is that of CR-APA in the immunoglobulin M (IgM) heavy chain gene. Use of two alternative poly(A) sites was found to alter the C-terminal protein sequence in such a way as to alter function and localisation (Alt et al., 1980; Early et al., 1980). Specifically, a switch to the more proximal poly(A) site upon B cell maturation excludes the C-terminal domain required for IgM membrane retention, leading to production of a secreted isoform (Alt et al., 1980; Early et al., 1980). Other examples of composite exon CR-APA have since been identified in a range of other genes, with varying consequences on protein function (Edwalds-Gilbert et al., 1997). For example, intronic APA in the cytokine receptor gp130 generates a soluble isoform, which, unlike its membrane-bound counterpart, inhibits trans-signalling through the IL-6 receptor (Sommer et al., 2014).

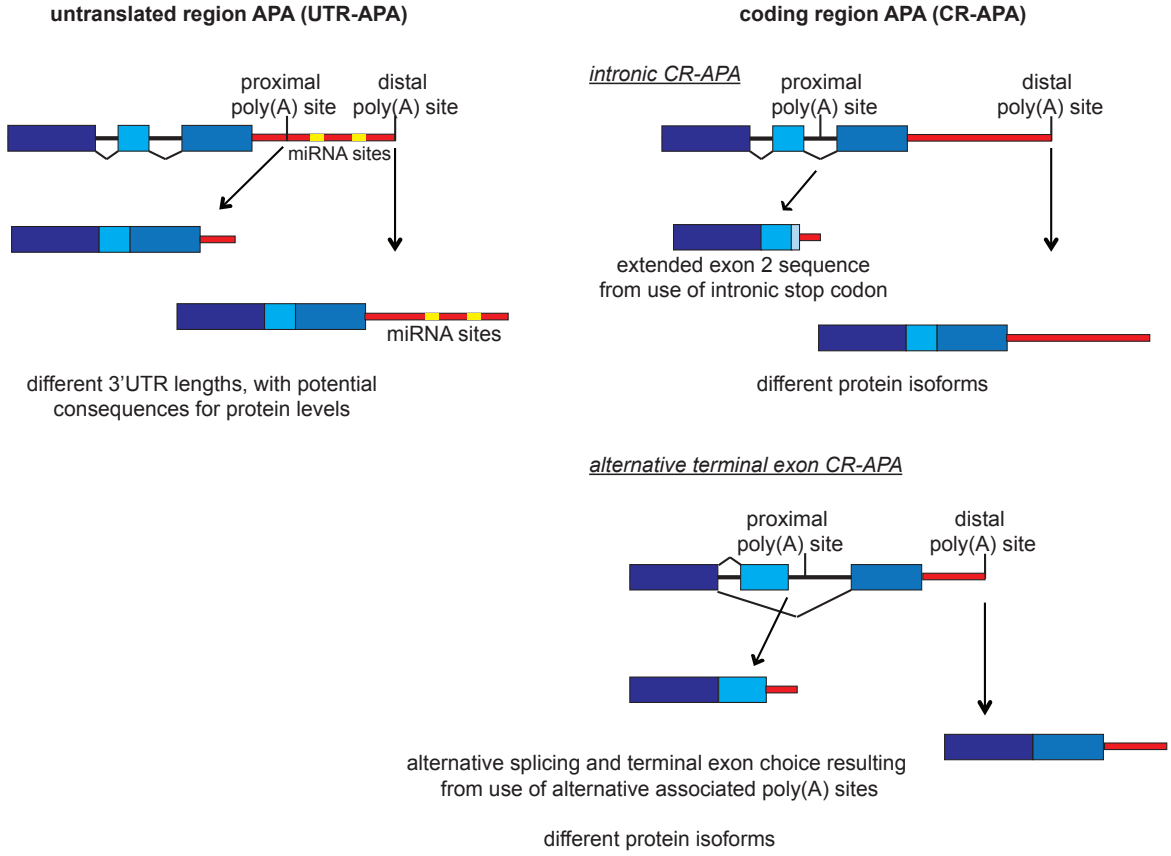


Figure 1.3: The two classes of alternative polyadenylation. Left - UTR-APA alters the length and content of the 3'UTR without altering protein coding sequence. This example shows the exclusion of miRNA sites, which is just one potential consequence of UTR-APA. Right - CR-APA alters the coding sequence of a transcript as well as the 3'UTR. Two of the major types of CR-APA are shown here. Blue boxes: protein coding sequence; black lines: introns; red boxes: 3'UTR; yellow boxes: miRNA sites. Adapted from (Hardy, J.G. and Norbury, C. J. (2016) Cleavage factor Im (CFIm) as a regulator of alternative polyadenylation. Biochem Soc Trans 44, 1051-1057).

A particularly well-studied example of alternative terminal exon CR-APA is in the calcitonin/calcitonin gene-related peptide (CGRP) gene. This gene undergoes tissue-specific RNA processing, such that neurons express CGRP while other cell types express calcitonin (Crenshaw et al., 1987). The difference in protein sequence is achieved through use of exon 4 as the terminal exon in all tissues except neuronal tissue. It was demonstrated that an intronic splicing enhancer promotes polyadenylation at the exon 4-associated poly(A) site in non-neuronal tissues, forcing the use of exon 4 as the terminal exon (Lou et al., 1996). This was shown to involve recruitment of SRp20 at the enhancer, which in turn recruits CstF-64 to the nearby poly(A) site (Lou et al., 1998).

Potential biological consequences of UTR-APA may not be as readily apparent as for CR-APA, as UTR-APA does not induce a change in protein identity. However, changes in the length and resulting content of the 3'UTR can have a significant impact on gene expression. 3'UTRs are platforms for a range of sequence elements that control aspects of mRNA fate such as stability, localisation and translation. This topic will be discussed in detail in Section 1.3.

1.2.3 Methods to identify APA sites

In order to investigate the extent of APA and its biological consequences, it is necessary to annotate APA sites in organisms of interest, ideally genome-wide.

Initial identifications of APA events were at the single-gene level, and often followed observation of multiple bands upon northern blotting. The earliest example was the identification of four dihydrofolate reductase (DHFR) transcripts differing only in 3'UTR length (Setzer et al., 1980). Although northern blotting is a low-throughput technique, it still remains useful for analysing regulation of APA at the single-gene level, for example to identify changes in APA upon RNA binding protein (RBP) knockdown (Katahira et al., 2013; Kubo et al., 2006). Northern blotting presents the advantage of allowing direct visualisation of mRNA sizes without the potentially bias-

introducing sample manipulation and data processing involved with more complex techniques. Quantitative reverse transcriptase polymerase chain reaction (qRT-PCR) provides a less laborious and more sensitive alternative for single-transcript APA analysis and has been used with success in some studies (Masamha et al., 2014). However, data interpretation can be complex due to the use of primer pairs with different amplification efficiencies for each poly(A) site, and the need for subtractive calculations to infer relative poly(A) site use.

The development of high-throughput, genome-wide technologies has allowed more extensive study of APA at the global level. The first attempts at genome-wide APA annotation exploited the growing database of expressed sequence tag (EST) data. EST reads containing poly(A) stretches were identified, aligned to the genome and validated by identification of a nearby PAS, allowing annotation of thousands of poly(A) sites (Tian et al., 2005; Yan and Marr, 2005). Such analyses revealed APA to be widespread, with APA sites identified in ~50% of human genes.

Microarray technology is another high-throughput method that has been exploited in APA analysis. While unable to identify new APA events, use of specifically designed probes allows the dynamics of known APA events to be investigated. This approach revealed, for example, generalised 3'UTR shortening in immune cell activation (Sandberg et al., 2008) and widespread changes in 3'UTR length during generation of induced pluripotent stem cells (Ji and Tian, 2009).

More recently, next-generation sequencing technologies have facilitated more accurate and sensitive mapping of APA sites. Whole RNA sequencing (RNA-seq) using platforms such as Illumina is highly quantitative, and theoretically allows quantification of APA site usage either by identification of reads with poly(A) stretches or by the use of algorithms such as ‘Dynamic analyses of APA from RNA-seq’ (DaPars) to predict locations of poly(A) sites based on read density. This particular algorithm was developed to compare APA site usage in RNA-seq datasets between hundreds of tumour/normal tissue pairs (Xia et al., 2014). However, the fact that only a very small fraction of reads in a standard RNA-seq experiment map to poly(A) sites limits the sensitivity of this approach for APA analysis.

Many 3' end-biased RNA sequencing methods have been developed for APA analysis to overcome this limitation of standard RNA-seq (Mueller et al., 2013). Most of these methods use an oligo(dT) reverse priming step following fragmentation in order to select and subsequently amplify polyadenylated 3' end fragments. Such methods include 3SEQ (Beck et al., 2010), SAPAS (Fu et al., 2011) and A-seq (Martin et al., 2012), which sequences in the sense direction and includes a split oligo(dT) primer to accurately pinpoint the cleavage location. The QuantSeq approach, used in this thesis (see Sections 3.5 and 4.5) is another example of this type of method, and the workflow is illustrated in Figure 1.4 (Moll et al., 2014).

One potential problem arising from oligo(dT) reverse priming is the potential for internal priming at A-rich sequences. A number of methods have been developed to overcome this limitation, including 3P-Seq (Jan et al., 2011) and 3'READS (Hoque et al., 2013). 3P-Seq involves initial ligation of a biotin-labelled adaptor to the intact poly(A) tail, while 3'READS uses long oligo(dT) primers and stringent hybridisation conditions, both of which eliminate internal priming.

The increased availability and sophistication of next-generation 3' end-biased mRNA sequencing approaches has led to a wealth of transcriptome-wide APA studies. Estimates from these studies suggest that 70-80% of mammalian mRNA genes are subject to APA (Derti et al., 2012; Hoque et al., 2013). However, similar analyses have also revealed pervasive APA across a wide range of non-mammalian organisms, including fission yeast, which have annotated APA sites in ~45% of genes (Mata, 2013), plants, which have APA sites in ~70% of genes (Wu et al., 2016) and zebrafish, which have APA sites in ~55% of genes (Ulitsky et al., 2012).

1.2.4 Modulation of APA under different physiological conditions

The availability of 3' end-biased sequencing approaches and complementary methods has allowed global APA patterns to be compared across a range of cell types and physiological states. Such studies

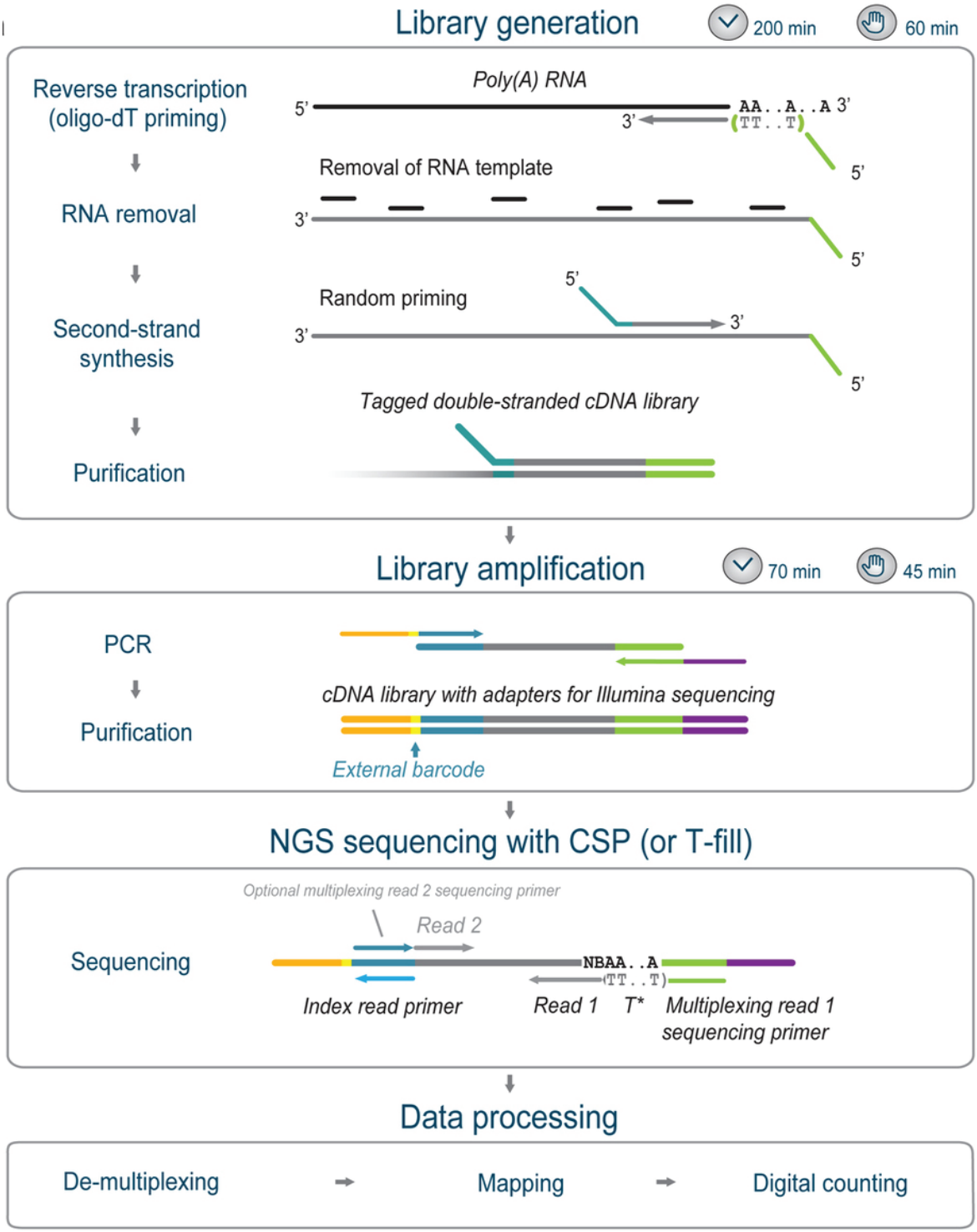


Figure 1.4: The QuantSeq method for generating 3' end mRNA sequencing libraries. This is an example of an oligo-dT reverse priming-based approach to generating 3' end sequencing libraries. Specific coverage of the 3' end is achieved through incorporation of the adapters during first- and second-strand synthesis - only the random primer hybridising closest to the 3' end will be fully extended to generate a product with adapters at both ends. Image reproduced with permission from: (Moll et al., (2014)) under a RightsLink licence.

have revealed that the usage frequency of specific poly(A) sites is highly changeable, and subject to tissue- and condition-specific regulation.

A number of studies have identified substantial differences in APA patterns between tissues (Lianoglou et al., 2013; Wang et al., 2008; Zhang et al., 2005). Tissue-specific APA variation was found to be two- to three-fold greater than variation between individual humans (Wang et al., 2008), suggesting potentially important regulatory roles. Indeed, gene ontology analysis demonstrated that genes showing distinct 3'UTR shortening in a specific tissue are often involved in a relevant tissue-specific process – for example, many genes with testis-specific 3'UTR shortening are implicated in spermatogenesis (Lianoglou et al., 2013). This suggests that ubiquitously expressed genes may employ APA to achieve tissue-specific expression patterns.

One particularly well-defined tissue-specific APA pattern is the striking prevalence of transcripts with unusually long 3'UTRs in neuronal tissues. Extensive use of distal poly(A) sites has been noted in brain tissue of mammals (Miura et al., 2013; Zhang et al., 2005) and *D. melanogaster* (Hilgers et al., 2012), and is often associated with genes important for neural function. Such sites are often exclusively used in the brain, and can be several kb downstream of poly(A) sites used in other tissues (Miura et al., 2013). Mechanisms driving neuronal distal poly(A) site selection will be discussed further in Section 1.2.5.

APA patterns are also highly variable between different physiological states, with a key example being the general association of more proliferative cell states with widespread APA-mediated 3'UTR shortening. This was first observed upon activation of murine $CD4^+$ T lymphocytes and further extended to other examples of immune cell activation (Sandberg et al., 2008). A negative correlation was observed between proliferative capacity of a cell type and its average 3'UTR length (Sandberg et al., 2008), and widespread 3'UTR shortening was shown to accompany the transition from quiescence into proliferation in human fibroblasts and epithelial cells (Elkon et al., 2012). Consistently, re-programming of somatic cells to generate induced pluripotent stem cells, which is associated with an

increase in proliferative capacity, also led to a global 3'UTR shortening, though the extent was dependent on the cell source (Ji and Tian, 2009).

Furthermore, global 3'UTR shortening has been observed upon cell transformation and cancer development. Pairwise comparisons between cancer cell lines and non-transformed lines of similar origin and proliferative capacity revealed an even greater 3'UTR shortening associated with the cancerous state for several oncogene-encoding transcripts (Mayr and Bartel, 2009). Clinical observations have generally supported these findings – for example, a comparison of five different tumour types with corresponding normal tissue highlighted an overall decrease in distal poly(A) site usage (Lin et al., 2012). However, another cell line study found no additional contribution of transformation to 3'UTR shortening above that expected from the increased proliferative state, which was perhaps a consequence of the different cell types and transformation processes studied (Elkon et al., 2012).

Development is another fundamental process demonstrating global APA changes. Specifically, mouse embryonic development involves progressive 3'UTR lengthening, and this has been recapitulated in a myoblast to myotube differentiation model (Ji et al., 2009; Li et al., 2015). This 3'UTR lengthening contrasts with the proliferation-linked 3'UTR shortening, which may reflect the progressively decreasing proliferative capacity of cells during development.

Global APA changes have also been observed in response to specific extracellular cues, including mTORC1 signalling and neuronal activation (Chang et al., 2015; Flavell et al., 2008). For example, activity-dependent induction of target genes of the MEF2 transcription factor in neurons also promotes increased proximal poly(A) site use in the MEF2 targets, leading to generation of alternative protein isoforms through CR-APA and 3'UTR shortening through UTR-APA (Flavell et al., 2008).

An important point when considering these so-called ‘global’ shifts in APA patterns is that often only a subset of genes with APA sites are involved, and that while the majority of events may be in one

direction (e.g. 3'UTR shortening), there are usually exceptions. For example, in the fibroblast proliferation model, ~35% of transcripts with APA sites showed a significant change in 3'UTR length, and 83% of these events involved 3'UTR shortening (Elkon et al., 2012). While clearly a directional phenomenon, the 17% of changes involving 3'UTR lengthening may also be biologically relevant.

Another key consideration when evaluating changes in APA between different conditions is whether such changes can definitively be attributed to altered poly(A) site selection, or whether they may alternatively reflect differential regulation of mRNA isoforms. For example, an increased relative abundance of a short 3'UTR isoform in proliferating cells could result from increased usage of the proximal poly(A) site, or might reflect decreased stability of the longer 3'UTR isoform in proliferating cells. One approach that has been developed to address this issue involves subcellular fractionation prior to 3' end sequencing (Neve et al., 2016). Using this method, APA changes that are observed in the nuclear fraction are likely to reflect changes in poly(A) site selection, while differing APA profiles between the cytoplasm and nucleus would indicate differential post-transcriptional regulation of isoforms.

In summary, APA patterns are subject to modulation, often of a global and directional nature, in response to a variety of a different physiological states and signals. The potential regulatory mechanisms underlying such global changes will be considered further in Section 1.2.5, and possible biological consequences will be discussed in Section 1.3.

1.2.5 Factors regulating APA

The observed plasticity of APA raises the key question of which factors and variables influence the choice of specific poly(A) sites, and how this choice is regulated in different physiological states. As discussed in Section 1.1.1, various sequence elements around a cleavage site, including the PAS, GU-rich sequence and upstream UGUA sequences, stimulate assembly of the cleavage machinery. Thus, at

the simplest level, the relative 'strength' of a cleavage site may be somewhat defined by the presence and positions of these elements, as well as their similarity to consensus sequences. Analysis of APA in human embryonic kidney 293 (HEK293) cells demonstrated that distal poly(A) sites are much more likely to possess a canonical PAS than proximal sites, and consistently are used substantially more frequently (Martin et al., 2012). Other studies confirmed the importance of the GU-rich sequence and UGUA sequences for stimulating cleavage, particularly in the absence of a canonical PAS (Nunes et al., 2010; Venkataraman et al., 2005). However, given the fact that the ratio of proximal to distal site CPA events for a given transcript is subject to condition-specific variation, cis elements alone cannot fully explain relative poly(A) site preferences.

Cleavage factors

As may be expected, trans-acting cleavage factors have been strongly implicated as regulators of poly(A) site choice. The factors studied and their effects are summarized in Table 1.1. One general observation is that the overall expression level of 3' end processing factors negatively correlates with 3'UTR length (Elkon et al., 2012; Ji et al., 2009; Mayr and Bartel, 2009), suggesting that increased cleavage factor expression may preferentially enhance cleavage at sub-optimal proximal poly(A) sites. A likely explanation for this is that increased cleavage factor concentration makes selection of proximal poly(A) sites on a 'first-come, first-served' basis more likely. One study identified up-regulation of members of the CPSF, CstF, CFIm and CFIIm complexes during the shift from quiescence to proliferation, correlated this with proliferation-linked 3'UTR shortening, and found some evidence of co-regulation of these factors by the proliferation-associated transcription factor E2F (Elkon et al., 2012). These findings suggest a potential functional contribution of overall changes in cleavage factor expression to proliferation-linked APA modulation, but at this stage the evidence remains mainly correlative.

Table 1.1: Factors involved in widespread directional selection of poly(A) sites.

Factors favouring proximal poly(A) sites (depletion = widespread 3'UTR lengthening)	**Factors favouring distal poly(A) sites (depletion = widespread 3'UTR shortening)**
CstF-64 and CstF-64τ (together)	CFIm25
hFip1	CFIm68
Pcf11	PABPN1
CPEB1 (for CPE-containing genes)	U1 snRNA
	ELAV (*Drosophila* brain)
	THOC5
	PAF1C

One factor conforming to the inverse correlation between cleavage factor expression and 3'UTR length is CstF-64, and knockdown experiments have provided more direct support for a causal role in poly(A) site choice. While depletion of CstF-64 alone had a negligible effect on global APA in HeLa cells, co-depletion of its homologue, CstF-64τ, led to a clear directional effect, with 85% of events involving 3'UTR lengthening (Yao et al., 2012). CLIP analysis of CstF-64 and CstF-64τ showed binding at proximal and distal poly(A) sites, but highlighted a slight preference at proximal poly(A) sites in CstF-64-regulated transcripts, suggesting a model whereby these factors preferentially enhance cleavage at proximal sites (Yao et al., 2013; Yao et al., 2012). CstF-64 has long been implicated in APA on an individual gene basis, with a suggested role in CR-APA at the IgM heavy chain transcript. The proliferation-linked switch from membrane-bound to secreted IgM, discussed in Section 1.2.2, appeared to be stimulated by increased CstF-64 expression (Takagaki et al., 1996). Consistently, subsequent studies identified a five-fold increase in CstF-64 expression during two other proliferative events - serum stimulation of fibroblasts and lipopolysaccharide stimulation of macrophages (Martincic et al., 1998; Shell et al., 2005). However, these studies also cast doubt on the causal role of the CstF-64 increase in the IgM APA switch, and also identified only a limited overlap between proliferation-linked and CstF-64 overexpression-induced APA events. In summary, while CstF-64 and

CstF-64τ together seem to generally enhance proximal poly(A) site selection, further evidence is needed to support causal links between these factors and global APA modulation.

Among the CPSF subunits, hFip1 expression has the clearest effect on APA, again demonstrating the negative correlation between expression level and 3'UTR length (Lackford et al., 2014; Li et al., 2015). hFip1 was found to be highly expressed in embryonic stem cells (ESCs) compared to differentiating cells, and shown to be essential for ESC self-renewal (Lackford et al., 2014). This was partly attributed to its role in enhancing proximal poly(A) site use in renewal-associated genes. This suggests that hFip1 may represent a key APA regulator in the context of stem cell differentiation. Individual knockdown of other CPSF subunits, perhaps surprisingly, had little effect on APA patterns (Li et al., 2015), suggesting that they may have limited or redundant roles in poly(A) site selection. Depletion of the Pcf11 subunit of CFIIm in the same study, however, did lead to clear 3'UTR lengthening (Li et al., 2015). In summary, certain members of the CstF, CPSF and CFIIm complexes conform to the general observation that increased cleavage factor expression preferentially enhances use of proximal poly(A) sites and leads to 3'UTR shortening.

Contrary to this general trend, knockdown of the CFIm25 or CFIm68 subunits of the CFIm complex leads to a widespread and highly directional 3'UTR shortening rather than lengthening (Gruber et al., 2012; Kubo et al., 2006; Masamha et al., 2014). This suggests that CFIm either preferentially stimulates the use of distal poly(A) sites, or alternatively or additionally suppresses the use of proximal sites. The fact that CFIm has an opposing poly(A) site preference to other cleavage factors and displays the most striking knockdown phenotype (Li et al., 2015) suggests that this complex may have a central regulatory role in APA. The evidence and implications surrounding the role of CFIm in APA will be fully discussed in Section 1.4.

Other RNA-associated factors

As well as components of the CPA machinery, other RBPs and RNA-associated factors have also been implicated in APA regulation (Table 1.1). One such factor is the nuclear poly(A) binding protein, PABPN1. The binding of PABPN1 to poly(A) tails necessarily occurs after CPA site choice has occurred. However, an RNA interference (RNAi) screen of RBPs coupled with a luciferase-based APA assay highlighted PABPN1 as an APA regulator, and 3' sequencing analysis revealed widespread 3'UTR shortening upon PABPN1 knockdown (de Klerk et al., 2012; Jenal et al., 2012). Identification of PABPN1 binding within 3'UTRs in these studies supported the concept that PABPN1 has an additional role independent of poly(A) tail length regulation. Mechanistic details, however, remain unclear - a CLIP study suggested PABPN1 binds at proximal sites and may competitively exclude cleavage factor binding (Jenal et al., 2012), but a more comprehensive RNA immunoprecipitation (RIP) experiment suggested preferential binding at distal sites and thus a possible stimulatory role in poly(A) site selection (de Klerk et al., 2012). The APA-associated role of PABPN1 has been further highlighted by study of the genetic disease oculopharyngeal muscular dystrophy (OPMD), which is caused by a triplet expansion in the *PABPN1* gene (tre*PABPN1*) that leads to PABPN1 protein aggregation and sequestration (Jenal et al., 2012). The resulting 'loss' of PABPN1 leads to pervasive changes in APA patterns, and mouse disease models have suggested that resulting changes in expression of key muscle function-associated genes may have a causative role in the disease, as well as age-related muscle wasting (Jenal et al., 2012; Riaz et al., 2016).

Another RBP strongly implicated in APA is the *D. melanogaster* neural-specific ELAV protein. ELAV was shown to be essential for promoting brain-specific 3'UTR extensions (see Section 1.2.4) through direct binding at and masking of proximal poly(A) sites (Hilgers et al., 2012). Intriguingly, this regulation was found to rely on recruitment of ELAV via specific promoter elements and paused pol II at these 'extended' genes (Oktaba et al., 2015). ELAV is clearly the central regulator of APA in this tissue-specific example, although there is as yet no evidence that the mammalian homologue ELAVL1/HuR plays a similar role in human brain-specific 3'UTR extensions.

Perhaps unsurprisingly given the links between splicing and 3' end processing, numerous splicing factors can also regulate APA. These include MBNL (Batra et al., 2014) and NOVA, a key APA regulator in mammalian brain (Licatalosi et al., 2008). U1 snRNP has a particularly well-defined role – through binding at 5' splice sites throughout transcripts, it suppresses premature CPA, for example at intronic sites near the start of genes (Kaida et al., 2010). Extending this observation, mild depletion of U1 snRNA was shown to cause more moderate upstream shifts in poly(A) site use, eliciting 3'UTR shortening – a process named 'telescripting' (Berg et al., 2012). Intriguingly, U1 depletion recapitulated the signature 3'UTR shortening observed upon neuronal activation, suggesting an elegant model whereby the transcriptional burst associated with neuronal activation (and perhaps other processes such as immune cell activation) creates a relative shortage of U1, thus eliciting widespread 3'UTR shortening (Berg et al., 2012).

Other RNA-associated factors implicated in APA regulation include THOC5, a member of the transcription/export (TREX) complex, which is thought to exert its effect via enhancing CFIm68 recruitment to genes (Katahira et al., 2013) and CPEB1, which via binding to cytoplasmic polyadenylation elements near proximal sites appeared to enhance CPSF recruitment to these sites (Bava et al., 2013). These examples further illustrate how factors not associated with CPA per se can influence poly(A) site selection, for example through interaction with CPA factors or masking of key sequence elements.

Transcription and chromatin context

As poly(A) site selection is co-transcriptional, the process of transcription itself can also exert an important influence on site choice. In particular, the speed of elongation has an important effect, as slower elongation will generally result in a greater lag between transcription of alternative poly(A) sites, thus increasing the probability of the proximal site being used. A study in *S. cerevisiae* demonstrated increased usage of proximal poly(A) sites upon mutation of elongation factors such as TFIIS, and it was suggested that this resulted from increased polymerase pausing (Cui and Denis,

2003). Use of a 'slow polymerase' mutant strain of *D. melanogaster* provided more direct evidence of this concept, showing increased proximal poly(A) site usage in the *polo* gene and others (Pinto et al., 2011). Moreover, depletion of human PAF1C, a complex known to regulate pol II elongation, was shown to increase pol II pausing and occupancy across the gene body, leading to increased proximal poly(A) site usage (Yang et al., 2016). The extent to which altered transcriptional elongation orchestrates changes in APA physiologically requires further investigation.

As well as elongation rate, greater transcriptional activity in general can impact upon APA. Specifically, more highly-expressed genes with stronger promoters were found to exhibit much greater use of proximal poly(A) sites in mouse and human cells (Ji et al., 2011). Two specific examples of this phenomenon, discussed earlier in the chapter, are the activity-dependent induction and resulting altered APA of MEF2 targets (Flavell et al., 2008) and the CR-APA-induced switch to the secreted IgM isoform, which is also associated with increased transcriptional activity (Martincic et al., 2009). A likely explanation for this finding is that transcription factors can promote recruitment of 3' end processing factors to the polymerase and enhance CPA, therefore enhancing the use of proximal poly(A) sites on a 'first-come, first-served' basis. The transcriptional co-activator Gal4-VP16, for example, has been shown to enhance CPA both in vitro and in vivo, via a process that requires PAF1C and the pol II CTD (Nagaike et al., 2011; Rosonina et al., 2003).

The chromatin context of poly(A) sites may also influence their usage. In general, poly(A) sites are nucleosome-depleted, likely owing to their AT-rich nature, but are often followed by a nucleosome-enriched region (Spies et al., 2009). Higher nucleosome density and the presence of heterochromatin-associated marks following poly(A) sites has been linked to increased usage, for example at the proximal poly(A) site of the *ETNK1* gene, which may be partly attributed to increased pol II pausing at the site (Neve et al., 2016; Spies et al., 2009). However, a study of APA in spermatogenesis suggested that the open chromatin mark H3K4me3 was associated with increased use of proximal poly(A) sites due to improved cleavage factor recruitment (Li et al., 2016). The interplay between

chromatin state and poly(A) site use is therefore likely to be relatively complex and gene-specific, and requires further dissection.

In summary, the choice between alternative poly(A) sites is clearly a complex process influenced by a range of factors, including the expression and recruitment of various RBPs and the level and rate of transcription. It remains of considerable interest to understand more fully the relative contribution and interplay of these factors and processes in APA regulation, and to identify the dominant factors involved in different physiological contexts.

1.3 Biological consequences of UTR-APA

As introduced in Section 1.2.2, alterations in the length and content of the 3'UTR resulting from UTR-APA may impact upon gene expression without changing the protein coding sequence. As this thesis focuses on the regulation of UTR-APA by CFIm, here the biological effects of UTR-APA are discussed in more detail.

1.3.1 Regulatory elements within 3'UTRs

The mRNA 3'UTR is defined as the section of sequence between the translation stop codon and the poly(A) tail, and often contains sequence elements able to regulate gene expression at the post-transcriptional level (Matoulkova et al., 2012).

A major class of regulatory element found predominantly in 3'UTRs is the microRNA (miRNA) target site. MiRNAs are single-stranded non-coding RNAs, around 22 nt in length, which can repress gene expression by binding to target mRNAs. Around 1,000 validated miRNAs have been identified in human cells (Friedlander et al., 2014), and they are able to target around 60% of genes (Friedman et

al., 2009). The minimum requirement for a miRNA to bind and regulate an mRNA is complementarity between a 6-8 nt 'seed' sequence near the 5' end of the miRNA and the target sequence in the 3'UTR (Lewis et al., 2003). This limited requirement for complementarity allows miRNAs to regulate multiple targets, and equally many 3'UTRs are targets for multiple miRNAs, or contain several binding sites for a single miRNA (Friedman et al., 2009; Hon and Zhang, 2007). Expression profiles of miRNAs vary substantially between different cell types and developmental stages, further increasing the complexity and regulation of post-transcriptional gene expression (Ludwig et al., 2016). To exert functional effects, miRNAs need to be incorporated into an RNA-induced silencing complex (RISC) containing an effector protein of the Argonaute (AGO) family. AGO and associated proteins can repress gene expression either by destabilising the transcript via deadenylation (Wu et al., 2006), inhibiting translation (Humphreys et al., 2005), or cleaving the transcript directly – this is mediated by AGO2, and generally requires greater miRNA:mRNA complementarity (Meister et al., 2004). In humans, initial work suggested that mRNA degradation is the major repression mechanism, accounting for >84% of decreased protein production (Guo et al., 2010). However, later studies have suggested that this may be an overestimate, and that generally translational repression is the dominant mechanism, preceding eventual decay (Bethune et al., 2012).

Another very common 3'UTR regulatory element is the AU-rich element (ARE), which was originally identified as a conserved sequence in cytokine transcripts able to target them for rapid degradation (Caput et al., 1986; Shaw and Kamen, 1986). AREs are typically defined as 3'UTR elements containing at least one copy of the AUUUA pentamer, and have now been identified in over 4,000 human genes of diverse functions (Bakheet et al., 2006). The function of AREs depends on the RBPs that bind them – most commonly, binding of repressive effector proteins such as AUF1 or tristetraprolin (TTP) promotes transcript degradation via deadenylation (DeMaria and Brewer, 1996; Lai et al., 2006), but binding of other proteins, such as HuR, can antagonize this activity and stabilise transcripts (Chen et al., 2002). Many other sequence elements may positively or negatively regulate gene expression via effects on mRNA stability or translation – these include GU-rich elements, CA-rich elements and iron responsive elements (Matoulkova et al., 2012).

Aside from influencing the amount of protein output, elements in the 3'UTR can regulate other aspects of mRNA fate, such as subcellular localisation, which can facilitate spatially restricted gene expression (Martin and Ephrussi, 2009). While there are not well-defined consensus sequences mediating particular localisation pathways, sequences important for specific transcripts have been identified. For example, an AC-rich 54 nt sequence in the *ACTB* mRNA was found to direct it to lamellipodia of motile fibroblasts (Lawrence and Singer, 1986), while an 11 nt sequence was sufficient for targeting of *MBP* mRNA to myelinating compartments of oligodendrocytes (Ainger et al., 1997).

1.3.2 Effects of UTR-APA at the single gene level

Through altering the length of the 3'UTR, UTR-APA can lead to the inclusion or exclusion of regulatory elements as discussed above, influencing various aspects of mRNA fate. As many 3'UTR elements have repressive effects, the prevailing view has been that APA-mediated shortening of 3'UTRs leads to general increases in gene expression (Figure 1.3), and this assumption has been found to hold for a number of genes studied, particularly in the context of disease.

One well-defined example of this phenomenon is in the cyclin D1 (*CCND1*) gene – a potent oncogene, which is characteristically overexpressed in mantle cell lymphoma (MCL) and associated with poor prognosis (Rosenwald et al., 2003). In a study of MCL cases, those with highest CCND1 expression showed truncated *CCND1* 3'UTRs with loss of destabilising AREs (Wiestner et al., 2007). Some of these truncations were generated by genomic deletions or mutations generating novel poly(A) sites, but for others no DNA sequence alteration was found, suggesting that altered APA may be involved. This is supported by a study which found greater use of the *CCND1* proximal poly(A) site in cancer cell lines, and demonstrated an increased mRNA half-life for the short 3'UTR isoform (Mayr and Bartel, 2009).

Similar findings have been reported for a number of other oncogenes, with luciferase assays showing increased expression from short vs. long 3'UTRs for *IMP-1*, *DICER1* and *CCND2*, which was partially attributable to loss of miRNA sites (Mayr and Bartel, 2009). Another study found prevalent APA-mediated 3'UTR shortening in the *NRAS* and *c-JUN* transcripts in triple-negative breast tumours, which was associated with increased expression of these oncogenes and increased tumour invasiveness (Miles et al., 2016). Luciferase assays suggested this may be attributable to loss of negative regulation by the RBP Pumilio. Increased gene expression through 3'UTR shortening is not limited to oncogenes and cancer development – for example, 3'UTR shortening of *HIP2* during T cell activation was associated with increased HIP2 expression, which was partially explained by loss of miRNA-mediated regulation (Sandberg et al., 2008).

The suggestion that 3'UTR shortening always leads to increased gene expression as seen in these examples, however, is an over-simplification. For other genes, 3'UTR shortening can down-regulate gene expression, as exemplified by the finding that the shorter isoform of *polo* in *D. melanogaster* is much less efficiently translated (Pinto et al., 2011). Interestingly, use of proximal poly(A) sites can also result in either shortening or lengthening of the poly(A) tail itself, with likely effects on mRNA decay and translation (Chang et al., 2014). Furthermore, it has been shown that miRNA sites preceding a proximal poly(A) site (i.e. present in both short and long 3'UTR isoforms) can become more active upon 3'UTR shortening due to more favourable positioning closer to the 3' end (Nam et al., 2014). Such increased potency of miRNA sites was identified in a cluster of anti-proliferative genes upon proliferation-associated shortening of their 3'UTRs, providing an elegant mechanism to down-regulate their expression (Hoffman et al., 2016). Similarly, the short 3'UTR isoform of the KGA isoform of glutaminase was found to be much more sensitive to repression by miRNAs miR-23a and miR-23b, while the longer isoform was subject to ARE-mediated decay (Masamha et al., 2016). In this case, the effect of altered UTR-APA and implications for cancer-promoting glutaminase expression would depend heavily on expression of effector miRNAs and RBPs, illustrating the sometimes complex relationship between alternative 3'UTRs and gene expression.

As well as these potential impacts on protein expression, UTR-APA-mediated exclusion or inclusion of localisation signals may have important effects. For example, the long 3'UTR isoform of brain-derived neurotrophic factor (*BDNF*) is required to direct its transport to neuronal dendrites, and loss of this isoform leads to less dendritic BDNF protein and resulting neuronal deficits (An et al., 2008). It has been postulated that the long 3'UTRs characteristic of neuronal tissue may be generally important in mediating transport along axons, though this is an area requiring further investigation (Hilgers et al., 2012). Remarkably, recent work has also revealed that alternative 3'UTRs may direct subcellular protein localisation independently of transcript localisation – a mechanism termed 3'UTR-dependent protein localisation (Berkovits and Mayr, 2015). It was shown that only the long 3'UTR isoform of *CD47* could be bound by HuR, which in turn locally recruited SET to the emerging CD47 protein during translation, ultimately mediating RAC1-mediated targeting to the plasma membrane and leading to increased protection of cells from phagocytosis. These observations were extended to other membrane proteins, hinting that APA may exert widespread effects on protein localisation and function (Berkovits and Mayr, 2015).

1.3.3 Global effects of UTR-APA

The potential for UTR-APA to exert significant effects on gene expression as discussed above raises the question of what proportion of transcripts are regulated in this way and what the physiological implications are. One prediction would be, for example, that widespread proliferation-associated 3'UTR shortening (see Section 1.2.4) might lead to a general increase in expression of affected genes, due to the bias of longer 3'UTRs towards negative regulation.

A handful of studies have analysed the effect of alternative 3'UTRs on gene expression genome-wide. Measurement of mRNA half-lives and translation rates of alternative 3'UTR isoforms in mouse 3T3 fibroblasts found a small but significant increase in stability (average 7.5% increase) and a small but significant decrease in translation rate (average 3% decrease) for shorter 3'UTR isoforms (Spies et al.,

2013). The small magnitude of these effects and clear correlation between half-lives and translation rates for short and long 3'UTR isoforms of the same gene led the authors to conclude that UTR-APA has a relatively limited effect on gene expression at the global level.

Another study looked directly at the effects of proliferation-associated 3'UTR shortening, measuring APA patterns in conjunction with global mRNA and protein abundance in resting and activated murine and human T cells (Gruber et al., 2014). Despite observing the expected global 3'UTR shortening in activated T cells, no global correlation was observed between the APA changes and mRNA or protein expression levels. Moreover, there was little overlap between the identities of transcripts showing 3'UTR shortening in mouse and human T cells, raising doubts about the biological significance of the APA events (Gruber et al., 2014). Similar observations were made in cell cycle entry from quiescence in fibroblasts and epithelial cells, where no overall effect of 3'UTR shortening was observed, though this study only considered mRNA abundance as a readout (Elkon et al., 2012).

Overall, these studies suggest that while APA has the potential to strongly influence gene expression at the single-gene level, overall effects of global APA modulation are relatively modest. A caveat to this is that some potentially important effects could be missed by these approaches – for example, impacts upon mRNA or protein localisation. These findings raise questions surrounding the biological significance of APA – for example, is 3'UTR shortening in proliferation important for regulating the expression of proliferation-associated genes, or just a secondary consequence of proliferation (Spies et al., 2013)? While one study found no enrichment of proliferation-associated genes in gene ontology analysis of those undergoing 3'UTR shortening in proliferation (Elkon et al., 2012), the findings relating to oncogenes (Mayr and Bartel, 2009; Wiestner et al., 2007) and tissue-specific 3'UTR shortening of tissue-specific genes (Lianoglou et al., 2013) suggest such events may be biologically important.

One possibility is that even where global APA changes exert only minor changes on gene expression transcriptome-wide, a small number of transcripts are affected in a physiologically important way. For

example, *CCND1* 3'UTR shortening in the context of widespread cancer-associated 3'UTR shortening could be sufficient to drive increased CCND1 expression that promotes enhanced tumour growth in MCL. Consistently, the observed growth-dependence of some glioblastomas on low CFIm25 expression level suggested that one or more of the 59 CFIm25-associated 3'UTR shortening events in these tumours was central to cancer development (Masamha et al., 2014). In the future, consideration of the effects of key single-gene events in the context of global APA modulation will be important to fully understand the physiological implications.

1.4 CFIm and its contribution to APA

As discussed in Section 1.1.3, the heterotetrameric cleavage factor I complex (CFIm) is one of the core mammalian cleavage factors, essential for pre-mRNA cleavage in vitro, and has also been strongly implicated in APA. As the function of CFIm in APA forms the central research topic of this thesis, the structure and function of the individual CFIm subunits is discussed in detail here, along with evidence relating to its role in APA and models for its function.

1.4.1 Structure and function of CFIm25

The smallest CFIm subunit has a molecular weight of 25 kDa, and is thus named CFIm25 (alternatively CPSF5 or NUDT21). The gene name is *NUDT21*, reflecting the fact that CFIm25 belongs to the superfamily of NUDIX hydrolases, which typically cleave nucleoside diphosphates linked to other moieties (McLennan, 2006). While the CFIm25 NUDIX domain was shown to bind ATP and diadenosine tetraphosphate (Ap_4A) in a crystal structure, it was found to lack hydrolase activity due to loss of the key catalytic residues and obstruction of the active site (Coseno et al., 2008).

CFIm25 is 227 amino acids long, with the NUDIX domain forming a large portion of the protein (Figure 1.2). As outlined in Section 1.1.3, CFIm25 is the main RNA-binding subunit of CFIm, with a preference for UGUAN sequences (Brown and Gilmartin, 2003; Dettwiler et al., 2004). Initial in vitro experiments suggested the region encompassing amino acids 1-160 was required for RNA binding (Dettwiler et al., 2004), and crystal structures of CFIm25 bound to a UGUA-containing oligonucleotide revealed the residues within this region responsible for sequence-specific recognition (Li et al., 2011; Yang et al., 2010). Specifically, CFIm25 F104 was shown to hydrogen bond with U1 of the RNA, E55 with G2 and R63 with U3, with further stabilisation and discrimination provided by stacking interactions of F103 with G2. Mutagenesis studies supported the contribution of these residues to UGUA recognition and, interestingly, RNA and AP_4A binding were found to be mutually exclusive, suggesting the potential for small molecule-mediated regulation of RNA binding (Li et al., 2011; Yang et al., 2010). While an important role of UGUA recognition in CFIm activity has been demonstrated (see Section 1.1.3), it remains unclear whether CFIm25 can also bind non-UGUA sequences, and how CFIm may be recruited in the absence of UGUA.

Substantial biochemical and structural evidence supports the ability of CFIm25 to dimerise (Coseno et al., 2008; Kim et al., 2010; Li et al., 2011; Yang et al., 2011a). The dimer interface is large, representing ~13% of the total surface, with around 20 residues participating in contacts between the two monomers (Coseno et al., 2008). Structures of the CFIm25 homodimer with bound UGUA-containing RNA revealed that each monomer could bind one 9mer oligonucleotide, with the sequences in anti-parallel orientation (Li et al., 2011; Yang et al., 2011a). This led to suggestions that CFIm25 may simultaneously bind two UGUA sequences on one pre-mRNA, perhaps associated with separate poly(A) sites, and loop out the intervening sequence (Yang et al., 2011b). This idea has been supported by data from electrophoretic mobility shift assays (EMSA) using a longer substrate containing two UGUA sequences, which also suggested a key role for CFIm68 in facilitating the looping (Yang et al., 2011a).

The CFIm25 homodimer interacts with two of the large CFIm subunits, CFIm68 or CFIm59, to form the functional heterotetramer (Kim et al., 2010). In vitro interaction studies suggested that the entire CFIm25 was required for interaction with CFIm68 (Dettwiler et al., 2004). Crystal structures of the CFIm25 homodimer in complex with RRMs from two CFIm68 molecules revealed that each CFIm68 RRM forms contacts with both CFIm25 subunits simultaneously (Li et al., 2011; Yang et al., 2011a). In both studies, tyrosines Y158 and Y160 on CFIm25 were shown to participate in the CFIm25-CFIm68 interaction, forming hydrogen bonds with CFIm68 via their hydroxyl groups, and their key role was supported by loss of the interaction in gel filtration analysis upon Y158A Y160A mutagenesis. H89 and F199 were shown to mediate interactions with the other CFIm68 molecule, though the importance of this interaction for complex formation remains unclear due to conflicting findings from mutagenesis studies (Li et al., 2011; Yang et al., 2011a).

As well as interacting with CFIm68/CFIm59, CFIm25 has defined interactions with PAP and PABPN1 (Dettwiler et al., 2004; Kim and Lee, 2001; Shimazu et al., 2007). A region within amino acids 81-160 was found to mediate both of these interactions (Dettwiler et al., 2004), though other studies suggested the interaction interface with PAP was closer to the N-terminus (Kim and Lee, 2001; Shimazu et al., 2007). These findings suggest that CFIm25 may have an important role in linking the cleavage and polyadenylation steps.

While the structural annotation of CFIm25 is relatively complete, all crystal structures lack the N-terminal amino acids 1-20, suggesting this region may be unstructured or highly mobile. Consistent with this, the subsequent amino acids, 21-32, form an extended loop that protrudes away from the structure (Coseno et al., 2008; Yang et al., 2010). The fact that K23 acetylation has been shown to modulate interaction with PAP (see Section 1.4.4) suggests that this region could represent an important protein interaction interface, and would benefit from further study.

1.4.2 Structure and function of CFIm68

CFIm68 (also CPSF6, gene name *CPSF6*) is one of the two 'large' CFIm subunits, with a molecular weight of 68 kDa. CFIm68 has a domain structure similar to that of SR proteins (see Section 1.5), and is classed as an 'SR-like' protein (Ruegsegger et al., 1998). The SR-like structure consists of an RRM and a C-terminal RS domain, with an additional proline-rich region (PRR) between these domains in CFIm68 (Figure 1.2). The RRM of CFIm68 does not seem to bind RNA, but was found to be sufficient for mediating interaction with CFIm25 (Dettwiler et al., 2004). Crystal structures coupled with mutagenesis identified the main residues involved, with residues 116-122 in the CFIm68 β_2/β_3 loop mediating interaction with CFIm25 Y158/Y160 in vitro and in vivo (Rasheedi et al., 2016; Yang et al., 2011a). Another study identified G86 and N87 in the RNP2 motif of the CFIm68 RRM as essential for the interaction in vitro (Dettwiler et al., 2004), though this is not supported by structural evidence.

Intriguingly, while the RRM of CFIm68 alone does not bind RNA, biochemical studies indicated that it strongly enhances RNA binding in complex with CFIm25 (Li et al., 2011; Yang et al., 2011a). Mutagenesis analysis showed that residues in the loops at the interface with CFIm25, rather than in the beta-sheets themselves (which are typically involved in RRM-mediated RNA binding), were responsible for this enhanced RNA binding (Yang et al., 2011a). This led to suggestion of a model whereby formation of entry and exit clefts at the CFIm68-CFIm25 interface may aid looping of RNA through the complex and stabilise binding (Yang et al., 2011b).

The C-terminal RS domain of CFIm68 is rich in positively charged arginine residues alternating with serines (potentially subject to phosphorylation) and negatively charged glutamate and aspartate residues. The general structure, function and phosphorylation of SR proteins is discussed further in Section 1.5. Biochemical experiments indicated that the CFIm68 RS domain could immunoprecipitate RNA in vitro (Dettwiler et al., 2004), though this may well reflect a non-specific ionic interaction and its relevance in vivo has not been explored. Interactions between the CFIm68 RS domain and various

other SR splicing factors (hTra-2ß, SRp20, and 9G8) have been identified, suggesting it may function as an important protein-protein interaction platform and promote links between splicing and 3' end processing (Dettwiler et al., 2004). In addition, microscopy studies with CFIm68 deletion mutants suggested that the RS domain was necessary and sufficient to achieve nuclear localisation of the protein, which may partially result from a putative nuclear localisation signal (NLS) in this domain (Dettwiler et al., 2004).

Analysis of the role of CFIm68 in human immunodeficiency virus (HIV) infection, where it is of interest due to its ability to bind viral capsid and direct HIV integration to active chromatin (Sowd et al., 2016), has provided further insight into RS domain-mediated subcellular localisation. A role for CFIm68 (referred to as CPSF6 in the field) in HIV infection was first identified through a screen for viral restriction factors, in which truncated CFIm68 lacking the RS domain accumulated in the cytoplasm and sequestered HIV there via capsid binding, preventing viral integration (Lee et al., 2010). Similar phenotypes were apparent upon depletion of the nuclear import receptor TNPO3 (De Iaco et al., 2013), suggesting a possible role for TNPO3 in CFIm68 nuclear import. A direct CFIm68-TNPO3 interaction was identified in vitro, and residues within an arginine-rich helix of TNPO3 were identified as essential for the interaction (Maertens et al., 2014). Abolishing the interaction through TNPO3 mutagenesis led to cytoplasmic accumulation of CFIm68, suggesting this represents a key nuclear import pathway for CFIm68. The study did not, however, define the residues of the CFIm68 RS domain involved or the importance of the putative NLS, other than to show that phosphorylation of the domain was not necessary for the interaction (Maertens et al., 2014).

Aside from its role in general nuclear import, the RS domain has also been implicated in subnuclear localisation of CFIm68. Fluorescence microscopy studies of both endogenous and GFP-tagged CFIm68 (and CFIm25) revealed a diffuse nuclear distribution with enriched foci (Cardinale et al., 2007; Dettwiler et al., 2004). Co-localisation analysis revealed that some foci overlapped partially with nuclear speckles, which are subnuclear structures enriched in splicing factors, while others corresponded to PSP1-containing paraspeckles – poorly understood features found adjacent to nuclear

speckles (Cardinale et al., 2007). The RS domain alone was found to target CFIm68 exclusively to nuclear speckles, while addition of the RRM recapitulated additional targeting to paraspeckles (Cardinale et al., 2007), which may reflect differing interaction partners of the RRM and RS domains. The importance of speckle localisation for CFIm function remains unexplored.

Heterokaryon experiments have demonstrated that, as well as being actively imported into the nucleus, CFIm68 undergoes nucleo-cytoplasmic shuttling (Ruepp et al., 2009). CFIm68 nuclear export may be mediated by interaction with the mRNA export factor NXF1/TAP, both directly via the CFIm68 RRM and indirectly via RS domain-mediated interaction with other SR proteins (Ruepp et al., 2009). This study also showed that CFIm68 tethering to mRNA enhanced its export, and that CFIm68 appears to act as a general mRNA export factor, probably via its interaction with NXF1/TAP.

1.4.3 Structure and function of CFIm59

The overall sequence and structure of the other large CFIm subunit, CFIm59 (or CPSF7, gene name *CPSF7*) is highly similar to that of CFIm68, suggesting a common origin via gene duplication. The domain structure, like CFIm68, incorporates an RRM, PRR and RS domain (Figure 1.2). A crystal structure of CFIm59 RRM-CFIm25 showed substantial similarity to CFIm68 RRM-CFIm25, and though not confirmed by mutagenesis, it appears that residues forming the interaction surface with CFIm25 are broadly similar (Yang et al., 2011b). Despite this general similarity, CFIm59 has important functional differences to CFIm68, most strikingly in its contribution to APA regulation (see Section 1.4.5).

One key difference noted in the crystal structure was the observation of a narrower ‘entry cleft’ and formation of an additional salt bridge between the CFIm59 RRM and CFIm25, which may be expected to preclude RNA threading (Yang et al., 2011b). This may suggest fundamental differences in the contribution of CFIm68 and CFIm59 to the RNA binding of CFIm, though a more detailed comparative analysis would be needed to probe this further. CFIm59 and CFIm68 also differ in their

protein-protein interactions – in particular, the RS domain of CFIm59 was not able to interact with the SR proteins identified as CFIm68 binding partners (Dettwiler et al., 2004). Conversely, a CFIm59-specific RS domain-mediated interaction with U2AF65, a splicing factor which binds intronic polypyrimidine tracts, was found to promote 3' end processing in vitro, and proposed to provide a crucial stimulatory link between terminal intron splicing and CPA (Millevoi et al., 2006).

These findings raise the possibility that differing protein-protein interactions of CFIm68 and CFIm59, particularly via their RS domains, provide a fundamental distinction between the function of CFIm59- and CFIm68- containing CFIm complexes.

1.4.4 Post-translational modification of CFIm

Many sites of post-translational modification (PTM), including phosphorylation, acetylation and methylation, have been identified on all three CFIm subunits, as summarised on the PhosphoSite Plus database online at phosphosite.org (Hornbeck et al., 2015). Many of these were identified in proteome-wide studies, with varying confidence, and the physiological relevance of most is unclear. However, many RBPs are known to be subject to PTM-mediated regulation (e.g. SR proteins – see Section 1.5), for example via alteration of protein-protein interactions or RNA binding (Lovci et al., 2016), and this may represent an important aspect of CFIm regulation.

Two PTM events on the CFIm complex have been studied in detail, the first of which is acetylation of lysine 23 (K23) on CFIm25. This PTM was identified by mass spectrometry (MS) following trichostatin A treatment and enrichment of acetylated proteins in COS-7 cells, and was found to be catalysed by CBP/p300, which was recruited and stimulated by interaction with CFIm68 (Shimazu et al., 2007). Furthermore, a potential functional role for this PTM was identified - K23 acetylation decreased the interaction of CFIm25 with PAP, suggesting that it may regulate disassembly of the 3' end processing complex (Shimazu et al., 2007).

Another well-studied PTM is methylation of three arginines, R202, R204 and R206, within the glycine/arginine-rich (GAR) motif of CFIm68 (which is absent from CFIm59). Both symmetric dimethylation by PRMT5 and asymmetric dimethylation by PRMT1 of these residues were identified in vitro (Martin et al., 2010), and methylation of R202 and R204 has since been confirmed in vivo (Guo et al., 2014). However, this methylation was found to be very stable, and no biological regulatory role has yet been identified (Martin et al., 2010).

No CFIm phosphorylation events have been well characterised, though some identified in proteome-wide studies have potential biological interest. For example, phosphorylation of tyrosine 40 (Y40) in CFIm25 has been identified in a number of cancer cell lines (Rush et al., 2005; Weber et al., 2012), suggesting potential relevance to proliferation-linked APA changes. Additionally, phosphorylation of serine 165 on CFIm68 (Matsuoka et al., 2007) may be of functional relevance, as phosphomimetic mutation at this site increased RNA binding affinity of the CFIm complex two-fold in vitro (Yang et al., 2011b).

Overall, while there are some hints of roles for PTMs in regulation of CFIm activity, this remains an underexplored area.

1.4.5 Effect of CFIm depletion on APA

As introduced in Section 1.2.5, CFIm is considered a central player in APA, as unlike other core cleavage factors, its depletion leads to widespread 3'UTR shortening through increased usage of proximal poly(A) sites. This phenomenon was first observed upon transient siRNA-mediated depletion of CFIm25 from HeLa cells, which led to increased proximal poly(A) site use in four genes analysed by northern blotting, including *TIMP2* and *SDC2* (Kubo et al., 2006). A subsequent study found that knockdown of CFIm68, though interestingly not CFIm59, had a very similar effect (Kim et al., 2010).

These observations were further generalised at the global level, with the demonstration that depletion of CFIm25 or CFIm68, but not CFIm59, from HEK293 cells, led to a transcriptome-wide increase in the use of proximal poly(A) sites (Gruber et al., 2012). These findings have since been validated by a number of groups and in a number of cell types, as summarised in Table 1.2.

Table 1.2: Summary of studies exploring the transcriptome-wide effect of CFIm depletion.

Cell type and method	Subunit	APA findings	Reference
siRNA in HEK293 cells A-seq	CFIm68	'Majority' of transcripts undergo 3'UTR shortening.	(Martin et al., 2012)
siRNA in HEK293 cells A-seq	CFIm25, CFIm59, CFIm68	'Widespread' 3'UTR shortening for CFIm25 and CFIm68, but not CFIm59.	(Gruber et al., 2012)
siRNA in HeLa cells Microarray	CFIm68	93% overlap with Thoc5 knockdown APA pattern, which showed 95% 3'UTR shortening.	(Katahira et al., 2013)
siRNA in HeLa cells RNA-seq with DaPars	CFIm25	1453 transcripts change; 99% 3'UTR shortening.	(Masamha et al., 2014)
siRNA in C2C12 myoblasts 3'READS sequencing	CFIm25, CFIm59, CFIm68	~1400-1500 transcripts change for CFIm25 and CFIm68, almost exclusively 3'UTR shortening. ~100 for CFIm59, no directional preference.	(Li et al., 2015)
CRISPR/Cas9 knockout in HEK293T cells RNA-seq with DaPars	CFIm68	12% of all transcripts change; 98.5% 3'UTR shortening	(Sowd et al., 2016)

The finding that CFIm25 and CFIm68 depletion have similar transcriptome-wide effects suggests that they may function together to favour the use of more distal poly(A) sites, and this is supported by the clear correlation (r=0.83) observed between 3'UTR changes upon CFIm25 and CFIm68 knockdown in C2C12 cells (Li et al., 2015). The observation that CFIm59 knockdown does not cause 3'UTR shortening was surprising considering its similarity to CFIm68, particularly in HeLa cells, where

CFIm59 is expressed at similar levels to CFIm68 (Kim et al., 2010). This suggests CFIm59-containing and CFIm68-containing complexes have fundamentally different roles, which may perhaps reflect different RNA-binding capabilities or interacting partners as discussed above (Section 1.4.3).

1.4.6 Models for regulation of APA by CFIm

The widespread and directional changes observed upon CFIm depletion suggest a general role for CFIm either in stimulating the use of distal poly(A) sites or in suppressing the use of proximal poly(A) sites, and there is some evidence supporting each of these mechanisms. A stimulatory role for CFIm is supported by its initial identification as a core factor necessary for in vitro cleavage (Takagaki et al., 1989). In addition, a combined study which mapped poly(A) site usage alongside CLIP analysis of cleavage factor binding positions found that for genes with APA sites, the position of the strongest binding peak of CFIm68 was often at the most distal site, and was highly predictive of the dominant poly(A) site used (Martin et al., 2012). Further evidence for a stimulatory role of CFIm at distal sites comes from the observed distribution of the preferred CFIm binding sequence, UGUA, around CFIm-regulated APA sites. In C2C12 myoblasts, transcripts showing 3'UTR shortening upon CFIm25/CFIm68 knockdown were significantly enriched in UGUA sequences in the -100 to -41 nt window relative to the distal PAS and significantly depleted of these sequences in the same window upstream of the proximal PAS (Li et al., 2015). This suggests that CFIm is likely to bind preferentially around distal poly(A) sites and presumably stimulates their usage, though this study did not directly analyse CFIm binding.

Evidence for a suppressive role of CFIm comes from single-transcript studies. One example is within the CFIm68 transcript itself, where a UGUA sequence overlaps the most proximal PAS (UGUAAUAAA); in vitro, binding of CFIm at this site decreased cleavage, possibly due to steric hindrance of CPSF recruitment to the PAS (Brown and Gilmartin, 2003). This may represent an autoregulatory mechanism, whereby increasing CFIm expression leads to altered 3' end processing of

CFIm transcripts. Another demonstration of a suppressive role for CFIm was in the glutaminase KGA isoform transcript, which has two poly(A) sites – use of the distal site yields a very unstable transcript, while use of the proximal site gives much higher expression (Masamha et al., 2016). A luciferase assay with either the short or long KGA 3'UTR surprisingly showed that both gave similarly low expression levels, which were relieved upon CFIm25 depletion (Masamha et al., 2016). This suggested that CFIm25 acts to repress cleavage at the proximal poly(A) site even in the absence of a distal site. These examples only cover two transcripts, and may not be representative – nonetheless, they highlight the potential of CFIm to suppress proximal poly(A) site usage.

Another model suggests that CFIm may stimulate CPA at distal sites and repress it at proximal sites simultaneously through RNA looping. Specifically, the finding that the CFIm heterotetramer can simultaneously bind two anti-parallel UGUA sequences prompted the suggestion that a single heterotetramer may simultaneously bind the proximal and distal sites, looping out the intervening sequence to obscure the proximal site (Yang et al., 2011a; Yang et al., 2011b) – see Figure 1.5B. These studies also suggest alternative situations, where CFIm may bind two UGUA sequences upstream of a single site to enhance its use (Figure 1.5A). These models, however, remain speculative.

It seems that a combination of stimulatory, suppressive and 'looping' modes of action may be involved in regulation of APA by CFIm. However, this is clearly an area requiring further investigation and clarification.

1.4.7 Physiological regulation and influence of CFIm

While the 3'UTR shortening phenotype seen upon CFIm depletion is striking, questions remain surrounding its physiological relevance. Specifically, do CFIm-induced changes in global APA patterns have important effects on gene expression and cell biology? Furthermore, does CFIm actually act as a physiological regulator of APA, for example in the shift from quiescence to proliferation?

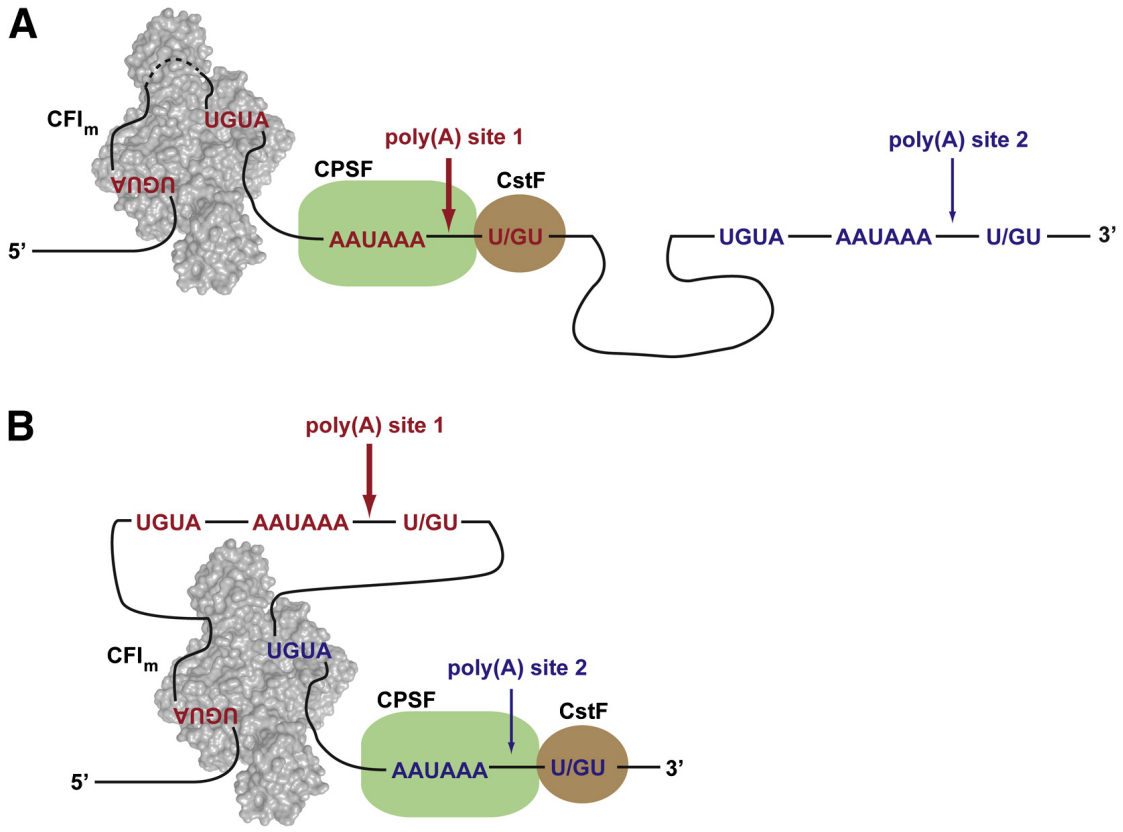


Figure 1.5: Models for potential roles of dual UGUA binding and RNA looping in regulation of APA by CFIm. (A) Here, CFIm simultaneously binds two UGUA sequences upstream of poly(A) site 1 to stimulate its use. (B) Here, CFIm simultaneously binds UGUA sequences upstream of poly(A) site 1 and poly(A) site 2, looping out and obscuring poly(A) site 1 while stimulating use of poly(A) site 2. Image reproduced with permission from: (Yang et al., (2011a)) under a RightsLink licence.

There are several examples, involving both natural and experimental systems, whereby changes in CFIm expression alter cell physiology via APA. At the simplest level, as discussed in Section 1.3.2, changes in poly(A) site use can alter protein output – for example, CFIm68 knockdown-induced shortening of the *RNF11* 3'UTR in HEK293 cells increased RNF11 protein expression (Martin et al., 2012). Such changes can have important physiological effects – for example, a small set of patients with neuropsychiatric disease have had the causative genetic change identified as copy number alterations in the *NUDT21* gene encoding CFIm25 (Gennarino et al., 2015). In these patients, CFIm25-dependent changes in poly(A) site choice led to over- or under-expression of the Rett syndrome-associated MECP2 protein and resulting neurological symptoms. Another example of a potential contribution of CFIm to neurological disease was found in a study of glioblastoma cell lines, many of which showed low CFIm25 expression levels (Masamha et al., 2014). Strikingly, restoration of normal CFIm25 levels in these cells attenuated their tumourigenic properties in a mouse model. This suggests that CFIm25 downregulation may have made a key contribution to oncogenic transformation in these cancers through APA-mediated changes in expression of oncogenes/tumour suppressor genes. The fact that these two examples of CFIm-mediated changes were both in brain disease suggests a particular sensitivity of neurons to CFIm expression levels, which may result from their unusually long 3'UTRs (see Section 1.2.4). Consistently, CFIm25 knockdown in rat hippocampal neurons was found to promote neurite outgrowth (Fukumitsu et al., 2012). While this was experimentally induced, this suggests potential for CFIm25-mediated regulation in neuronal development.

Potential roles for CFIm have also been observed outside of neuronal tissues. For example, a particularly high expression level of CFIm25 and CFIm68 has been observed in mouse male germ cells, where many spermatogenesis-associated transcripts use non-canonical poly(A) sites (Sartini et al., 2008). However, this was only a correlative observation, with no direct evidence that CFIm is needed for use of these sites. Furthermore, identification of a *CPSF6*-null mutant medaka fish embryo defective in primordial germ cell migration suggested a role for CFIm68-mediated APA in this aspect

of development (Sasado et al., 2017). The defect was partially attributed to 3'UTR shortening and increased expression of the SDF1A chemokine induced by the CFIm68 depletion.

While the above examples demonstrate the potential for CFIm-regulated APA changes to influence gene expression and physiology, they mostly relate to rare diseases or artificially induced CFIm depletion. The question remains as to whether, and in what situations, altered CFIm expression/activity acts as a master regulator of APA. The effects of decreased CFIm25 expression in glioblastoma (Masamha et al., 2014) suggest the potential for a general contribution of CFIm downregulation to proliferation-linked APA. However, aside from this study, evidence for decreased CFIm expression in proliferating cells is lacking, and in fact one study found slightly increased expression of CFIm25 and CFIm68 in proliferating compared to quiescent fibroblasts (Elkon et al., 2012). Another study found that genes undergoing 3'UTR shortening upon knockdown of CFIm68 showed significant overlap with those undergoing 3'UTR lengthening during C2C12 myoblast differentiation, suggesting CFIm may be a key driver of APA in this process (Li et al., 2015). However, counter-intuitively, CFIm68 was down-regulated in the differentiated cells, suggesting again that simple changes in CFIm expression cannot account for the observed APA changes.

Further studies are clearly needed in different cell types and states to understand the contribution of CFIm to physiological APA, with an added focus on other potential regulatory mechanisms aside from up-/down-regulation. For example, the possibility remains that in the absence of changes in CFIm expression, PTMs on the complex may modify its activity through altering interactions with other cleavage factors or RNA. Cell cycle-associated phosphorylation events would be of particular interest when considering possible links between CFIm activity and proliferation-associated APA.

This idea, that altered PTM profiles or other functional changes in CFIm may regulate its activity to bring about global alterations in APA, represents a central hypothesis of this study.

1.5 Arginine/serine-rich (RS) domains in SR and SR-like proteins

As introduced in Sections 1.4.2 and 1.4.3, CFIm68 and CFIm59 have a structure reminiscent of SR proteins, possessing both an RRM and a C-terminal arginine/serine-rich (RS) domain. The core SR protein family members are important regulators of both constitutive and alternative splicing, while 'SR-like' proteins such as CFIm68/CFIm59 have diverse roles in various aspects of RNA metabolism. The RS domain of CFIm68 is investigated in this thesis, with a particular focus on its potential contribution to APA regulation and its phosphorylation sites. To place these investigations in a wider context, here relevant current knowledge about the structure, function, dynamics and phosphorylation of SR proteins will be discussed.

1.5.1 Domain organisation and structure of SR proteins

The so-called SR proteins were first identified as splicing regulators in *D. melanogaster* genetic screens (Amrein et al., 1988). Subsequently, proteins with a similar arginine/serine-rich (RS) domain - ASF/SF2/SRSF1 and SC35/SRSF2 – were identified in human cells based on their ability to restore splicing activity to splicing-deficient nuclear extract (Fu and Maniatis, 1992; Krainer et al., 1990). Further members of the human SR protein family were identified on the basis of recognition by the mAb104 monoclonal antibody, which bound a phosphorylated epitope associated with sites of active pol II transcription (Zahler et al., 1992).

In the years since, numerous other RS domain-containing proteins have been identified, and the definition of what constitutes a 'core' SR protein has been tightened. The 12 bona fide SR proteins, now renamed SRSF1-12, all possess an N-terminal RRM (and in some cases a second RRM-homology domain) as well as an RS domain, and have known splicing-associated activity (Manley and Krainer, 2010). The RS domains within this group are defined as sequences of >50 amino acids in length with an R/S content of >40% and numerous RS dipeptides (Manley and Krainer, 2010).

However, they are highly variable in length, with SRSF4 having a 316 amino acid RS domain, for example. 'SR-like' or 'SR-related' proteins are those which possess as RS domain, but may lack the classical RRM (e.g. ZNF265, an alternative splicing regulator with zinc finger domains) or lack a known direct role in splicing (Long and Caceres, 2009).

Three-dimensional structural data for SR proteins is relatively limited, with challenges presented by their high degree of intrinsic disorder, particularly within the RS domain (Haynes and Iakoucheva, 2006). Most available structures are from in-solution nuclear magnetic resonance studies of SRSF1, and encompass the RRM only (Tintaru et al., 2007). There has, however, been some success at capturing conformations of the SRSF1 RS domain in complex with other proteins. A crystal structure of a small section of the RS domain bound to the nuclear import receptor TNPO3 showed it to be in an extended, relatively ordered conformation, with phosphorylated serines forming numerous salt bridges with an arginine-rich helix in the receptor (Maertens et al., 2014). In addition, a structure of ASF/SF2 bound to its kinase SRPK1 again captured a small section of the RS domain in an extended conformation, this time with basic arginines of the domain interacting with an acidic docking site on the kinase at the pre-phosphorylation stage (Ngo et al., 2008). These studies suggest that the disordered RS domain tends to adopt a more ordered, extended conformation upon interaction with binding partners, giving insight into its potential importance as a flexible protein-protein interaction platform. In addition, it has been suggested that the potential of disordered RS domains to contribute to hydrogel formation and phase separation may play an important role in nuclear signalling (Csizmok et al., 2016).

1.5.2 Functions and localisation of SR proteins

SRSF1 and SRSF2 were first identified as proteins essential for activating constitutive splicing (Fu and Maniatis, 1990; Krainer et al., 1990). By binding to exonic splicing enhancers (ESEs) via their RRM and serving as an interaction platform, they promote recruitment of U1 and U2 to the 5' and 3'

splice sites respectively (Long and Caceres, 2009; Wu and Maniatis, 1993), and may bridge interactions between the two sites (Fu and Maniatis, 1992). SR proteins are also well-known regulators of alternative splicing, which can enhance or suppress exon inclusion depending on a complex (and not fully defined) interplay of SR protein binding positions and interactions (Zhou and Fu, 2013). Binding of an SR protein in an exon generally promotes its inclusion, whereas binding of the same protein in the neighbouring intron suppresses inclusion of the exon (Erkelenz et al., 2013). As well as this major role in splicing regulation, SR proteins have been implicated in a diverse range of other RNA-associated processes, from transcription elongation to translation (Zhong et al., 2009).

RS domains have generally been considered the key mediators of protein-protein interactions involved in SR-mediated splicing regulation. In vitro co-IP experiments identified RS domain-dependent interactions of SRSF1 and SRSF2 with both U1-70K and U2AF35, and these were supported by yeast two-hybrid experiments (Wu and Maniatis, 1993). In addition, in vitro recruitment of U2AF35 to 3' splice sites by ESE-bound SR proteins was shown to depend upon the RS domain of U2AF35, suggesting RS-RS interactions are involved (Zuo and Maniatis, 1996).

These in vitro experiments have, however, not provided direct evidence that such RS-RS interactions actually occur in the spliceosome, and other studies have supported alternative roles for the RS domain. In particular, there have been convincing demonstrations of RS domain binding to RNA – specifically, an interaction between the RS domain of SRSF1 and the intronic branchpoint was shown to be essential for splicing, and this finding was extended to other RS-intron interactions (Shen and Green, 2004; Shen et al., 2004). In addition, another study found that the SRSF1-U1-70K interaction is actually mediated by interactions between the RRMs of the two proteins, and that the role of the SRSF1 RS domain is to regulate this interaction in a phosphorylation-dependent manner (Cho et al., 2011) – this is discussed further in Section 1.5.3. These findings challenge the traditional view of the RS domain as simply a platform for interactions with other SR proteins, though the different roles identified are not necessarily mutually exclusive.

The role of the RS domain as a protein-protein interaction platform has been better defined in relation to its contribution to nuclear import, and for some proteins, nucleo-cytoplasmic shuttling. Consistent with their role in splicing, SR proteins are predominantly nuclear, and their nuclear import was shown to rely on an interaction of the phosphorylated RS domain with the nuclear import receptor TNPO3 (Kataoka et al., 1999; Lai et al., 2000). Structural evidence and mutagenesis identified key interacting residues, highlighting salt bridges between phosphorylated RS serines and an arginine-rich helix in TNPO3, and the import mechanism was confirmed in vivo (Maertens et al., 2014). Some SR proteins undergo nucleo-cytoplasmic shuttling at a high dynamic rate, with this process seemingly dependent on the composition and phosphorylation state of the RS domain (Caceres et al., 1998). However, the RS domain alone was found to be insufficient to promote shuttling, with RNA binding by the RRM also playing a critical role (Caceres et al., 1998). Active export is likely explained by the RRM-mediated interaction of SR proteins with the nuclear export factor NXF1/TAP, which also allows them to act as an adaptor to enhance export of some mRNAs (Huang et al., 2003). Contrastingly, the RS domain of SRSF2 was found to contain a dominant proline-rich nuclear retention signal, which was sufficient to completely inhibit nucleo-cytoplasmic shuttling when transferred to other SR proteins (Cazalla et al., 2002). These observations highlight the important and diverse contributions of RS domains to determining the subcellular localisation of SR proteins.

RS domains have also been strongly implicated in SR protein localisation at the subnuclear level. SR proteins and other splicing-associated factors are concentrated in irregular, punctate nuclear foci, known as nuclear speckles or interchromatin granule clusters (Spector et al., 1991; Spector and Lamond, 2011). Evidence from Br-UTP pulse labelling experiments and the observation that nuclear speckles increase in size upon transcriptional inhibition suggests that these compartments are not sites of active transcription, and therefore are more likely to represent storage and assembly sites for splicing factors (Cmarko et al., 1999; Melcak et al., 2000). Fluorescent labelling experiments have supported this view, revealing recruitment of RS proteins from speckles to active transcription sites in a phosphorylation-dependent manner (Misteli et al., 1998; Misteli et al., 1997). The contribution of the RS domain to speckle localisation of SR proteins is variable – while ΔRS variants of SRSF1 and

SRSF5 still localised to nuclear speckles in an RRM-dependent manner, ΔRS variants of SRSF2 and SRSF3 completely lost targeting to speckles (Caceres et al., 1997). Such differences may reflect differential reliance on RS domain-mediated protein-protein interactions for recruitment to nuclear speckles.

1.5.3 RS domain phosphorylation

Extensive serine phosphorylation within RS domains was identified soon after the discovery of SR proteins, as initial antibodies raised against them recognised a phosphoepitope (Zahler et al., 1992). Phosphatase over-expression and inhibition experiments have identified a crucial role for this phosphorylation in splicing regulation. Specifically, SR protein phosphorylation is essential for spliceosome assembly, while subsequent de-phosphorylation is a pre-requisite for splicing catalysis (Mermoud et al., 1992; Mermoud et al., 1994). More recently, analysis of the SRSF1-U1-70K interaction has provided a possible mechanistic explanation for the phosphorylation requirement (Cho et al., 2011). In hypophosphorylated SRSF1, which can bind efficiently to ESEs, an intramolecular interaction between the RS domain and the RRM was identified. Phosphorylation of the RS domain was found to disrupt this interaction and unmask a U1-70K binding site on the SRSF1 RRM, facilitating U1-70K recruitment and spliceosome assembly (Cho et al., 2011).

Phosphorylation of the RS domain also has important roles in SR protein localisation and recycling. As discussed above, phosphorylation of SRSF1 is crucial for its nuclear import, though some SR proteins show phosphorylation-independent import (Maertens et al., 2014; Yun et al., 2003). In addition, phosphorylation promotes SR protein recruitment to active transcription sites, and nuclear re-phosphorylation may allow participation in repeated rounds of splicing (Misteli et al., 1998). Alternatively, following participation in the splicing reaction, de-phosphorylated SR proteins can remain associated with the mRNA and interact with NXF1/TAP to promote mRNA export (Lai and Tarn, 2004).

Clearly, the phosphorylation state of SR proteins has a considerable impact on their function, and needs to be tightly regulated. The kinases involved in RS domain phosphorylation belong to the SR protein-specific kinase (SRPK) and CDC-like kinase (CLK) families, while protein phosphatase 1 (PP1) and protein phosphatase 2 (PP2A) are thought to be the major RS domain phosphatases (Zhou and Fu, 2013). The ubiquitously expressed SRPK1 was the first human SRPK to be identified, based on its ability to promote cell cycle-dependent nuclear redistribution of splicing factors (Gui et al., 1994a; Gui et al., 1994b). This was followed by identification of SRPK2, which is mainly expressed in the nervous system (Wang et al., 1998), and SRPK3, which has not been extensively studied in humans but is muscle cell-specific in mice (Nakagawa et al., 2005). The SRPKs are kinases in the serine-threonine kinase family, with a characteristic two-lobed kinase domain (Giannakouros et al., 2011). They localise in both the cytoplasm and nucleus, and phosphorylate serines within RS dipeptides, with HSR/RSR representing preferred motifs (Wang et al., 1998). Biochemical and structural studies have suggested that SRPK1-mediated phosphorylation of SRSF1 is processive, with sequential phosphorylation of eight serines occurring as a result of sliding of the RS domain through a docking groove on the kinase in a C-to-N-terminal direction (Ngo et al., 2008).

The CLK family comprises four dual-specificity (serine-threonine/tyrosine) kinases (CLK1-4), with CLK1 (Clk/STY) the first to have been implicated in SR protein phosphorylation, again through its ability to promote redistribution of splicing factors (Colwill et al., 1996b). CLKs also phosphorylate only serines in RS domains, but have broader substrate specificity than SRPKs, efficiently phosphorylating SP and SK dipeptides as well as SR dipeptides (Colwill et al., 1996a). The other major difference between SRPKs and CLKs is that CLKs are exclusively localised to the nucleus, showing extensive co-localisation with nuclear speckle components (Colwill et al., 1996b; Duncan et al., 1998). Like SRPKs, CLKs demonstrate a processive phosphorylation mechanism (Velazquez-Dones et al., 2005).

The distinct localisation and substrate specificities suggest that SRPKs and CLKs may play different and possibly sequential roles in SR protein phosphorylation, and their relative contributions are yet to be fully defined. The mainly cytoplasmic localisation of SRPKs suggests they are responsible for the phosphorylation-dependent nuclear import of SR proteins, and this has been supported by the findings that deletion of SKY1, the only SRPK in *S. cerevisiae*, inhibited nuclear import of SRSF2, and over-expression of catalytically inactive SRPK2 led to cytoplasmic accumulation of human SRSF1 (Koizumi et al., 1999; Yeakley et al., 1999). In the nucleus, SRPK1 and SRPK2 are components of the spliceosome, and therefore may be involved in the phosphorylation events needed to promote spliceosome assembly (Mathew et al., 2008). It is not clear whether CLKs are also spliceosome components, but their main role may be to further fine-tune SR protein phosphorylation and activity in the nucleus. This is supported by the finding that SRPKs and CLKs both phosphorylate SRSF1, but at different target sites within the RS domain (Colwill et al., 1996a; Velazquez-Dones et al., 2005). Furthermore, structural studies highlighted a docking motif responsible for restricting SRPK1-mediated phosphorylation of SRSF1 to the N-terminal region of the RS domain, and further experiments demonstrated that subsequent phosphorylation of the C-terminal region of the RS domain by CLK1 led to release of SRSF1 from nuclear speckles (Ngo et al., 2005). This highlights the potential for SRPKs and CLKs to phosphorylate SR proteins differentially and sequentially in order to regulate their activity. The importance of CLKs is further underlined by the observation that small molecule CLK inhibitors can induce widespread changes in alternative splicing and cell growth (Araki et al., 2015).

1.5.4 CFIm68 and CFIm59 as SR-like proteins

CFIm68 and CFIm59 are considered ‘SR-like’ proteins, and were first classified as such based upon their reactivity with the 16H3 monoclonal antibody, which recognises several de-phosphorylated SR proteins (Neugebauer et al., 1995; Ruegsegger et al., 1998). Both proteins show the general domain structure of bona fide SR proteins, but do not show the splicing activity characteristic of the core group (Manley and Krainer, 2010).

CFIm68 and CFIm59 share several features with other SR proteins – they are able to interact with other SR proteins (Dettwiler et al., 2004; Millevoi et al., 2006), and CFIm68 is a component of the purified spliceosome (Zhou et al., 2002). In addition, CFIm68 has been shown to undergo nucleo-cytoplasmic shuttling (Ruepp et al., 2009), TNPO3-mediated nuclear import via the RS domain (Maertens et al., 2014) and partial targeting to nuclear speckles via the RS domain (Cardinale et al., 2007).

A key way in which CFIm68 differs from other SR proteins, aside from its implication in 3' end processing rather than splicing, is in the composition of its RS domain, which is particularly rich in RD and RE dipeptides as well as RS dipeptides. Aspartate and glutamate are negatively charged, phosphomimetic residues, and it could be speculated that functional mimicry of constitutive phosphorylation has been selected for in CFIm68 evolution. While RD and RE dipeptides are found in other SR proteins, they are much less frequent, and this may reflect an important functional or regulatory difference between CFIm68 and other SR proteins. The prevalence and possible function of phosphorylation in the CFIm68 RS domain also remains unclear – while CFIm68 appears to be a substrate for SRPK1 phosphorylation in vitro, in contrast to SRSF1, the phosphorylation does not appear necessary for its interaction with TNPO3 (Maertens et al., 2014). The possibility that reversible serine phosphorylation within the RS domain may regulate 3' end processing, analogously to regulation of splicing by phosphorylation of core SR proteins, remains an underexplored area.

1.6 Overview of the thesis

The aim of the work presented in the following chapters was to more fully characterise the contribution of CFIm to the regulation of APA in human cells, with a particular focus on the hypothesis that CFIm activity may be regulated by post-translational modification.

In Chapter 3, the development of a CFIm complementation assay is described. CRISPR/Cas9-mediated depletion of CFIm25 or CFIm68 yielded 3'UTR shortening in representative transcripts, which was reversible by CFIm25/CFIm68 re-expression.

In Chapter 4, this assay was exploited to investigate the functionality of a range of altered CFIm isoforms in order to gain insights into the requirements for efficient poly(A) site selection by CFIm in vivo. The importance of the CFIm25-CFIm68 interaction for the function of both subunits was confirmed, and the capacity of CFIm25 to recognise UGUA sequences on RNA was found to make a key contribution to poly(A) site selection at the transcriptome-wide level. In addition, CFIm68 mutants lacking portions of their C-terminal RS domain were unable to fully restore distal poly(A) site use despite retaining the ability to form a nuclear CFIm complex, highlighting a novel role for this domain in poly(A) site selection. Alongside complementation analysis, ChIP-seq profiling of pol II occupancy in CFIm25/CFIm68-depleted cells showed a surprising lack of changes at the 3' ends of genes, but did reveal an interesting reduction in pol II signal at transcription start sites, raising further questions about the mechanism of action of CFIm.

In Chapter 5, the potential impact of post-translational modification on CFIm activity was explored. MS and western blotting-based experiments identified a number of novel PTM sites on CFIm25 and CFIm68, and in particular highlighted extensive serine phosphorylation in the CFIm68 RS domain. Potential functions for these novel PTMs, as well as well-defined PTMs from public databases, were assessed using the complementation assay following modification-ablating or modification-mimicking mutagenesis. These investigations revealed that multi-site phosphomimetic mutagenesis in the CFIm68 RS domain leads to impaired distal poly(A) site selection, suggesting a potential role for RS domain phosphorylation in APA regulation.

Overall, the results presented in this thesis extend current understanding of the key mechanistic determinants of CFIm activity, and provide particular insights into the function of the CFIm68 RS domain.

Chapter 2
Materials and Methods

2.1 DNA methods

2.1.1 Determination of concentration

Concentration and purity of DNA was determined by measuring absorbance at 260 nm using a NanoDrop spectrophotometer, with an A_{260} of 1 representing 50 µg/ml DNA. An A_{260}/A_{280} greater than 1.8 was deemed to indicate acceptable DNA purity.

2.1.2 Genomic DNA extraction

Genomic DNA was extracted from human cells using the Wizard SV Genomic DNA Purification System (Promega). Briefly, cells were grown to 60-80% confluence in a 6-well plate and resuspended directly in >150 µl of lysis buffer. Lysates were applied to silica spin columns and the centrifugation-based protocol was used to bind the DNA and carry out ethanol-based washes. DNA was eluted in 250 µl dH_2O.

2.1.3 Taq PCR

For standard PCR reactions (colony PCR, northern probe synthesis, amplification of short sequences from cDNA, amplification of DNA for TA cloning), Simpler Red Taq polymerase (ThermoScientific) was used. Template or colony was mixed with 10X Taq buffer (1/10 dilution), dNTPs (200 µM each), $MgCl_2$ (2 mM), forward and reverse oligonucleotides (0.8 µM each) and Simpler Red Taq polymerase (40 U/ml), to a final volume of 50 µl.

Thermal cycling steps: 94°C for 2 min, 30-35 cycles of (94°C for 30 s, 55°C for 30 s, 72°C for 1 min per kb template), 72°C for 5 min, chilled to 4°C.

2.1.4 Phusion PCR

For PCR reactions requiring higher fidelity (e.g. amplification of longer sequences, such as CFIm68 coding region), Phusion High-Fidelity polymerase (New England Biolabs) was used. Template was mixed with 5x Phusion HF buffer (1/5 dilution), dNTPs (200 µM each), forward and reverse oligonucleotides (0.8 µM each) and Phusion High-Fidelity polymerase (20 U/ml), to a final volume of 50 µl.

Thermal cycling steps: 98°C for 30 s, 30-35 cycles of (98°C for 10 s, 55°C for 15 s, 72°C for 15 s), 72°C for 5 min, chilled to 4°C.

2.1.5 PCR-based site-directed mutagenesis

To generate amino acid point mutations and insertions/deletions in CFIm subunits, complementary oligonucleotides of 30-40 nucleotides in length containing the desired mutation were designed (Appendix 2) and used in a PCR reaction. The exception was the CFIm68 RS1 S-to-D mutant, for which a DNA fragment containing the desired mutations was purchased from VectorBuilder (Cyagen Biosciences). 10-15 ng of template plasmid (see Table 2.1 for vector list) was mixed with the oligonucleotides (0.25 µM each), dNTPs (200 µM each), Turbo Pfu polymerase (Agilent; 50 U/ml) and 10X Turbo Pfu buffer (1/10 dilution) in a volume of 50 µl.

Thermal cycling steps: 95°C for 1 min, 18 cycles of (95°C for 15 s, 55°C for 50 s, 68°C for 1 min per kb template), 68°C for 7 min, chilled to 4°C.
1 µl *Dpn*I enzyme was added to products and incubated at 37°C for 1 h to digest the parental methylated strand before transformation.

2.1.6 Quantitative PCR (qPCR)

For qPCR, 1 µl DNA sample (from pol II ChIP) was mixed with 5 µl SensiMix SYBR No-ROX mastermix (Bioline) and appropriate oligonucleotides (0.5 µM each) in MicroAmp optical tubes

(Applied Biosystems). Thermal cycling was carried out on the StepOnePlus Real-Time PCR System (Applied Biosystems), including a melt curve analysis after the PCR. Each sample was run in triplicate, and input dilutions were included to generate a standard curve for quantification.

Thermal cycling steps: 95°C for 10 min, 40 cycles of (95°C for 15 s, 60°C for 15 s, 72°C for 15 sec).

2.1.7 Generation of labelled northern probes

Taq PCR (see 2.1.3) was used to generate the templates for synthesis of labelled DNA probes for northern blotting. Oligonucleotides (Appendix 2) were designed to amplify ~400 bp of coding region in the gene of interest from a cDNA template. Products were gel extracted (see 2.1.9) and cloned into pCR2.1-TOPO (see Table 2.1) with TA cloning (see 2.1.11). Midiprep-purified preparations (see 2.2.4) were digested with *Eco*RI when necessary to re-excise the template for labelling, and the product was purified by gel extraction.

To generate radioactively labelled probes, the Ambion DECAPrime II Kit (ThermoScientific) was used. Briefly, 25 ng of gel-extracted template was diluted to ~2 ng/µl and random decamer oligonucleotides were annealed by boiling and snap freezing. A polymerase extension reaction with Exo$^-$ Klenow was then used to incorporate [α^{32}-P] dCTP (PerkinElmer, ~1.85 MBq added).

2.1.8 Agarose gel electrophoresis

1-2% agarose gels were prepared by dissolving agarose in 1X TBE buffer and Midori Green stain (Nippon Genetics) was added (1/20,000 dilution) before pouring to allow DNA visualisation. 0.2 volumes of 6X DNA loading dye were added to samples before loading, and a marker lane of HyperLadder I or II (Bioline) was included. Gels were run in 1X TBE buffer at 80V, and DNA was visualised with a UV transilluminator or blue light box.

1X TBE buffer: 89 mM Tris base, 89 mM boric acid, 2 mM EDTA pH 8.0.

6X DNA loading dye: 30% glycerol, 0.25% bromophenol blue or xylene cyanol (depending on DNA size).

2.1.9 DNA purification from PCR or gel extraction

To purify DNA from gel bands or PCR reactions, the Q-spin Gel Extraction/PCR Purification Kit (GeneFlow) was used. Briefly, bands were excised with a razor blade under blue light and dissolved in a binding buffer (or PCR products were mixed with this buffer), then the solution was applied to a silica spin column. Ethanol-based washes were carried out on the column and DNA was eluted in 25 µl dH_2O.

2.1.10 Restriction enzyme digests

Restriction digests were generally carried out in a 10-20 µl reaction comprising template (PCR product or ~2 µg plasmid), appropriate enzymes (10 U) (Roche/New England Biolabs), appropriate buffer (1X) and BSA where required (100 µg/ml). Digests were incubated at 37°C for 1-3 h.

2.1.11 Ligation (including TA cloning)

For ligation, ~50 ng digested vector was mixed with digested insert at a 1:3–1:5 molar ratio. 1-2 U T4 DNA ligase (Roche) and 10X T4 ligase buffer (1/10 dilution) were added, and reactions (final volume 10-20 µl) were incubated at 16°C overnight.

Some Taq PCR products were directly ligated into a linearised vector with T overhangs, without a restriction digest step, using a TA cloning kit (Invitrogen). The provided pCR2.1-TOPO plasmid (0.5 µl) was mixed with the Taq-generated insert (2.5 µl) and ligation was carried out as described above.

2.1.12 Generation of CFIm25/CFIm68 expression plasmids

For CFIm25 expression, the cDNA representing the CFIm25 coding sequence was amplified from the Basinger human primary fibroblast cDNA library (kindly provided by Hiroto Okayama) by Rupen

Hargreaves using *Hin*DIII-CFIm25 Fwd and CFIm25-*Bam*HI Rvs oligonucleotides (Appendix 2). Following *Hin*DIII and *Bam*HI digest, the fragment was ligated into pcDNA3.1+ (Table 2.1).

For CFIm68 expression, the cDNA representing the CFIm68 coding sequence was amplified from the Basinger human primary fibroblast cDNA library as described above by Dr. Marianne Christensen using *NotI*-CFIm68 Fwd and CFIm68-*XhoI* Rvs oligonucleotides. *Kpn*I/*Xho*I digest was ultimately used to excise the *CPSF6*-containing fragment from an intermediate plasmid and ligate it into pcDNA5/FRT (Table 2.1).

N-terminal GFP tags were added where required to CFIm68 isoforms in the pcDNA5/FRT plasmid by insertion of a GFP cDNA as a *Kpn*I fragment.

Table 2.1: Vectors used in this study

Vector	Use	Source
pSpCas9(BB)-2A-Puro (pX459)-*NUDT21* gRNA	CRISPR/Cas9 targeted disruption of the *NUDT21* gene	Dr Andrew Bassett
pSpCas9(BB)-2A-Puro (pX459)-*CPSF6* gRNA	CRISPR/Cas9 targeted disruption of the *CPSF6* gene	Dr Andrew Bassett
pCR2.1-TOPO	TA cloning, e.g. of northern probe sequences or gDNA loci for sequencing	TOPO TA cloning kit (Invitrogen)
pcDNA3.1+	Backbone for CFIm25 cloning/expression, negative control in CFIm25 transfections	Lab stocks
pcDNA3.1-CFIm25 (and mutant derivatives)	Transient expression of CFIm25 (WT/mutant)	Lab stocks (WT) This study (mutants)
pSUPERIOR.puro	Co-transfection with pcDNA3.1-CFIm25 to allow puromycin enrichment	Oligoengine
pcDNA5/FRT	Backbone for CFIm68 cloning/expression, integrated at FRT site in 68KD cells as negative control (68KD+empty)	Adam Volanakis (originally Invitrogen)
pcDNA5/FRT-CFIm68 (and mutant derivatives)	Stable expression of CFIm68 (WT/mutant) by integration at FRT site of 68KD cells	This study
pOG44	Expresses FLP recombinase - co-transfection with pcDNA5/FRT-derived plasmids to catalyse integration	Adam Volanakis (originally Invitrogen)

2.1.13 Sanger sequencing

Sanger sequencing of plasmids was performed using the BigDye Terminator v3.1 kit (Applied Biosystems). ~400 ng DNA was mixed with 1.5 µl 5X BigDye buffer, 3.2 pmol of the appropriate oligonucleotide and 1 µl BigDye mix, in a volume of 10 µl.

Thermocycler steps: 96°C for 4 min, 25 cycles of (96°C for 10 s, 50°C for 5 s, 60°C for 4 min), cooled to 4°C.

The reaction was terminated with 0.25 volumes of 125 mM EDTA and DNA was precipitated by adding 3 volumes chilled ethanol and centrifuging at 16,100 × g for 10 min. Ethanol was removed and the pellet was air-dried. Analysis was performed by Source BioScience, University of Oxford, and the data were analysed using ApE software.

2.2 Bacterial methods

2.2.1 Preparation of competent cells

The XL1-Blue strain of *E. coli* was used in bacterial transformations. To make competent cells, a polyethylene glycol (PEG)/dimethyl sulphoxide (DMSO)-based method was used. An XL1-Blue scrape was grown up in 5 ml of LB medium in a shaking 37°C incubator overnight, and the next morning 200 µl of the culture was diluted into 50 ml of SOB medium. The culture was incubated at 37°C, shaking, for around 3 h until an OD_{600} of 0.2-0.5 was reached.

The culture was cooled on ice for 30 min, then centrifuged at 2000 × *g* at 4°C for 10 min. The pellet was resuspended in 5 ml of chilled TSB buffer and incubated on ice for 10 min. The suspension was split into 100 µl aliquots, which were flash-frozen in liquid nitrogen and stored at -80°C.

LB medium: 1% tryptone, 0.5% yeast extract, 0.17 M NaCl. pH adjusted to 7 with NaOH. Sterilised.

SOB medium: 2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$. Sterilised.

TSB buffer: 10% PEG 6000, 5% DMSO, 20 mM $MgCl_2$, in LB.

2.2.2 Transformation of competent cells

For transformations, DNA (2-10 µl of ligations or mutagenesis reaction, 2 ng for re-transformations) was mixed with 20 µl 5X KCM and diluted up to 100 µl with H_2O. This solution was added to 100 µl of competent XL1-Blue cells and incubated on ice for 30 min. The bacteria were then heat shocked at 42°C for 1 min and briefly cooled on ice. 700 µl LB medium (see 2.2.1) was added, and the bacteria were incubated, shaking, at 37°C for 45 min.

Bacteria were then plated onto LB agar plates containing 100 µg/ml ampicillin, using sterile glass beads to coat the plate, before incubating overnight at 37°C to allow colony growth.

5X KCM: 0.5 M KCl, 0.15 M $CaCl_2$, 0.25 M $MgCl_2$. Sterilised.

2.2.3 Plasmid minipreparation ('miniprep')

To purify small quantities of plasmid (~3 µg) from transformed bacteria, a phenol/chloroform-based miniprep method was used. Colonies were picked into 2 ml of LB medium (see 2.2.1) with 100 µg/ml ampicillin and incubated overnight in a shaking 37°C incubator. 1.5 ml of the suspension was centrifuged at 5000 × *g* for 3.5 min, and the pellet was resuspended in 200 µl of miniprep solution. 200 µl of phenol/chloroform/isoamyl alcohol 25:24:1 ratio, water-saturated, pH 8.0 (Fisher Scientific) was added and mixed by vortexing.

Following centrifugation at 16,100 × g for 2 min, the top (aqueous) layer was transferred to a new tube. 400 µl cold 100% ethanol was added and mixed by vortexing. Precipitated DNA was pelleted by

centrifugation at 16,100 × g for 5 min, and the pellet was washed by replacement of the ethanol with 70% ethanol and further centrifugation at 16,100 × g for 3 min. After removing ethanol, the pellet was air-dried, then resuspended in 10 µl of 10 mM Tris containing 100 µg/ml RNase A.

Miniprep solution: 50 mM Tris pH 8.0, 2.5 M LiCl, 62.5 mM EDTA, 4% Triton X-100.

2.2.4 Plasmid midipreparation ('midiprep')

To purify larger quantities of plasmid DNA (~300 µg), the GenElute HP Plasmid Midiprep kit (Sigma-Aldrich) was used. 50 µl of a starter culture was diluted into 50 ml LB medium (see 2.2.1) with 100 µg/ml ampicillin and grown up in a shaking 37°C incubator overnight. Bacteria were pelleted by centrifugation at 5000 × *g* for 10 minutes and the centrifugation-based midiprep method was used according to instructions, to obtain 1 ml of plasmid DNA at ~200-400 ng/µl.

2.3 RNA methods

2.3.1 RNA extraction and quantification

Total RNA was extracted by resuspension of cell pellets in TRI reagent (Sigma-Aldrich) followed by bromochloropropane extraction and isopropanol precipitation according to manufacturer's instructions. The RNA pellet was washed with 80% ethanol and resuspended in MilliQ water. Concentration and purity was determined by measuring absorbance at 260 nm using a NanoDrop spectrophotometer, with an A_{260} of 1 representing 40 µg/ml RNA. An A_{260}/A_{280} greater than 1.9 was deemed to indicate acceptable RNA purity.

2.3.2 RNA agarose gel electrophoresis

15–20 µg of RNA was mixed with 2.7 volumes of RNA loading buffer and denatured by heating at 65°C for 10 min. After cooling on ice, samples were loaded and separated on an agarose-formaldehyde gel. The gel was run at 5 V/cm (50 V) for 4 h in 1X MOPS buffer. The RNA was

visualised using a UV transilluminator to confirm RNA integrity by observation of the 18S and 28S rRNA bands.

RNA loading buffer: 50% formamide, 6% formaldehyde, 1X MOPS buffer, 10% glycerol, 20 µg/ml ethidium bromide, several drops bromophenol blue.

1X MOPS buffer: 20 mM MOPS, 5 mM sodium acetate, 1 mM EDTA, pH adjusted to 7 with NaOH.

Agarose-formaldehyde gel: 1.2% agarose, 0.66 M formaldehyde, 1X MOPS buffer

2.3.3 Northern blotting

Transfer to nitrocellulose membrane

Following electrophoresis, the gel was rinsed in water for 10 min and twice in 10X SSC for 10 min. RNA was transferred to the membrane using capillary transfer. The gel was placed upside down on a Whatman 3MM paper wick soaking in 10X SSC, and a piece of Hybond-N+ membrane (GE Healthcare) was placed on top of the gel. Two further pieces of Whatman 3MM paper soaked in 10X SSC were placed on top of the membrane, followed by a stack of paper towels and a weight. Following overnight transfer, the RNA was fixed to the membrane by UV crosslinking using a UVC 500 Crosslinker (Hoefer) at 700 mJ/cm^2. Membranes were stored in Whatman 3MM paper at room temperature for later use.

Blocking and probe hybridisation

Membranes were blocked and prepared for hybridisation by placing in a hybridisation bottle with 10 ml of hybridisation buffer and incubating in a rotating hybridisation oven at 42°C for 2 h. The labelled probe was synthesised as described in Section 2.1.7, then denatured by heating at 95°C for 5 min. After cooling on ice, the probe was added directly into the hybridisation buffer and left to hybridise overnight at 42°C.

Washes and signal visualisation/analysis

The probe-containing buffer was poured away and the membrane was washed in wash buffer 1 for 15 min at 42°C and 15 min at 65°C, then twice in wash buffer 2 at 65°C (~50 ml each wash). The washed membrane was wrapped in Saran wrap and exposed to a Phosphor screen for 10–50 min, depending on the probe. The screen was scanned using a FLA5000 Image Reader (Fuji). Where necessary, membranes were stripped by incubating in boiling 1% sodium dodecyl sulphate (SDS) for 30 min and rinsed in 2X SSC before re-blocking and re-probing.

Data analysis

Images were analysed using AIDA Image Analyzer (Raytest), using the 2D densitometry tool with subtraction of a single background reference square (not lane-specific). Proximal/distal (P/D) band ratios were calculated by dividing the density of the lower-most band (shortest transcript) by the sum of the other, upper band densities (longer transcripts). Statistical analysis of differences in P/D ratios between samples was performed using an unpaired, two-tailed Student's t-test. Asterisks on graphs indicate the following cut-offs: * $p<0.05$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$.

10X SSC: 1.5 M NaCl, 0.15M sodium citrate, adjusted to pH 7 with HCl.

Hybridisation buffer: 5X SSC, 50% formamide, 1% SDS, 1X Denhardt's solution (0.02% BSA, 0.02% Ficoll, 0.02% Polyvinylpyrrolidone), 10% dextran sulphate, 100 µg/ml boiled salmon sperm DNA.

Wash buffer 1: 1X SSC, 0.1% SDS

Wash buffer 2: 0.1X SSC, 0.1% SDS

2.4 Protein methods

2.4.1 Preparation of protein lysates

Harvested cell pellets were resuspended in an appropriate volume of NET-N150 buffer (with additives) depending on the pellet size. Lysates were cleared by centrifugation at 16,100 × g for 18

min at 4°C. Protein concentration was measured using a Bradford-based assay. Protein assay dye reagent concentrate (Bio-Rad) was diluted 1/5 and 1 ml was mixed with 1 µl of lysate. After 3 min incubation, absorbance at 595 nm was measured with a spectrophotometer. Values were compared to measurements from a bovine serum albumin (BSA) standard curve to estimate total protein concentration.

NET-N150 buffer: 50 mM Tris-HCl pH 7.8, 150 mM NaCl, 0.5% IgePal.

Added fresh: 1X cOmplete EDTA-free protease inhibitors (Roche), 10 mM benzamidine.

Added fresh, for immunoprecipitations only: 2.5 mM $MgCl_2$, 125 U/ml benzonase (Sigma-Aldrich).

Added fresh, in phosphorylation detection experiments only: 1X PhosSTOP phosphatase inhibitor (Roche).

2.4.2 Immunoprecipitation

For CFIm immunoprecipitation (IP), lysates were prepared as described above. Generally, 800 µg lysate was used per IP at a concentration of 1–1.5 mg/ml. After removing an input sample, an appropriate amount of primary antibody was added (see Table 2.2) and samples were mixed on a rotating wheel for 2 h at 4°C.

Protein G-Sepharose fast flow beads (Sigma-Aldrich) were prepared (35 µl per sample) by washing 3 times in lysis buffer (centrifugation at 500 × *g*, 2.5 min) and blocking. Beads were blocked in 3 mg/ml BSA on a rotating wheel for 1 h at 4°C, followed by removal of BSA and two lysis buffer washes.

Lysate/antibody mix was added to the prepared beads and mixed on a rotating wheel for a further 2 h at 4°C. After removing supernatant, beads were washed 3 times in 700 µl NET-2 wash buffer – twice

by inversion of the tube, and once with a 10 min incubation on the rotating wheel (recovering beads by centrifugation at 500 × *g*, 2.5 min each time).

NET-2 wash buffer: 50 mM Tris-HCl pH 7.8, 150 mM NaCl, 0.05% IgePal.
Added fresh: 1X cOmplete EDTA-free protease inhibitors (Roche), 10 mM benzamidine, 1 mM $MgCl_2$, 125 U/ml benzonase (Sigma-Aldrich).
Added fresh, in phosphorylation detection experiments only: 1X PhosSTOP phosphatase inhibitor (Roche).

2.4.3 SDS-polyacrylamide gel electrophoresis (PAGE)

Whole cell protein lysates (25–50 µg) or IP beads were mixed with 4X NuPAGE LDS Sample Buffer (ThermoFisher) (10 µl for lysates, 30 µl for beads). Samples were heated for 5 min at 95°C and loaded into wells of a freshly made 11% polyacrylamide gel. 3.5 µl of BLUeye prestained protein ladder (Geneflow) was run alongside samples. Gels were run at 100–130 V for ~90 min in SDS running buffer, until the dye front reached the bottom.

For proteomics experiments only, pre-cast NuPAGE Novex 4-12% Bis-Tris polyacrylamide gels and NuPAGE MES SDS running buffer (ThermoFisher) were used instead, to minimise keratin contamination.

Polyacrylamide gel: 75 mm x 85 mm x 1.5 mm gel. Separating gel (7 ml) – 11% acrylamide:bis-acrylamide (37.5:1 solution, Geneflow), 375 mM Tris pH 8.8, 0.1% SDS, 0.1% ammonium persulphate, 0.16% TEMED. Stacking gel (2.4 ml) – 5% acrylamide:bis-acrylamide (37.5:1 solution, Geneflow), 125 mM Tris pH 6.8, 0.1% SDS, 0.05% ammonium persulphate, 0.1% TEMED.
SDS running buffer: 25 mM Tris base, 192 mM glycine, 0.1% SDS

2.4.4 Coomassie staining

To stain protein bands in preparation for mass spectrometry, gels were incubated in the Coomassie-based InstantBlue Protein Stain (Expedeon) for 1 h at room temperature. Stained gels were imaged by scanning using the LI-COR Odyssey SA scanner in the 700 nm channel.

2.4.5 Western blotting

For western blotting, proteins were transferred from the gel to an Amersham Protran 0.45µm nitrocellulose membrane (GE Healthcare) by wet transfer in 1X transfer buffer using the Mini Trans-Blot Cell (BioRad). Transfer was performed at 100 V for 1 h.

Membranes were stained with Ponceau S solution to confirm successful and even transfer. Blocking was performed at room temperature for 30 min or overnight at 4°C. 5% milk powder in TBS-T was used for standard western blots. For experiments involving phosphorylated protein detection, 5% BSA in TBS-T or pIMAGO blocking buffer (for pIMAGO experiments) was used.

Primary antibodies (see Table 2.2) were diluted in blocking buffer and incubated with the membrane either overnight at 4°C or for 2 h at room temperature. After washing with TBS-T (3x 5 min washes), LI-COR secondary antibodies (see Table 2.2) were diluted in blocking buffer and incubated with the membrane for 1 h at room temperature. The membrane was washed again with TBS-T (3x 5 min washes) and scanned using the LI-COR Odyssey SA fluorescence scanner, detecting signal in the 700 nm or 800 nm channel (or both) as appropriate. Band intensities were quantified using the Odyssey SA software, by manually drawing lanes/bands and including a lane-specific background subtraction method.

Transfer buffer: 25 mM Tris base, 192 mM glycine, 20% ethanol

TBS-T: 25 mM Tris base, 150 mM NaCl, 2 mM KCl, pH adjusted to 7.8 with HCl. 0.1% Tween 20 added immediately before use.

Table 2.2: Antibodies used in this study.

Target	Type	Source	Application/dilution
CFIm25	Mouse monoclonal	Santa Cruz NUDT21 (2203C3): sc-81109	Immunoprecipitation, 2.5 µg per mg lysate Western blot, 1/1000
CFIm25	Rabbit polyclonal	Gift from Georges Martin, University of Basel	Western blot, 1/2000
CFIm25	Rabbit polyclonal	Proteintech NUDT21 (10322-1-AP)	Western blot, 1/1000
CFIm59	Rabbit polyclonal	Bethyl Laboratories Anti-CPSF59 (A301-359A)	Western blot, 1/1000
CFIm68	Rabbit polyclonal	Bethyl Laboratories Anti-CPSF68 (A301-356A)	Western blot, 1/1000
CFIm68	Rabbit polyclonal	Bethyl Laboratories Anti-CPSF68 (A301-358A)	Immunoprecipitation, 4 µg per mg lysate
GFP	Mouse monoclonal	Santa Cruz GFP (B-2): sc-9996	Immunoprecipitation (negative control), 2.5 µg per mg lysate
GFP	Rabbit polyclonal	Sigma-Aldrich Anti-GFP, N-terminal (G1544)	Western blot, 1/10,000
GAPDH	Rabbit polyclonal	Sigma-Aldrich Anti-GAPDH (G9545)	Western blot, 1/5000
α-tubulin	Mouse monoclonal	Sigma-Aldrich Anti-α-Tubulin, clone DM1A (T9026)	Western blot, 1/10,000
α-tubulin	Mouse monoclonal	TAT-1 – gift from Keith Gull, University of Oxford	Western blot, 1/10,000
Pol II RPB1	Rabbit polyclonal	Santa Cruz Pol II (N-20X): sc-899 X	Western blot, 1/1000 ChIP, 2 µg (for 80 µg chromatin)
None (isotype control)	Rabbit polyclonal	Santa Cruz Normal rabbit IgG: sc-2027	Immunoprecipitation (negative control), 4 µg per mg lysate ChIP, 2 µg (for 80 µg chromatin)
Phospho-SR proteins	Mouse monoclonal	Millipore Anti-phosphoepitope SR, clone 1H4 (MABE50)	Western blot, 1/1000
Rabbit IgG	Goat polyclonal	LI-COR IRDye 680 conjugated anti-rabbit (926-32221)	Fluorescent secondary antibody for western blot LI-COR detection, 1/10,000
Mouse IgG	Goat polyclonal	LI-COR IRDye 800CW conjugated anti-rabbit (926-32210)	Fluorescent secondary antibody for western blot LI-COR detection, 1/10,000

2.4.6 pIMAGO detection

Following CFIm IP and transfer of proteins to the membrane, the pIMAGO-Biotin Fluor 800 kit (Tymora Analytical) was used to detect phosphorylated proteins. Blocking was performed with pIMAGO blocking buffer overnight at 4°C, and detection using biotinylated pIMAGO and avidin-fluor 800 was carried out as specified. Signal was detected using the LI-COR Odyssey SA scanner as described above.

2.4.7 Mass spectrometry (MS) analysis of PTMs

For mass spectrometry experiments, CFIm protein bands were excised from the gel with a scalpel and diced into 1 mm cubes, which were stored at 4°C.

Sample preparation

Diced bands were de-stained in 50% acetonitrile diluted in 50 mM tetraethylammonium bromide (TEAB) by incubation, shaking at 37°C, for 15-20 min, then washed in 100% acetonitrile until fully dehydrated. Bands were incubated in 100 µl tris(2-carboxyethyl)phosphine (TCEP) for 30 min, shaking at 37°C, to reduce disulphide bonds, followed by two washes with 50 mM TEAB to remove the TCEP. Bands were then incubated in 100 µl 2-chloroacetamide at RT in the dark for 1 h to alkylate cysteines, followed by two washes with 100% acetonitrile.

For protein digestion, 200 ng sequencing grade modified trypsin (Promega), diluted in 50 mM TEAB, was added to completely cover the bands and incubated overnight at 37°C, shaking. The following day, 1 µl trifluoroacetic acid was added to stop the digest and the supernatant was removed to a new tube. 20 µl of 10% formic acid was added to the gel pieces and incubated, shaking, at RT for 20 min to increase peptide extraction, and the supernatants were pooled.

Data collection and analysis

The sample was analysed by the University of Oxford Proteomics Facility with liquid chromatography tandem-MS (LC-MS/MS) using the Q-Exactive Hybrid Quadrupole-Orbitrap Mass Spectrometer (ThermoScientific). Spectra were processed and analysed using the Mascot Server (Matrix Science), with the following search parameters: fixed carbamidomethyl modification, variable oxidation (M), acetyl (protein N-term), acetyl (K), phospho (ST) and phospho (Y) modifications, peptide mass tolerance ± 10 ppm, fragment mass tolerance ± 0.05 Da, maximum of 2 missed cleavages.

2.4.8 Chromatin immunoprecipitation (ChIP)

For ChIP experiments, cells (two 15 cm dishes at 80% confluence for each line) were cross-linked at room temperature with 1% formaldehyde and quenched with 125 mM glycine for 5 min. Following two PBS rinses, cells were harvested by scraping into ice-cold PBS, then resuspended in ChIP lysis buffer and incubated (10 min) on ice. After centrifugation (5 min, 1500 × *g*, 4°C), nuclear pellets were washed in ChIP wash buffer and resuspended in ChIP sonication buffer. Cells were sonicated twice (15 min, high amplitude, 30 s ON/30 s OFF) using a Bioruptor (Diagenode). Extracts were cleared by centrifugation (16,100 × *g*, 15 min, 4°C). 80 µg of chromatin was incubated overnight at 4°C with 2 µg of antibody against IgG (IP negative control) or pol II. Immunocomplexes were immobilized for 1 h at 4°C by adding to 15 µl of BSA-saturated protein G Dynabeads (ThermoFisher). IgG supernatant was retained as total input.

After extensive, stringent washes, DNA was eluted in a total volume of 100 µl (15 min, 25°C, shaking). For inputs, 90 µl elution buffer was added to 10 µl total input. Following RNase A treatment (30 min, 37°C), protein-DNA crosslinks were reversed by incubation at 65°C for 5 h. After ethanol precipitation, proteins were degraded with proteinase K (2 h, 45°C). DNA was purified using a PCR Purification Kit (Qiagen), with elution in a total volume of 15 µl nuclease-free H_2O.

ChIP lysis buffer: 10 mM Tris-HCl pH 8.0, 0.25% Triton X-100, 10 mM EDTA, 1X cOmplete EDTA-free protease inhibitors (Roche)

ChIP wash buffer: 10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 1X cOmplete EDTA-free protease inhibitors (Roche)

ChIP sonication buffer: 10 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA, 1X cOmplete EDTA-free protease inhibitors (Roche)

Elution buffer: 100 mM $NaHCO_3$, 1% SDS, 10 mM DTT

2.5 Mammalian cell culture and analysis methods

2.5.1 Cell lines and routine maintenance

Starting cell lines used in this project were Human Embryonic Kidney (HEK) 293T, HEK 293 Flp-In T-REx (gifts from Fumiko Esashi) and BJ-5ta hTert-immortalised human fibroblasts (purchased from ATCC, catalogue number CRL-4001).

Cells were cultured in 37°C incubators with 5% CO_2, in Dulbecco's Modified Eagle Medium (DMEM, Sigma-Aldrich) supplemented with 10% Fetal Bovine Serum (FBS, Gibco) and 100 units/ml penicillin, 100 µg/ml streptomycin (PenStrep, Gibco). Additional antibiotics were added to the medium to maintain expression of resistance cassettes as follows: 10 µg/ml hygromycin B for BJ-5ta fibroblasts; 100 µg/ml zeocin and 15 µg/ml blasticidin for HEK 293 Flp-In T-REx parent and CRISPR-targeted lines; 150 µg/ml hygromycin B and 15 µg/ml blasticidin for HEK293 Flp-In T-REx stable lines with expression cassette integrated at FRT site.

Cells were passaged using TrypLE Express dissociation reagent (ThermoFisher) following a PBS rinse. Cells were counted using Glasstic slides with grid (KOVA) after diluting 5-fold in 0.4% trypan blue solution (Sigma-Aldrich).

2.5.2 Cell harvesting and cryopreservation

Cells were harvested in a cold room by scraping into ice-cold PBS following an initial PBS wash. Cells were pelleted by centrifugation at 1000 × g for 4 min at 4°C and either stored as pellets at -80°C, or resuspended in TRI reagent before storing at -80°C.

For cryopreservation, cells were harvested by TrypLE detachment and centrifugation at 1000 × g for 4 min. The pellet was resuspended in 1 ml FBS containing 10% DMSO, and a CoolCell container (Biocision) was used to slowly cool the suspension to -80°C. Cells were either kept at -80°C, or in liquid nitrogen for longer-term storage.

2.5.3 Transfection

Transfections were performed using jetPRIME reagent (Polyplus-transfection) according to the manufacturer's instructions. Cells were seeded the day before transfection and transfected at 60–90% confluence. Plasmid DNA was purified by midiprep (see 2.2.4). DNA (2 µg per well in a 6-well plate or 4 µg per 6 cm dish) was mixed with 200 µl jetPRIME buffer and jetPRIME reagent at a ratio of 1 µg DNA:2 µl jetPRIME. After a 10 min incubation, the suspension was added drop-wise to the cell growth medium.

2.5.4 CRISPR/Cas9-mediated gene disruption

For CRISPR/Cas9 editing of CFIm genes, single-guide RNA (sgRNA) sequences targeting *NUDT21* and *CPSF6* (see Figure 3.1) were designed, validated by high resolution melt analysis (HRMA) and cloned into the pSpCas9(BB)-2A-Puro (PX459) plasmid by Dr Andrew Bassett, Genome Engineering Oxford.

The plasmids were transfected into cells (see 2.5.3) in a 6-well plate. 24 h later, 2.5 µg/ml puromycin was added to the medium, and after a further 24 h, cells were passaged into a 10 cm dish with 2.5 µg/ml puromycin. 24 h later (72 h after transfection), cells were diluted to isolate single cells and

obtain clonal populations. This was achieved either by plating at low density in 10 cm dishes, or by serial dilution in 96-well plates. 1–2 weeks later, single colonies were identified, passaged using TrypLE (using cloning rings to isolate colonies in dishes) and expanded.

2.5.5 CFIm25 complementation experiments

Complementation with wild type (WT) and mutant CFIm25 in the 25KD cell line was carried out by transient transfection (see 2.5.3) in 6 cm dishes with 3 μg of the desired pcDNA3.1 plasmid and 1 μg of pSUPERIOR.puro (see Table 2.1). 24 h after transfection, cells were passaged into 10 cm dishes with 2.5 μg/ml puromycin to enrich for transfected cells. Cells were harvested for analysis 72 h later.

2.5.6 Stable cell line generation for CFIm68 complementation experiments

Stable cell lines re-expressing WT or mutant CFIm68 (or relevant controls) were generated using the HEK 293 Flp-In T-REx CFIm68 knockdown line (68KD) as a starting point. Cells in antibiotic-free medium were transfected (see 2.5.3) in 6-well plates with 0.3 μg of the desired pcDNA5/FRT-derived plasmid and 1.7 μg of the pOG44 FLP (flippase) recombinase plasmid (see Table 2.1).

The medium was replaced 24 h later, and after a further 24 h, cells were passaged into 10 cm dishes with medium containing 150 μg/ml hygromycin B and 15 μg/ml blasticidin. The hygromycin/blasticidin medium was replenished every 3–4 days, and hygromycin-resistant colonies were observed after 1–2 weeks. These were batch-passaged into small flasks for further growth and maintenance.

2.5.7 Live cell confocal microscopy

Cells were seeded in CELLview 35 mm glass-bottomed dishes (Greiner Bio-One) for imaging. GFP and Hoechst 33342 signals were imaged on an Olympus FV1000 IX81 confocal microscope system using a 60x 1.35NA oil immersion lens and FluoView software (Olympus). Hoechst 33342 (1 μg/ml)

was added directly to the medium 15 min before imaging. Images were edited and merged using Photoshop CS6 (Adobe).

2.5.8 Cell cycle analysis by flow cytometry

Harvested cells were transferred to a Falcon 2058 tube and fixed by addition of 2 ml cold 70% ethanol while vortexing, then stored at 4°C for later analysis. Immediately before use, fixed cell pellets were washed twice in cold PBS by centrifugation (2000g, 5 min) and resuspended in 1 ml propidium iodide (PI) staining solution. Samples were incubated in the dark for 15 min before analysis.

The FACSCalibur cell analyser and CellQuest Pro software (BD Biosciences) were used to analyse samples, using the FL-3 channel to measure PI fluorescence of 10,000 cells per sample.

PI staining solution: 40 µg/ml PI, 100 µg/ml RNase A, in PBS.

2.6 Whole genome/transcriptome methods

2.6.1 3' mRNA sequencing analysis

3' mRNA-seq library preparation

5 µg of total RNA was used to prepare sequencing libraries using the QuantSeq 3' mRNA-Seq Library Prep Kit for Ion Torrent (Lexogen) according to manufacturer's instructions, with 13 cycles used for the PCR amplification step.

Quality control of sequencing libraries

The quality of libraries was assessed following QuantSeq library preparation using a High Sensitivity DNA Chip (Agilent) and the Agilent 2100 Bioanalyzer system. DNA concentration was determined from the output trace by quantifying the area under the curve between 50 bp and 1000 bp.

Ion Torrent sequencing

Libraries were diluted to a final concentration of 300 pM and the Ion PI Hi-Q Chef Kit and Ion Chef system (ThermoFisher) were used for template preparation and chip loading according to manufacturer's instructions. Prepared chips were then sequenced using the Ion Proton system (ThermoFisher). FASTQ files were aligned to the human genome (hg19) using STAR - Spliced Transcripts Alignment to a Reference (Dobin et al., 2013). BAM files used for further processing had a minimum of one million reads.

Bioinformatic analysis of 3' mRNA-seq data

The 3'READS bioinformatics pipeline (Hoque et al., 2013; Neve et al., 2016) was used to process sequencing data and perform gene expression and APA analysis. Comparison of data with those obtained from previous 3'READS experiments allowed internal priming events to be removed. Significant APA changes in pairwise comparisons between the two most abundantly expressed isoforms were defined based on a minimum of 5% change in APA isoform and a value of $p<0.01$ in Fisher's exact test.

2.6.2 Pol II ChIP-sequencing analysis

ChIP was performed as described in Section 2.4.8. Samples were analysed by deep sequencing using Illumina HiSeq 4000 75 bp paired-end reads (Wellcome Trust Centre for Human Genetics, University of Oxford).

For bioinformatics analysis of data, adapters were first trimmed with Cutadapt v. 1.9.1 (Martin, 2011) with the following constant parameters: --minimum-length 10 –q 15, 10 –-max-n 1. Obtained sequences were mapped to the human hg19 reference sequence with Bowtie2 v. 2.2.5 (Langmead and Salzberg, 2012). Unmapped reads were removed with SAMtools v. 1.3.1 (Li et al., 2009). Mapped reads were then de-duplicated using Picard to remove PCR-duplicates. Bam files were sorted and indexed with SAMtools. Data were normalized to Reads Per Genomic Content (RPGC) by employing

deepTools2 v. 2.2.4 (Ramirez et al., 2016) bamCoverage tool with the following parameters: -bs 10 –normalizeTo1x -e –p max. Data was deposited at the Gene Expression Omnibus (GEO) with the GSE number 99955.

For generation of the metaprofiles, TSS and poly(A) site locations were obtained from GENCODE (which annotates the most distal poly(A) sites). Proximal poly(A) sites were as annotated in the 3'end sequencing data from this thesis. For pairwise comparisons as represented in the box-and-whisker plots, statistical analysis was performed with the Wilcoxon matched-paired signed rank test, with definitions used as follows:

Gene body (GB): 500 nt from TSS up to GENCODE poly(A) site

Poly(A) site (pA): GENCODE poly(A) site up to 2 kb after poly(A) site.

Transcription start site (TSS): -50 to +250 nt around GENCODE TSS

Pol II occupancy (at GB/pA/TSS): log2((GB or pA or TSS) IP x RPM)-((GB or pA or TSS) Input x RPM)/ length (GB or pA or TSS). Only genes with an occupancy value >0 after this calculation were included in the pairwise comparisons.

Chapter 3
Establishment of a CFIm complementation assay

3.1 Introduction

As discussed in Chapter 1, alterations in CFIm activity may influence polyadenylation site choice, as demonstrated by the widespread and directional shift towards use of proximal polyadenylation sites upon knockdown of CFIm25 or CFIm68. However, the extent to which CFIm modulates APA under physiological conditions remains unclear. Little is known about how the activity of CFIm itself is regulated, and in particular the potential contribution of PTMs to CFIm regulation remains underexplored.

In order to investigate determinants of CFIm activity and their influence on cleavage site choice, a CFIm complementation assay was established. The starting point for development of the assay was the clearly observable distal-to-proximal APA shift upon CFIm knockdown. If such a shift could be observed and then reversed by re-expression of wild type (WT) CFIm (complementation), this would provide the basis to investigate the functionality of a range of CFIm mutants by comparison with the wild type.

To generate cell lines with a permanent CFIm depletion, CRISPR/Cas9-based gene editing was used to target the CFIm25 (*NUDT21*) and CFIm68 (*CPSF6*) genes separately. Clonal cell lines were identified which showed substantial depletion of CFIm25 or CFIm68 and the expected 3'UTR shortening in two representative transcripts, as assessed by northern blotting. This 3'UTR shortening was reversible upon re-expression of CFIm25 or CFIm68 respectively, providing the basis for the complementation assay.

3.2 Generation and identification of CFIm25-depleted cell lines using CRISPR/Cas9 gene editing

Transient reduction of CFIm25 levels using siRNA or short hairpin RNA (shRNA) has been shown to cause global 3'UTR shortening in HeLa and HEK293 cells (Gruber et al., 2012; Kubo et al., 2006; Masamha et al., 2014). Using a transient knockdown method would, however, present various technical difficulties for a complementation assay due to the need for simultaneous knockdown of endogenous CFIm25 and re-expression of exogenous CFIm25. It was therefore decided that a permanent CFIm25 depletion approach would be more appropriate.

Initial attempts to generate CFIm25-depleted cell lines utilised stable integration of doxycycline-inducible shRNA constructs targeting a sequence in the *NUDT21* 3'UTR. These constructs were successfully integrated into HEK293 Flp-In T-REx cells using the FRT site. However, analysis of protein lysates from these cell lines by western blot revealed no reduction in CFIm25 expression upon doxycycline induction, despite some success using the same shRNAs in transient transfection (data not shown). It was therefore decided that stable shRNA expression was not a viable option for extended CFIm knockdown, and the advent of CRISPR/Cas9-based gene editing technology provided an alternative methodology.

The bacterial CRISPR/Cas9 immunity system has been adapted to allow RNA-guided targeting of a DNA endonuclease (Cas9) to a desired genomic location in human cells, as reviewed in (Ran et al., 2013). Cas9 introduces a double-strand break at the target site, which in the absence of a homologous template is repaired by non-homologous end joining (NHEJ). NHEJ is an inaccurate process, often leading to small insertions or deletions (indels) at the cut site that have the potential to cause frameshifts and ablate protein production. This approach was exploited in an attempt to ablate CFIm25 protein production.

A 20 bp sgRNA targeting exon 1 of *NUDT21* was designed and cloned into the pX459 plasmid, which also encodes the Cas9 nuclease coupled to a puromycin resistance gene (Figure 3.1). HRMA was used to confirm efficient cutting within the *NUDT21* locus at the batch population level (data not shown). The plasmid was transfected into HEK293T cells, and in a later round into HEK293 Flp-In T-REx ('Flp-In') cells. Following puromycin enrichment of transfected cells, clonal populations were isolated, expanded and screened by western blotting to identify any clones with a reduction in CFIm25 levels. In total, 140 clones were screened (77 HEK293T-derived and 63 Flp-In-derived).

Figure 3.2 shows a selection of these western blots, which are representative of the data obtained. A number of clones showed reduced CFIm25 expression (e.g. 293T clones 27 and 56, Flp-In clones 30 and 37), with quantification indicating that 78 clones (56%) had more than a 50% reduction. Several clones also showed a shift in the size of the CFIm25 band observed (e.g. 293T clone 27, Flp-In clones 10 and 66), suggestive of in-frame indels (this was later confirmed by sequencing; see Section 3.3). These results indicated that *NUDT21* gene editing was successful and relatively efficient. Despite this, no clone was identified with a total loss of CFIm25, suggesting that full CFIm25 knockout may be lethal. This view is supported by the identification of CFIm25 in genome-wide CRISPR screens for essential genes (Hart et al., 2015; Wang et al., 2015) and by the finding that re-targeting of substantially CFIm25-depleted clones with a different sgRNA targeting exon 2 did not lead to silencing of the remaining allele (data not shown).

A number of clones were identified with very low residual levels of CFIm25 – for example, HEK293T clones 27 and 73 had <10% of the parental protein level. Several such clones were selected for further analysis to investigate the genetic basis of the observed phenotypes.

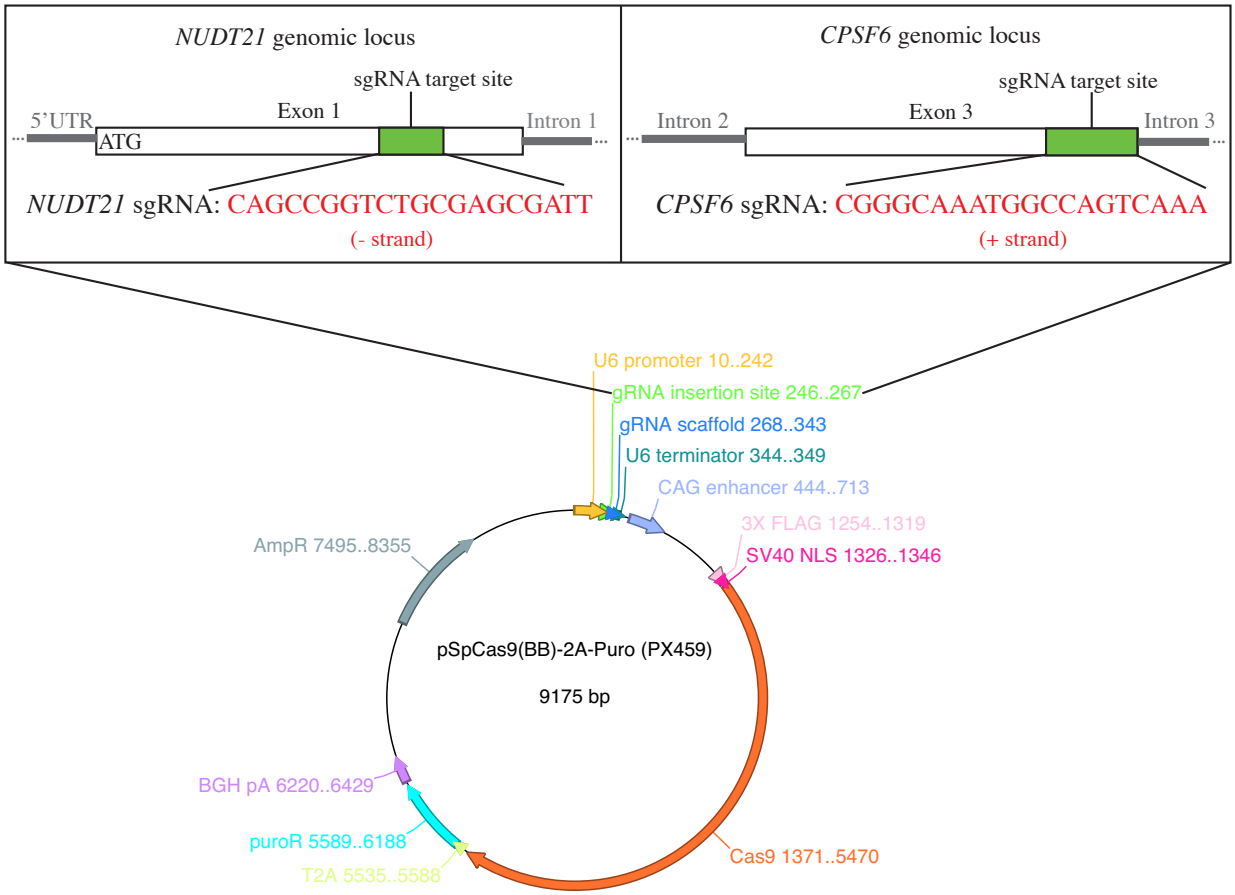


Figure 3.1: The pX459 plasmid and sgRNAs used for CRISPR targeting of *NUDT21* and *CPSF6*. Top: a depiction of the sequence of the sgRNAs used and the position of their target sites within the context of the *NUDT21* (left) and *CPSF6* (right) genomic sequences. Bottom: an illustration of the pX459 plasmid into which the sgRNAs were cloned, with other key features such as the Cas9-2A-puro cassette and sgRNA scaffold also highlighted.

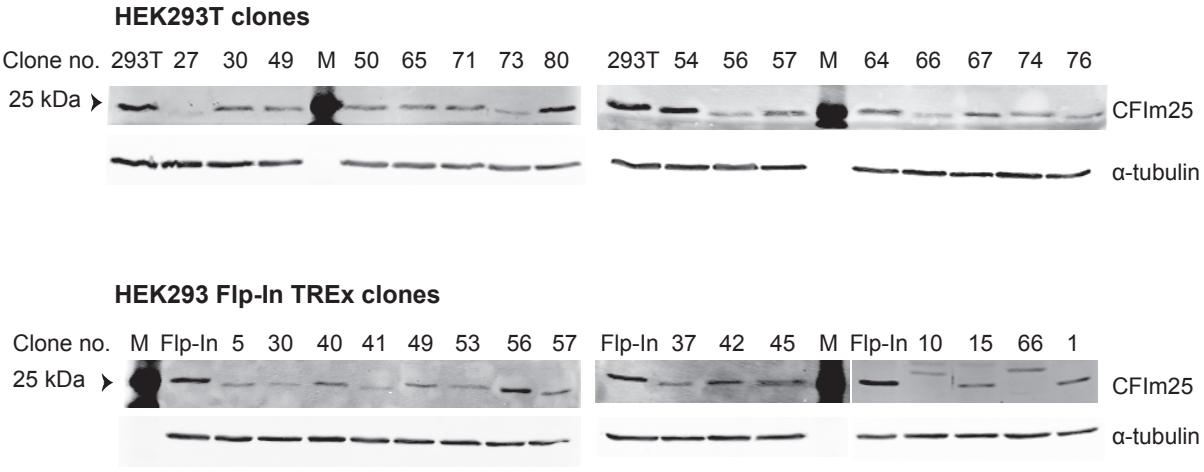


Figure 3.2: Western blot screening of clones following *NUDT21* CRISPR targeting. Whole-cell lysates from 77 293T-derived CRISPR lines (numbered 1-81 – some failed to grow) and 63 293 Flp-In-derived CRISPR lines (numbered 1-72 – some failed to grow) were screened by anti-CFIm25 western blotting to identify clones with reductions in CFIm25 expression. α-tubulin was used as a loading control. A representative selection of western blots is shown here to illustrate the typical phenotypes observed. 293T/Flp-In = parental cell line, M = marker lane.

3.3 Sequence analysis of CFIm25 knockdown cell lines

The observation that clones such as HEK293T 27 and 73 had very low, but nonetheless detectable, levels of CFIm25 was surprising. It was expected that editing of one of two alleles would lead to a reduction of around 50% in protein level. A possible explanation for this phenotype was found by referring to a study analysing the ploidy of HEK293 cells (Lin et al., 2014), which suggested that both lines used here have three copies of *NUDT21*. Disruption of two of three alleles may lead to particularly low protein levels, with >90% depletion perhaps possible if the remaining allele is weakly expressed.

Alternative explanations include the possibility that the residual band visible by western blot was non-specific signal, or that some of these clones were in fact knockouts, but with a small number of contaminating cells from another clonal population responsible for this 'ghost' band. To rule out the latter possibility, clone 27 was subjected to repeated clonal isolation – all 14 re-isolated clones still had the faint CFIm25 band visible (data not shown), suggesting that this was unlikely to result from a small number of contaminating parental cells.

To provide a more definitive explanation for the observed patterns of CFIm25 expression, genotyping was carried out on selected clones of interest. The genomic locus surrounding the sgRNA target site was amplified by PCR and sequenced. For most clones, despite three bands often being visible when products were separated by gel electrophoresis (data not shown), sequencing was only successful for one or two alleles (Table 3.1). The sequencing data provides explanations for many of the observed western blot band patterns. For example, Flp-In clone 66 had an 84 bp (28 amino acid) insertion, explaining the increased molecular weight of the western blot band (Figure 3.2). The data also confirmed that for clones 27 and 73, despite the faintness of the CFIm25 band, there was indeed a potentially functional allele present (with an in-frame deletion in each case).

Whilst these sequencing results were incomplete, they did confirm that at least three *NUDT21* alleles were present in these cells, that none of these clones were genetic knockouts, and in several cases defined the sequence of the remaining CFIm25 isoform present.

Table 3.1: A summary of *NUDT21* genotypes in selected CRISPR knockdown clones. The region of exon 1 targeted by the *NUDT21* sgRNA was PCR amplified from genomic DNA extracted from given clones. The product from gel-extracted bands or total PCR reaction was cloned by TA cloning and sequenced. Cells highlighted in red represent alleles expected to be unable to produce functional protein; those in green are alleles expected to produce altered CFIm25 protein isoforms. Where loss of amino acids is described (e.g. N7 to N22), this encompasses all amino acids between and inclusive of the given residues.

Clone	Allele 1 *NUDT21* genotype	Allele 2 *NUDT21* genotype	Allele 3 *NUDT21* genotype
HEK293T			
27 (lane 2, Fig. 3.2)	32 bp deletion including start codon.	Different 32 bp deletion including start codon.	48 bp in-frame deletion – loss of N7 to N22.
66 (lane 15, Fig. 3.2)	1 bp insertion - premature stop codon.	29 bp deletion - premature stop codon.	
73 (lane 8, Fig. 3.2)	1 bp insertion - premature stop codon.	27 bp in-frame deletion - loss of S2 to Q10.	
HEK293 Flp-In			
10 (lane 33, Fig. 3.2)	60 bp in-frame insertion – 20 amino acids inserted between N7 and R8.		
15 (lane 34, Fig. 3.2)	9 bp in-frame deletion - loss of R8 to Q10.		
66 (lane 35, Fig. 3.2)	1 bp insertion - premature stop codon.	84 bp in-frame insertion – 28 amino acids inserted between N7 and R8.	

3.4 Analysis of APA in selected CFIm25 knockdown cell lines

The level of CFIm25 depletion achieved in some of these cell lines (up to 90%) was comparable to that achieved by siRNA knockdowns (Gruber et al., 2012), but on a permanent basis. The next step was to establish whether this permanent CFIm25 depletion led to a distal-to-proximal poly(A) site shift comparable to that observed upon transient knockdown.

This was investigated by northern blotting using radioactively-labelled DNA probes for the coding regions of two transcripts with well-defined APA sites – tissue inhibitor of metalloproteinase 2 (*TIMP2*) and syndecan 2 (*SDC2*) (Figure 3.3A). These had both been previously been validated as transcripts that demonstrate the characteristic 3'UTR shortening in response to CFIm25 knockdown (Kubo et al., 2006). Total RNA from a number of 293T- and Flp-In-derived clones with varying extents of CFIm25 depletion was analysed using this northern blotting approach, and band intensities were quantified and used to calculate proximal-to-distal (P/D) polyadenylation ratios (Figure 3.3B).

Unexpectedly, the majority of clones screened showed very little difference in APA pattern to the parental cells (Figure 3.3B). In particular, it was surprising that clone 27, with <10% of the wild-type CFIm25 level remaining, showed no change in P/D ratio for *TIMP2* and only a three-fold increase in P/D ratio for *SDC2* (not readily obvious by visual inspection). This suggested that with long-term CFIm25 depletion, as opposed to transient knockdown, cells may somehow compensate for CFIm25 loss and revert to the original APA pattern of the parental line.

However, one cell line, 293T clone 73, despite having a similar extent of knockdown to clone 27 (~90%), clearly showed the expected switch to an increased relative usage of the more proximal poly(A) sites in both transcripts. Quantification of repeated northern blots after extended passaging of this clone confirmed that the change in P/D ratio was permanent and significant ($p<1x10^{-4}$ for both transcripts; Figure 3.4B).

It is unclear why clone 73 behaved differently from other CFIm25-depleted clones – possible explanations include a clone-specific inactivation of a compensatory pathway needed to reverse the APA change, or an impaired functionality of the remaining mutant allele in this clone. Another possibility was that this phenotype might be a clone-specific off-target effect, unrelated to CFIm25 depletion. This was explored further by testing the complementation ability of re-expressed CFIm25 in

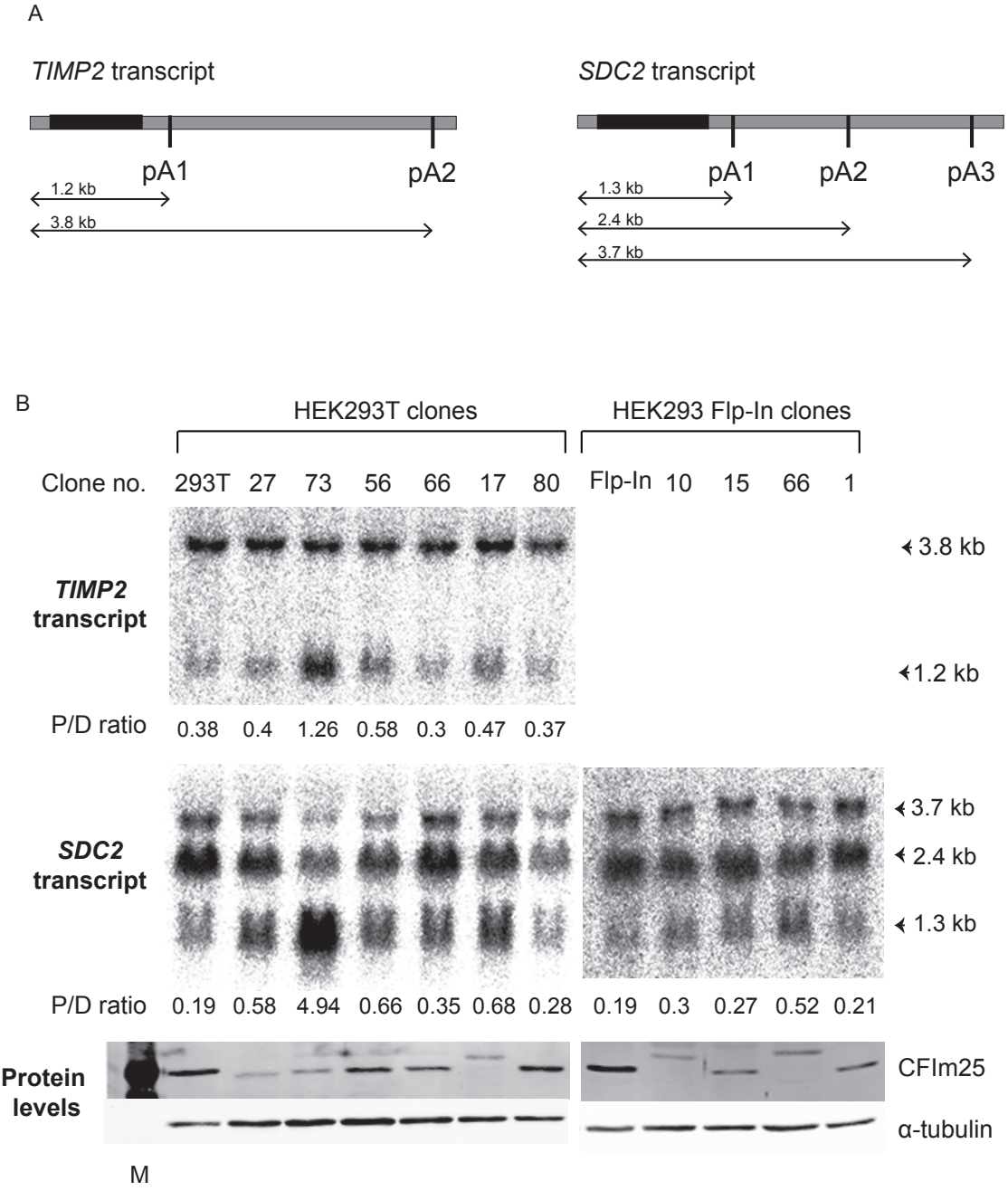


Figure 3.3: Northern blot analysis of APA in CFIm25-depleted clones. (A) An illustration of APA site positions and resulting transcript sizes in the *TIMP2* and *SDC2* genes. Black shading = coding region, grey shading = UTRs (not to scale). (B) Northern blotting against the protein-coding region of the *TIMP2* and *SDC2* transcripts was performed on total RNA isolated from selected CFIm25-depleted CRISPR clones. 293T/Flp-In refer to the non-targeted parental cell lines. Proximal/distal ratios (P/D ratios) were calculated from band intensities measured using 2D densitometry (smallest transcript divided by sum of longer transcripts). Corresponding western blots at the bottom show CFIm25 expression level in the clones; α-tubulin was used as a loading control. (M = marker lane). Clone 80 served as a CRISPR-targeted but non-CFIm25-depleted control.

this clone. Experiments were therefore carried forward with clone 73, which from hereon will be referred to as the '25KD' ('knockdown') cell line.

3.5 Complementation of APA in the 25KD cell line by re-expression of wild type CFIm25

To re-express CFIm25 in the 25KD line, a transient transfection was carried out using a pcDNA3.1-derived plasmid encoding the *NUDT21* cDNA (non-tagged) under the control of the human cytomegalovirus (CMV) immediate early promoter. To improve the uniformity of CFIm25 expression, this was co-transfected with a plasmid encoding a puromycin resistance gene (pSUPERIOR.puro), allowing enrichment of efficiently transfected cells by means of a 72 h period of puromycin selection. This method gave a robust re-expression of CFIm25, with expression levels comparable to or greater than those seen in the parental cell line (Figure 3.4A). It is perhaps surprising that the levels of expression achieved were not higher, as the CMV promoter usually supports considerable overexpression – this may hint at the potential existence of regulatory mechanisms that limit CFIm protein levels (discussed further in Section 3.10).

Analysis of APA upon CFIm25 re-expression by northern blotting showed a partial reversal towards the original 293T APA pattern, which did not occur in the control transfection with the parent pcDNA3.1+ plasmid (Figure 3.4A). This reversal was not complete, probably due to incomplete transfection – even with puromycin selection, it is unlikely that all cells re-expressed CFIm25. Although partial, the APA reversal was robust and significant ($p<0.05$ for *TIMP2* and $p<0.001$ for *SDC2*; Figure 3.4B). Importantly, this confirmed that the APA change was CFIm25-dependent.

To assess whether the 3'UTR shortening and complementation phenotypes observed for *TIMP2* and *SDC2* were globally representative, a transcriptome-wide approach was employed. 3' end sequencing libraries were generated using the QuantSeq method (Figure 1.4) and sequenced using the IonProton platform. This approach was applied to three RNA samples isolated from a transfection experiment as

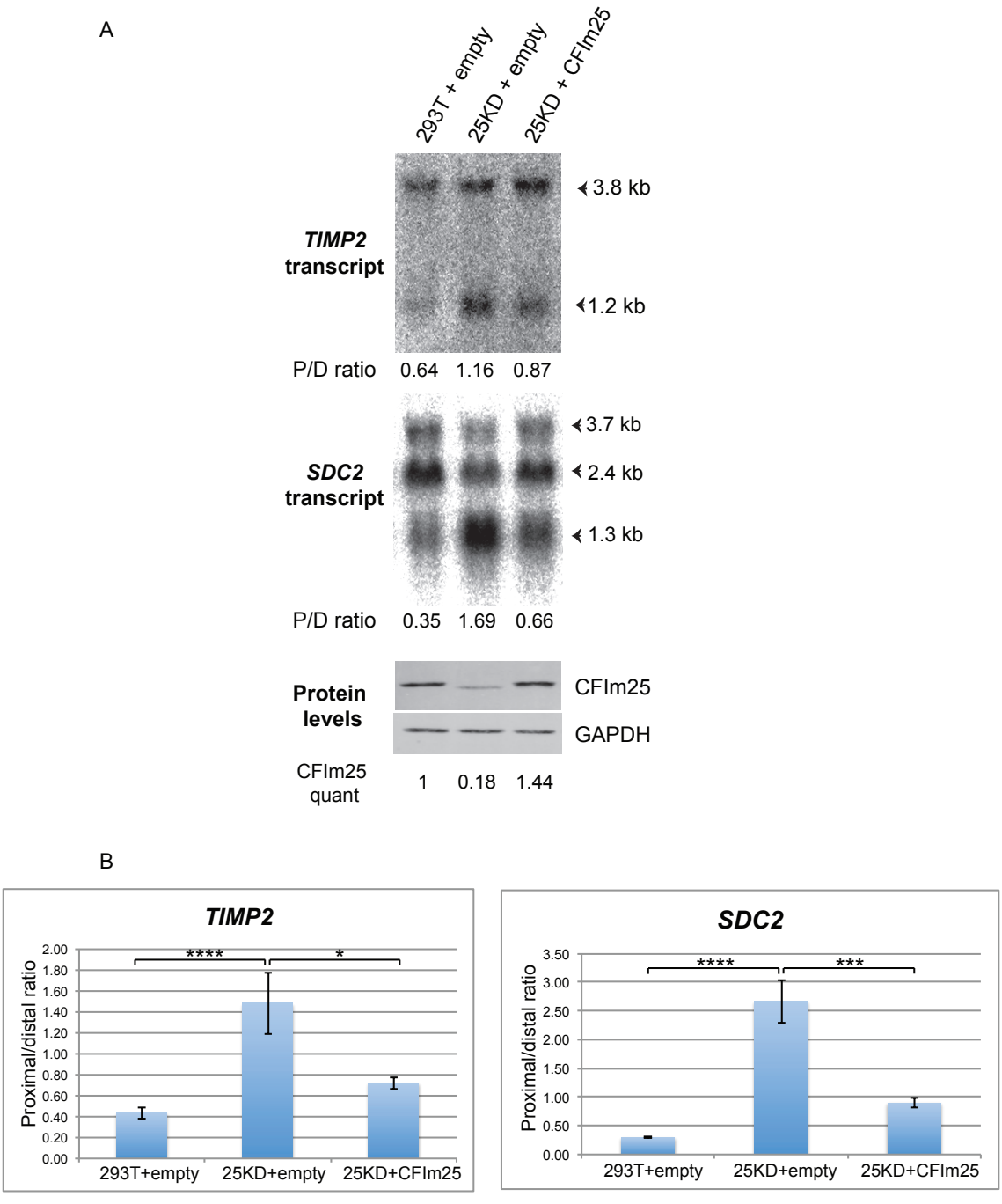


Figure 3.4: Complementation of the APA shift by re-expression of CFIm25 in the 25KD cell line. (A) Northern blotting and P/D ratio calculations were performed as previously described (Figure 3.3) on the given samples. For all samples, cells were transiently transfected with parent pcDNA3.1+ as a control (+empty) or pcDNA3.1-CFIm25 (+CFIm25) along with pSUPERIOR.puro to allow puromycin enrichment of transfected cells. The corresponding western blot at the bottom shows CFIm25 expression in the transfected clones. Fluorescence signals were used to quantify CFIm25 expression ('CFIm25 quant'), with CFIm25 band intensities normalised to GAPDH (loading control) band intensities. (B) Mean P/D ratios from repeats of this analysis (n=7). Error bars depict ± standard error of the mean (SEM) and statistical significance, calculated using a two-tailed Student's t-test, is illustrated with * symbols.

described above - 293T and 25KD (each transfected with the empty vector control) and 25KD+CFIm25 (transiently transfected with theCFIm25-expressing plasmid). Data were analysed using the 3'READS pipeline (Hoque et al., 2013) and pairwise comparisons were performed to highlight transcripts undergoing significant changes ($p<0.01$) in poly(A) site usage between samples.

Comparison of poly(A) site usage between the parental 293T line and the 25KD line revealed a widespread 3'UTR shortening in the 25KD line at the transcriptome-wide level (Figure 3.5A). Of 6700 transcripts with APA sites identified here, 797 (11.9%) underwent a significant change in APA site usage in the 25KD line, and 746 (93.6%) of these events were 3'UTR shortening. This supported previously published findings (Gruber et al., 2012) and confirmed that the phenotype observed by northern blotting with two transcripts was representative of the overall trend.

Furthermore, comparison of poly(A) site usage between transiently re-expressed CFIm25 (25KD+CFIm25) and 25KD demonstrated that there was widespread complementation on CFIm25 re-expression, with 6.7% of transcripts showing significant APA changes, 80.1% of which were 3'UTR lengthening upon CFIm25 re-expression (Figure 3.5B). While the number of affected transcripts was slightly lower here than in the 293T vs. 25KD comparison, there was still a clearly observable directional bias, confirming that re-expressed CFIm25 could reverse the 25KD-induced 3'UTR shortening at a global level.

A biological repeat of this particular comparison gave similar results (5.3% of transcripts showing APA changes, 78.1% lengthening upon re-expression; Figure 3.5D). However, the overlap between the identities of transcripts showing complementation, while significant, was not as large as expected, with only 77 transcripts showing significant 3'UTR lengthening in both repeats (Figure 3.5E). This may be a result of the significance boundary used, as some transcripts underwent a similar APA change in both experiments but narrowly missed the $p<0.01$ cut-off in one. It may also reflect some randomness in transcript-specific responses to CFIm25 re-expression, particularly as the extent and uniformity of re-expression achieved in each transient transfection experiment is variable.

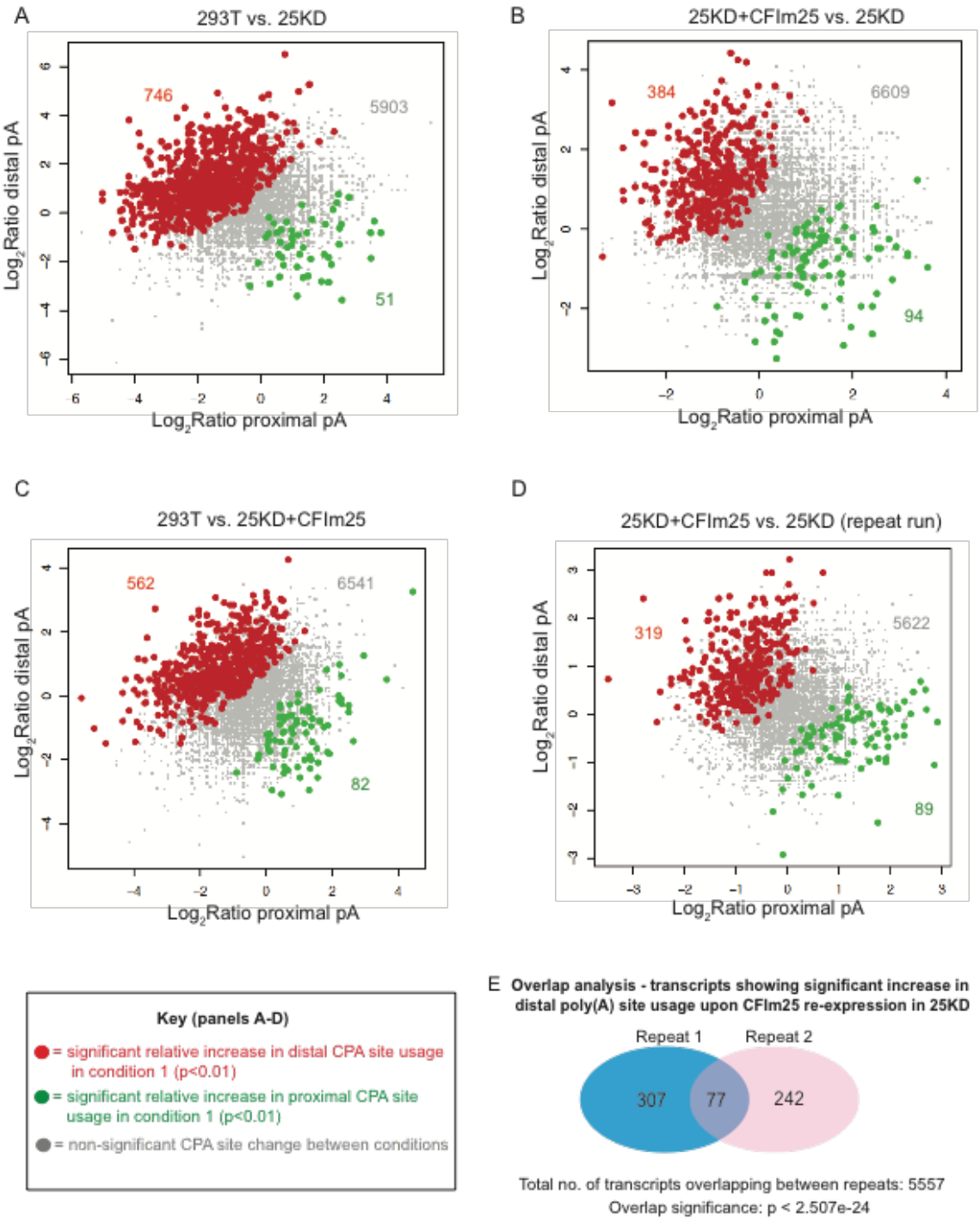


Figure 3.5: Transcriptome-wide analysis of APA upon CFIm25 depletion and re-expression. Total RNA was isolated following a transfection experiment (as in Figure 3.4) and 3'end sequencing libraries were prepared and analysed by IonProton sequencing. Reads were mapped to the human genome and aligned to known poly(A) sites (Derti et al., 2012). The 3'READS bioinformatic pipeline (Hoque et al., 2013) was used to calculate APA changes. (A)-(D) Pairwise comparisons of transcriptome-wide poly(A) site use between the given sample pairs. 293T or 25KD alone refer to samples transfected with the empty pcDNA3.1+ plasmid; 25KD+CFIm25 was transfected with pcDNA3.1-CFIm25. Individual transcripts are plotted as dots with reference to the log2ratio of proximal and distal pA site usage in condition 1 vs. condition 2 (as ordered in figure titles). Grey dots - no significant difference in poly(A) site usage; red dots - increased relative distal site usage in condition 1; green dots - increased relative proximal site usage in condition 1 (>5% change in APA isoform and p<0.01, Fisher's exact test). (D) is a biological repeat of (B). (E) Venn diagram of overlap between transcripts showing 3'UTR lengthening in CFIm25 re-expression experiment in the two repeats (red dots in (B) and (D)). Overlap significance was calculated using an online tool at http://nemates.org/MA/progs/overlap_stats.html with the total number defined as transcripts identified as subject to APA in both repeats.

Although the CFIm25 re-expression did result in a widespread complementation phenotype, comparison of this condition with the 293T parental line demonstrated that the extent of this complementation was only partial (Figure 3.5C). The 293T line showed a clear 3'UTR lengthening compared with re-expressed CFIm25 (9.0% of transcripts showed differing APA patterns, 87.2% of which were lengthening events in 293T). This is consistent with the observations made by northern blotting, whereby re-expression of CFIm25 led to a significant but partial 3'UTR lengthening for *TIMP2* and *SDC2*, with the original 293T APA pattern not fully restored (Figure 3.4).

Overall, the transcriptome-wide data confirmed the generality of observations made by northern blotting with two transcripts. This demonstration of APA complementation by CFIm25 re-expression in the 25KD cell line provided the possibility to screen the functionality of CFIm25 mutants, as discussed in Chapters 4 and 5.

3.6 Generation and identification of CFIm68-depleted cell lines using CRISPR/Cas9 gene editing

Following the partial success of the CRISPR/Cas9-based CFIm25 depletion approach, the same method was applied to generate CFIm68-depleted cell lines, as transient knockdown of CFIm68 induces a similar APA phenotype to that seen on CFIm25 knockdown (Kim et al., 2010; Martin et al., 2012). In this instance, targeting was only attempted in Flp-In cells. These cells contain a genomic FRT recombination site, which can be used to stably integrate and constitutively express a gene of interest (see Section 3.9. and Figure 3.8), facilitating a stable CFIm complementation approach. The same targeting and screening methodology was used, as described in Section 3.2, and the sgRNA used here targeted exon 3 of *CPSF6*, in a region close to the following intron (Figure 3.1).

A total of 43 clones were screened, with representative western blots shown in Figure 3.6. There was again evidence of successful gene editing, although this was considerably less efficient than for

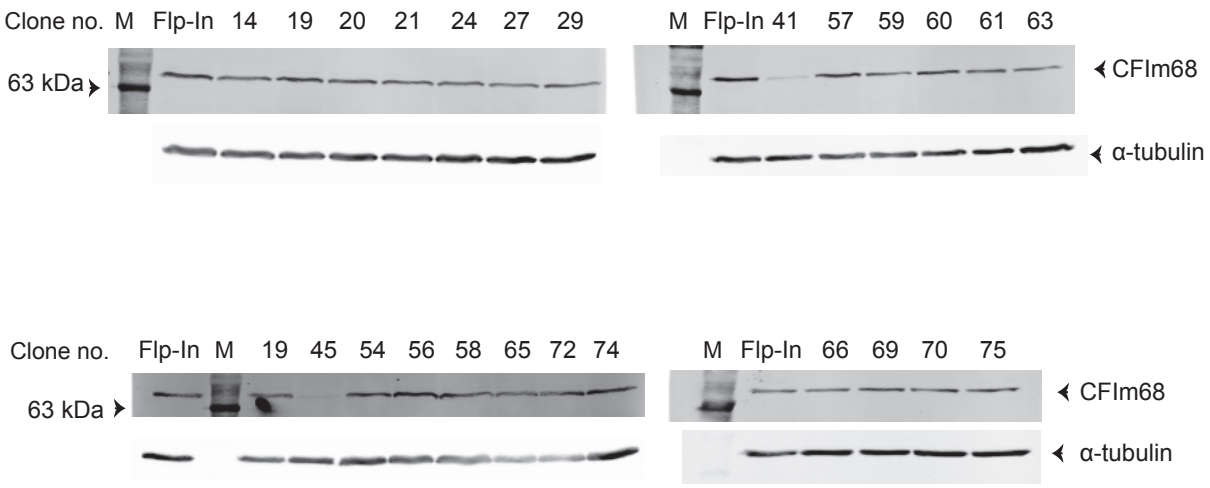


Figure 3.6: Western blot screening of clones following *CPSF6* CRISPR targeting. Whole-cell lysates from 43 Flp-In-derived CRISPR lines (numbered 1-75 – some failed to grow) were screened by anti-CFIm68 western blotting to identify clones with reductions in CFIm68 expression. α-tubulin was used as a loading control. A representative selection of western blots is shown here to illustrate the typical phenotypes observed. Flp-In = parental cell line, M = marker lane.

NUDT21 - quantification indicated that seven clones (16%) had more than a 50% reduction in CFIm68 expression. Changes in molecular weight from in-frame indels were this time not apparent on western blot, possibly because of the larger size of CFIm68. Two clones, numbers 41 and 45, showed a very substantial CFIm68 depletion (~4% and 8% of WT level respectively). These were considered as possible CFIm68 knockouts, and these and other clones of interest were taken forward for sequencing.

3.7 Sequence analysis of CFIm68 knockdown cell lines

Genotyping of CFIm68 was carried out in the clones of interest as described in Section 3.3 (Table 3.2). Surprisingly, in some cases, up to five PCR products were visible by agarose gel electrophoresis (data not shown), suggesting that up to five alleles of *CPSF6* may be present in these cells. Although the HEK293 ploidy study suggests that there are only three copies of the gene (Lin et al., 2014), it cannot be ruled out that more have arisen through amplification or chromosomal instability in this isolate. The fact that CFIm68 is often found overexpressed in cancer cells (http://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=CPSF6), possibly due to co-amplification with the neighbouring oncogene *MDM2*, supports this possibility.

As before, it was not possible to ensure complete coverage of all alleles in these clones. Analysis was also complicated by the fact that gene editing often disrupted the 5' splice site of intron 3, making the interpretation of the impact on protein production slightly more complex. However, the data were still informative in several respects.

In particular, for clone 41, three different alleles were sequenced, supporting the idea that there are at least three *CPSF6* alleles in these cells. Amongst these alleles, two had deletions spanning the splice donor site, which would presumably lead to intron retention and premature translation termination. In the case of the 8 bp deletion, even if a nearby alternative downstream splice donor were used, a premature stop codon would arise. The remaining allele had a 3 bp deletion, but would be expected to

produce protein. This therefore suggests that this clone is not a knockout, but, like the 25KD clone, has all but one allele disrupted and thus very low CFIm68 levels remaining. The particularly low level of CFIm68 protein in clone 41 is striking, and raises the possibility that CFIm68 produced from the remaining in-frame allele, which has a single amino acid deletion of lysine 124 (K124), could be particularly unstable. There are certainly examples of a single amino acid deletion conferring substantial instability on a protein, with the most notable being the effect of the cystic fibrosis-associated ΔF508 deletion on the CFTR protein (Lukacs and Verkman, 2012). Again, the possibility that a small number of contaminating non-edited cells were responsible for the residual CFIm68 here was ruled out by repeating clonal isolation on clone 41 (data not shown).

Table 3.2: A summary of CFIm68 genotypes in selected CFIm68 CRISPR knockdown clones. The region of exon 3/intron 4 targeted by the CFIm68 sgRNA was PCR amplified from genomic DNA extracted from given clones. The product from gel-extracted bands or total PCR reaction was cloned by TA cloning and sequenced. Cells highlighted in red represent alleles expected to be unable to produce functional protein; those in green are alleles expected to produce WT or altered CFIm68 protein isoforms. Where loss of amino acids is described (e.g. S123 to G125), this encompasses all amino acids between and inclusive of the given residues.

Clone	Allele 1 CFIm68 genotype	Allele 2 CFIm68 genotype	Allele 3 CFIm68 genotype
24 (lane 6, Fig. 3.5)	S123L substitution.	3 bp in-frame deletion - silent mutation in Q122 and loss of S123.	
27 (lane 7, Fig. 3.5)	Wild type.	S123L substitution.	15 bp deletion and 59 bp insertion - premature stop codon.
29 (lane 8, Fig. 3.5)	S123 silent mutation.	1 bp insertion - premature stop codon.	
41 (lane 10, Fig. 3.5)	3 bp in-frame deletion - loss of K124.	18 bp deletion spanning S123-G125 and splice donor site - premature stop codon inclusion through intron retention.	8 bp deletion spanning K124-G125 and splice donor site - premature stop codon inclusion through intron retention.
45 (lane 18. Fig. 3.5)	9 bp in-frame deletion – loss of S123-G125.	166 bp insertion - premature stop codon.	

These results again suggest that perhaps CFIm68 knockout is lethal, though it is possible that screening more clones may have resulted in the identification of a full knockout. Indeed, editing on the whole was less efficient for *CPSF6* than for *NUDT21*. Moreover, a recent study did claim to have achieved *CPSF6* knockout in HEK293 cells (Sowd et al., 2016).

3.8 Analysis of APA in selected CFIm68 knockdown cell lines

As in Section 3.4, northern blotting was used to assess the impact of permanent CFIm68 depletion on APA profiles in a selection of clones (Figure 3.7). It was found that clones 41 and 45, which had the most substantial CFIm68 knockdown, showed a dramatic 3'UTR shortening for both transcripts tested. The shift here was more obvious than in 25KD (compare to Figure 3.3), with a marked decrease in use of the *TIMP2* distal poly(A) site and a complete loss of detectable use of the middle and distal sites in *SDC2* (Figure 3.7). Some clones with a lesser depletion still showed a subtle APA change – for example clones 17, 24 and 27 showed a three- to four-fold increase in P/D ratios - but this was far from the striking change seen for clones 41 and 45.

The response of cells to changes in CFIm68 level was generally more dose-dependent and consistent than that seen for CFIm25 depletion. Whereas with CFIm25 CRISPR cells only clone 73 showed the expected 3'UTR shortening despite others having similar levels of CFIm25 depletion, here the two clones with >90% CFIm68 depletion both showed the same dramatic shift towards proximal poly(A) site use, with lesser-depleted clones showing a subtler phenotype. The reason for this difference is unclear, but may reflect an inherent difference in the compensatory responses to CFIm25 and CFIm68 depletion.

Clones 41 and 45 showed a similarly clear APA response to CFIm68 depletion. Clone 41 was selected for development of the complementation assay, and will from hereon be referred to as ‘68KD’.

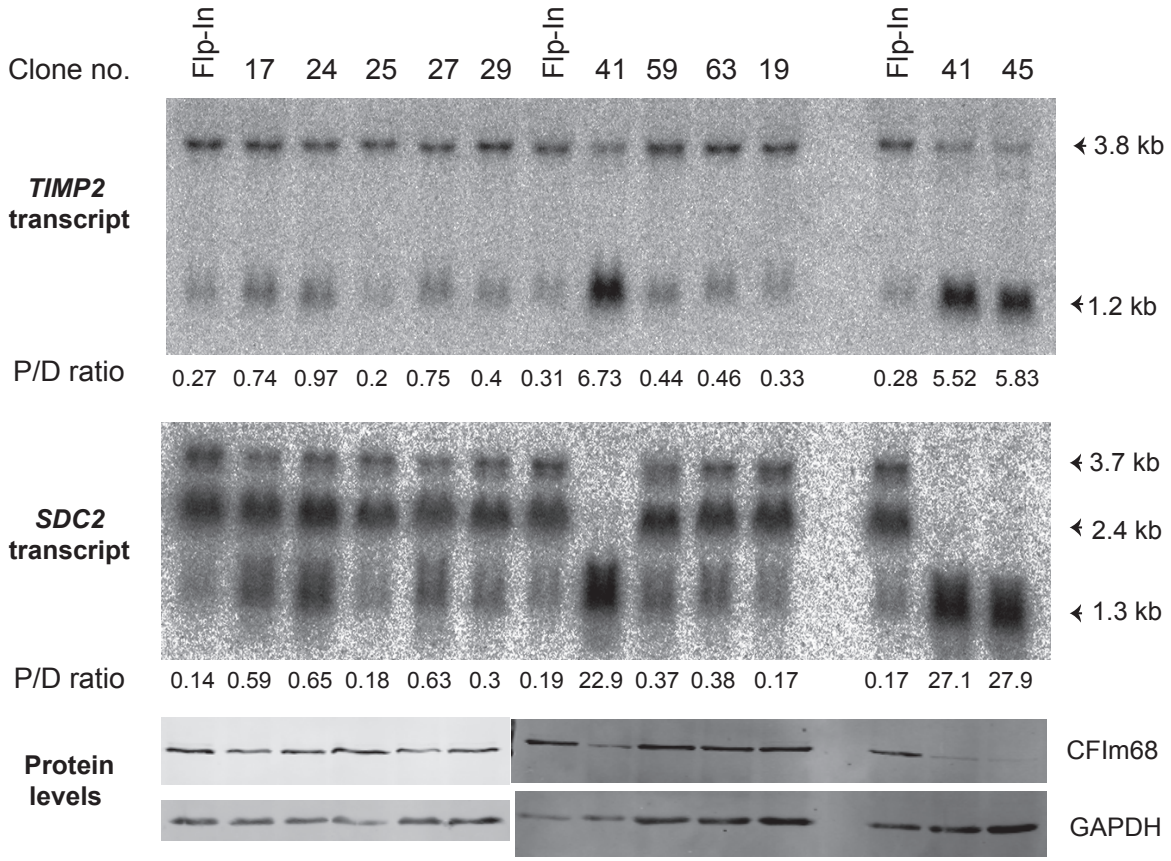


Figure 3.7: Northern blot analysis of APA in CFIm68-depleted clones. Northern blotting against the *TIMP2* and *SDC2* transcripts (see Figure 3.3 for illustration) was performed on total RNA isolated from selected *CPSF6*-targeted CRISPR clones. 'Flp-In' refers to the parental non-targeted HEK293 Flp-In cell line. P/D ratios were calculated as described in Figure 3.3. Corresponding western blots at the bottom show CFIm68 expression level in the clones; GAPDH was used as a loading control.

3.9 Complementation of APA in the 68KD cell line by re-expression of wild type CFIm68

As the 68KD cell line was derived from Flp-In cells, the presence of the FRT recombination site could be exploited for CFIm68 re-expression by FLP recombinase-mediated integration of a CFIm68 expression cassette. This offered clear advantages over the transient expression approach described in Section 3.5. Firstly, it allows a more permanent, stable CFIm68 re-expression such that cells can be studied over a long time period without repeated transfection. In addition, it promotes uniformity of expression – stable integrants are selected on the basis of conferred hygromycin resistance, such that all selected cells should re-express CFIm68 at comparable levels.

A *CPSF6* cDNA (non-tagged) was cloned into the pcDNA5/FRT plasmid, and this was transfected into the 68KD cell line along with a large excess of the pOG44 plasmid expressing FLP recombinase (Figure 3.8A). Stable integrants were hygromycin-selected and expanded, and then analysed by western blotting. This approach yielded a cell line with clear re-expression of CFIm68, which was absent from the control line with only the parent plasmid integrated (Figure 3.8B; 68KD+CFIm68 vs. 68KD+empty). The level of CFIm68 expression was 1.26-fold higher than in the parental cells (Figure 3.8B).

Analysis of this cell line by northern blotting (as described in Section 3.4) revealed a complete reversion to the parental APA phenotype upon CFIm68 re-expression, which was not observed in the control line with the empty plasmid integrated (Figure 3.8B). This was an improvement on the partial complementation profile observed with the CFIm25 re-expression (Figure 3.4), presumably owing to the more stable and uniform re-expression of CFIm68. This complementation was robust and significant (Figure 3.8C, $p<0.00001$ for *TIMP2* and $p<0.01$ for *SDC2*), remaining stable upon repeated harvests after many passages.

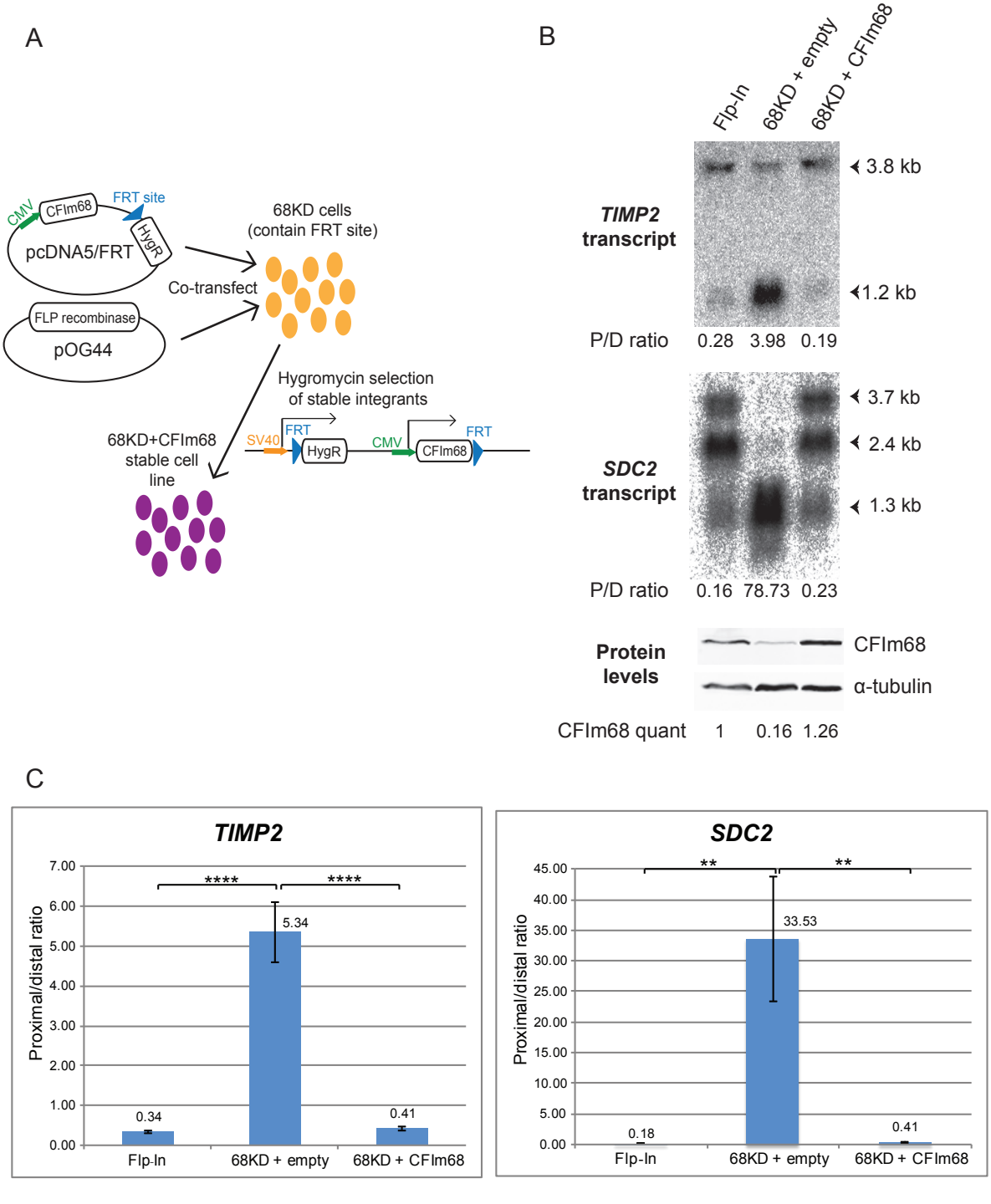


Figure 3.8: Complementation of APA by re-expression of CFIm68 in the 68KD cell line. (A) Schematic illustrating generation of the 68KD+CFIm68 cell line. (B) Northern blotting and P/D ratio calculations were performed as previously described (Figure 3.3) on the given RNA samples. 68KD+empty and 68KD+CFIm68 are stable cell lines generated by integration of the control pcDNA5/FRT empty plasmid or pcDNA5/FRT-CFIm68 at the genomic FRT site in 68KD. The corresponding western blot (bottom) shows CFIm68 expression in the transfected clones, which was quantified as described in Figure 3.4. (C) Mean P/D ratios from biological repeats (n=6). Error bars depict ±SEM and statistical significance, calculated using a two-tailed Student's t-test, is illustrated with * symbols.

The CFIm68 complementation observed here confirmed that the APA shift observed in 68KD was CFIm68-dependent, and provided the basis for a CFIm68 complementation assay.

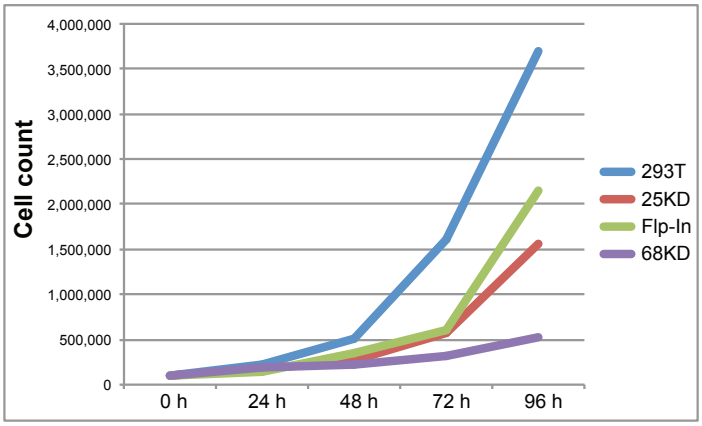
3.10 Phenotypic analysis of 25KD and 68KD cell lines

As well as using the 25KD and 68KD cell lines as the basis for CFIm complementation assays, further investigations were undertaken to establish whether there were any notable phenotypes resulting from CFIm25 or CFIm68 knockdown. One aspect studied was the growth of the knockdown cell lines compared to their parental counterparts, as particularly for 68KD, a clear increase in doubling time was observed during routine culture.

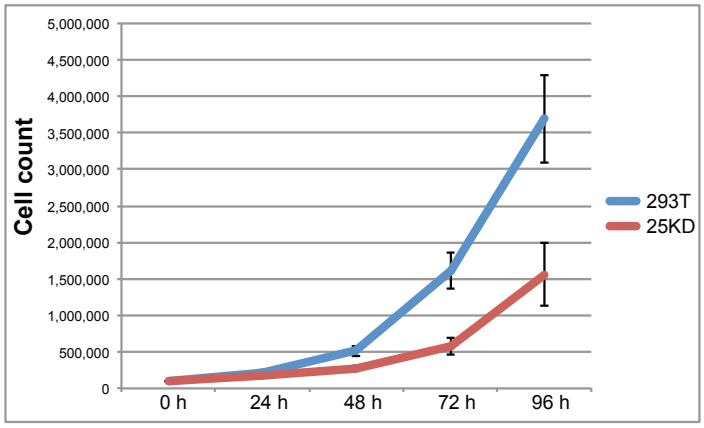
A growth timecourse was carried out for the four cell lines by plating 100,000 cells of each at day 0 and counting cells daily for a further four days. These results, averaged over three experiments, are plotted in Figure 3.9. The first point of interest is that the HEK293T-based lines grow faster in general than the Flp-In-based lines (Figure 3.9A), which is likely to be a result of the expression of the SV40 large T antigen in the former. This highlights the differing biology of the two parental lines used for CRIPSR gene editing, and thus the importance of considering the pairs separately, by comparison to the appropriate parental line.

Looking at the pairwise comparisons, it was apparent that both CFIm25 and CFIm68 knockdown led to decreased cell growth rate (Figure 3.9B, 3.9C). The effect of CFIm68 knockdown was particularly substantial, with more than a 2.5-fold increase in the cell doubling time (Figure 3.9C) – this is consistent with previous reports (Sowd et al., 2016). The stronger effect of CFIm68 knockdown may be a result of the more effective CFIm depletion in 68KD than 25KD, or it may be a CFIm68-specific phenotype. There are number of possible explanations that may account for the observed slow growth phenotype, including increased apoptosis, impaired cell cycle entry or slower progression through the

A

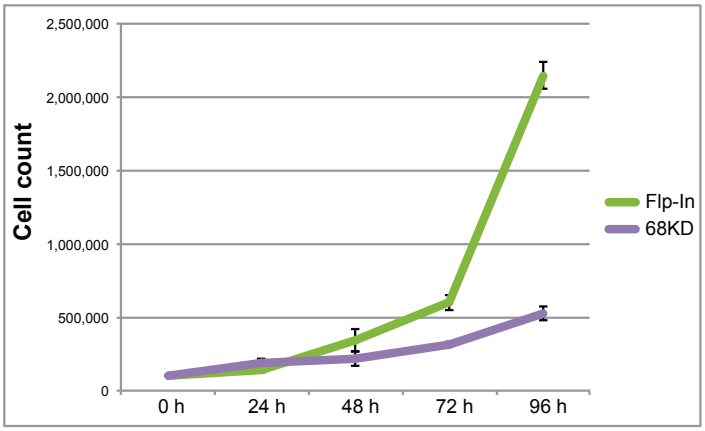


B



Doubling time (24 h - 96 h)	
293T	17.8 h
25KD	22.7 h

C



Doubling time (24 h - 96 h)	
Flp-In	18.4 h
68KD	48.4 h

Figure 3.9: Growth rates of 25KD and 68KD cell lines. For each given cell line, 4 wells of 100,000 cells each were seeded in a 6-well plate at 0 h. Every 24 h, a well of cells was harvested by trypsinisation and cells were stained with trypan blue and counted using a KOVA Glasstic slide grid. (A) Plot of mean cell counts for the 4 cell lines across the 96 h timecourse, from 3 independent experiments. (B) Separate plot of the 293T/25KD pair from (A). Error bars depict standard error of the mean (SEM). (C) As in (B), for the Flp-In/68KD pair.

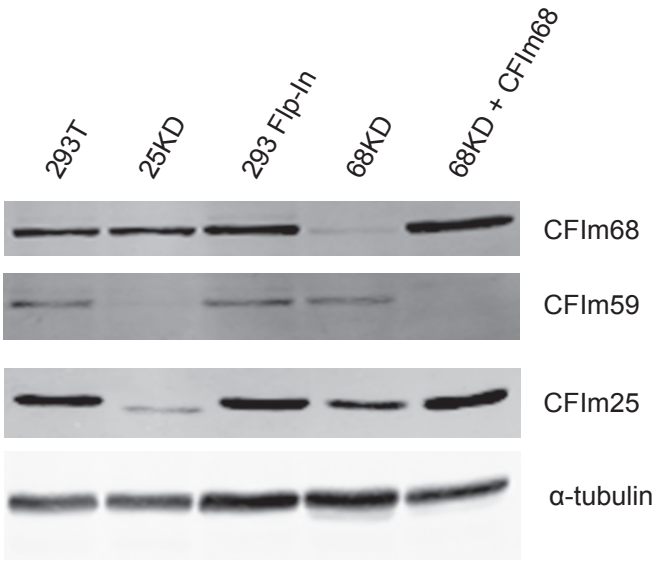
cell cycle. These possibilities were not explored here, although a previous report indicated that cells lacking CFIm68 did not show impaired progression through the cell cycle stages (Sowd et al., 2016).

As well as changes in growth rate, another point of interest was whether any interdependency of CFIm subunit levels could be observed, such that depletion of CFIm25 or CFIm68 individually led to co-depletion of the other subunits. This was assessed for all three subunits by western blotting of the different cell lines (Figure 3.10), with quantification across four repeat blots following separate cell harvests.

In the 25KD line, there was a near-complete (95%) loss of CFIm59 expression, while CFIm68 levels were not affected (Figure 3.10A, Figure 3.10B left graph). This suggests that expression or protein stability of CFIm59 is particularly sensitive to CFIm25 abundance, and is consistent with previous reports of the effect of transient CFIm25 knockdown on CFIm59 expression (Gruber et al., 2012; Kim et al., 2010). The observation that CFIm68 levels were not depleted in the same way suggests an interesting difference in regulation between these two large subunits. Previous studies had shown a small decrease in CFIm68 level upon CFIm25 transient knockdown (Gruber et al., 2012; Kim et al., 2010). This may have been a result of better knockdown efficiency, or alternatively may reflect a long-term compensation mechanism here that restores CFIm68 levels in response to CFIm25 loss.

In the 68KD line, there was a substantial (79%) co-depletion of CFIm25, while CFIm59 levels were not affected (Figure 3.10A; Figure 3.10B right). This suggests a lack of interdependent regulation between the two large subunits (although CFIm59 knockdown was not studied here), but a clear sensitivity of CFIm25 to CFIm68 abundance. Again, this is broadly similar to previous reports (Gruber et al., 2012; Kim et al., 2010). One interesting potential implication of this finding is the suggestion that the 68KD APA phenotype could be indirectly attributable to the concomitant decrease in CFIm25 expression. However, there are clear indications that this is not the case. Firstly, the 68KD phenotype (Figure 3.7) was much stronger than the 25KD phenotype (Figure 3.4), with a more

A



B

Mean band quantification (as proportion of parental levels)			
	CFIm25	CFIm59	CFIm68
293T	1.00	1.00	1.00
25KD	0.14	0.05	1.08
293 Flp-In	1.00	1.00	1.00
68KD	0.21	0.86	0.08
68KD+CFIm68	0.90	0.11	1.62

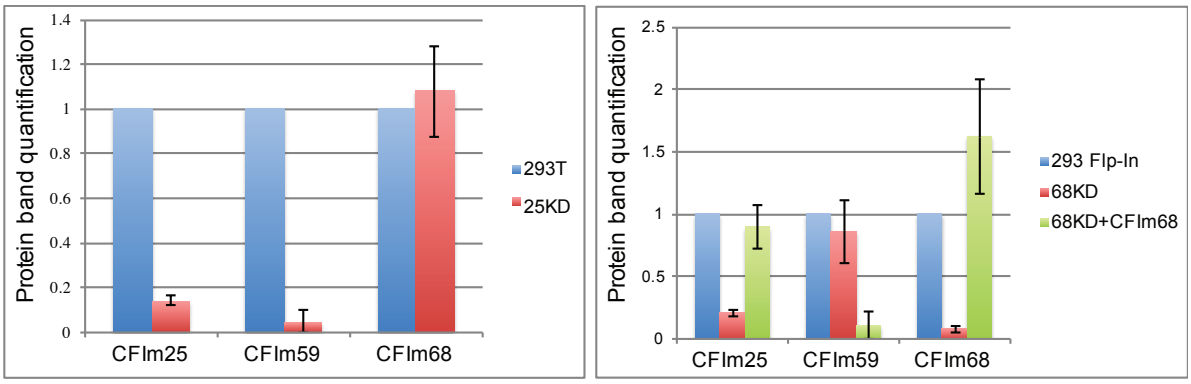


Figure 3.10: Effects of CFIm25/CFIm68 depletion on levels of other CFIm subunits. (A) Representative western blot showing CFIm expression levels. Whole-cell lysates were prepared from the given cell lines and subjected to western blotting against all 3 CFIm subunits, with α-tubulin acting as a loading control. (B) Table summarising the mean band fluorescence quantification across 4 experiments, normalised to α-tubulin and standardised to a value of 1 for the 293T or 293 Flp-In parental cell line as appropriate. (C) A bar chart representation of the mean quantification data from (B), split into 293T-derived cell lines (left) and 293 Flp-In-derived cell lines (right). Error bars represent ±SEM, n=4.

complete switch to proximal poly(A) sites, suggesting that there is a separable effect of CFIm68 knockdown. Secondly, expression of CFIm68 RS deletion mutants restored CFIm25 to wild type levels but did not complement the APA phenotype (see Chapter 4, Figure 4.14). These observations suggest that the effect seen in 68KD is CFIm68-specific, rather than a result of the co-depletion of CFIm25.

Finally, the effect of CFIm68 overexpression in the 68KD+CFIm68 stable cell line was investigated. While overexpression did not affect CFIm25 levels, it led to a substantial (89%) loss of CFIm59 (Figure 3.10A, Figure 3.10B right). This is a similar effect to that achieved with CFIm25 depletion, and may result from inhibition of CFIm25-CFIm59 complex formation due to increased incorporation of CFIm68 into CFIm.

In summary, these results indicate that there is clear homeostatic regulation between subunits of the CFIm complex – specifically, between small and large subunits. This could result from protein stability effects – for example, decreased availability of CFIm68 may lead to increased amounts of free, potentially monomeric CFIm25, which may be more vulnerable to proteolytic degradation. Alternatively, or in addition, there may be regulation at the transcriptional or post-transcriptional levels. The data presented here do not allow discrimination between these different possibilities, but this would be an interesting area for further study.

Chapter 4

Mechanistic determinants of CFIm activity

4.1 Introduction

The development of APA complementation assays based on CFIm25 and CFIm68 function (Chapter 3) provided a platform to compare the complementation abilities of altered CFIm25 and CFIm68 isoforms. This approach enabled various aspects of CFIm function to be explored, by assessing the contribution of specific domains and sequences within CFIm subunits to APA regulation.

The first aspect of the CFIm structure-function relationship investigated was the CFIm25-CFIm68 interaction interface. CFIm is thought to be a heterotetrameric complex, but there had not yet been direct demonstration that the interaction between small and large subunits within this heterotetramer was necessary for regulation of APA by CFIm. Here, the role of key residues in mediating the CFIm25-CFIm68 interaction was confirmed in vivo, and it was shown that loss of this interaction led to a loss of complementation in the APA assay. This suggests that CFIm complex formation is essential for normal regulation of APA.

Another aspect of CFIm function investigated was the importance of recognition of UGUA sequences on RNA by CFIm25. While UGUA is a known preferred binding sequence of CFIm25, the contribution of this sequence-specific recognition to APA regulation had not been fully explored in vivo. Here, a CFIm25 mutant lacking UGUA-binding residues showed a widespread failure to restore distal poly(A) site use, suggesting a central role for UGUA binding in CFIm function. There were, however, some transcript-specific exceptions to this general finding, and further study of such

transcripts may provide insights into non-UGUA-associated CFIm25 regulation at alternative poly(A) sites.

Experiments were also carried out to assess the contribution of the CFIm68 C-terminal RS domain to APA. This domain has been implicated in nuclear and subnuclear localisation of CFIm68 and interactions with other SR proteins, but a specific role in APA had not been explored. Deletion of portions of the RS domain led to impaired complementation, despite no obvious loss of CFIm68 nuclear localisation or interaction with other CFIm subunits. This suggests a specific role for the RS domain in regulation of poly(A) site selection by CFIm68.

Finally, experiments were performed to explore the mechanism of CFIm function more generally, as it still remains unclear whether the observed CFIm knockdown phenotype results from lost stimulation of cleavage at distal poly(A) sites, lost inhibition at proximal poly(A) sites, or both. Transcriptome-wide pol II ChIP-seq was applied to the parental and CFIm knockdown cell lines in order to search for potential transcriptional read-through in the knockdown lines, which may be expected if CFIm is required for efficient distal poly(A) site cleavage. No such read-through was detected, but a clear decrease in pol II occupancy at the transcription start site (TSS) was apparent, particularly in the 68KD line, suggesting that CFIm may be required for efficient transcription initiation. This raises the interesting possibility that effects of CFIm knockdown on APA may be, at least in part, attributable to alterations in transcription.

In summary, this chapter provides evidence that the CFIm25-CFIm68 interaction, UGUA recognition by CFIm25 and the CFIm68 RS domain all make key contributions to normal APA regulation, as well as providing the first analysis of the genome-wide transcriptional consequences of CFIm depletion.

4.2 The CFIm25 Y158A Y160A mutant is unable to interact with CFIm68 in vivo

As a first step to investigating the CFIm25-CFIm68 interaction, residues with a reported role in the interaction interface were identified from the literature. For CFIm25, two tyrosine residues, Y158 and Y160, had been identified as forming hydrogen bonds with the CFIm68 β_2/β_3 loop in a CFIm25-CFIm68 RRM crystal structure, and Y158A Y160A mutagenesis led to loss of the interaction in vitro (Yang et al., 2011a) – see Figure 4.1A.

To explore this interaction further, the 25KD cell line was used as a platform to investigate whether these in vitro findings also hold in vivo. Site-directed mutagenesis was used to generate a CFIm25 Y158A Y160A mutant using the previously generated pcDNA3.1-CFIm25 plasmid. This plasmid was transfected into the 25KD cell line alongside control transfections with the empty pcDNA3.1 plasmid and the wild type pcDNA3.1-CFIm25 plasmid, as described in Section 3.5. Western blotting confirmed that the Y158A Y160A mutant was expressed at similar levels to the wild type protein (Figure 4.1B).

CFIm25 was immunoprecipitated from the transfected cells and western blotting was used to assess the interaction with CFIm68. As can be observed from a representative blot (Figure 4.1B), less CFIm68 was immunoprecipitated by the Y158A Y160A mutant than by WT CFIm25. The CFIm68 band intensity in the Y158A Y160A IP appeared similar to the intensity in the empty plasmid IP, suggesting that this represented only CFIm68 immunoprecipitated by the low level of endogenous CFIm25 in the 25KD cell line, and that there was no further CFIm68 pull-down by the Y158A Y160A mutant.

Band quantification across three biological repeats confirmed that the quantity of CFIm68 immunoprecipitated per unit of CFIm25 Y158A Y160A was around half that immunoprecipitated by

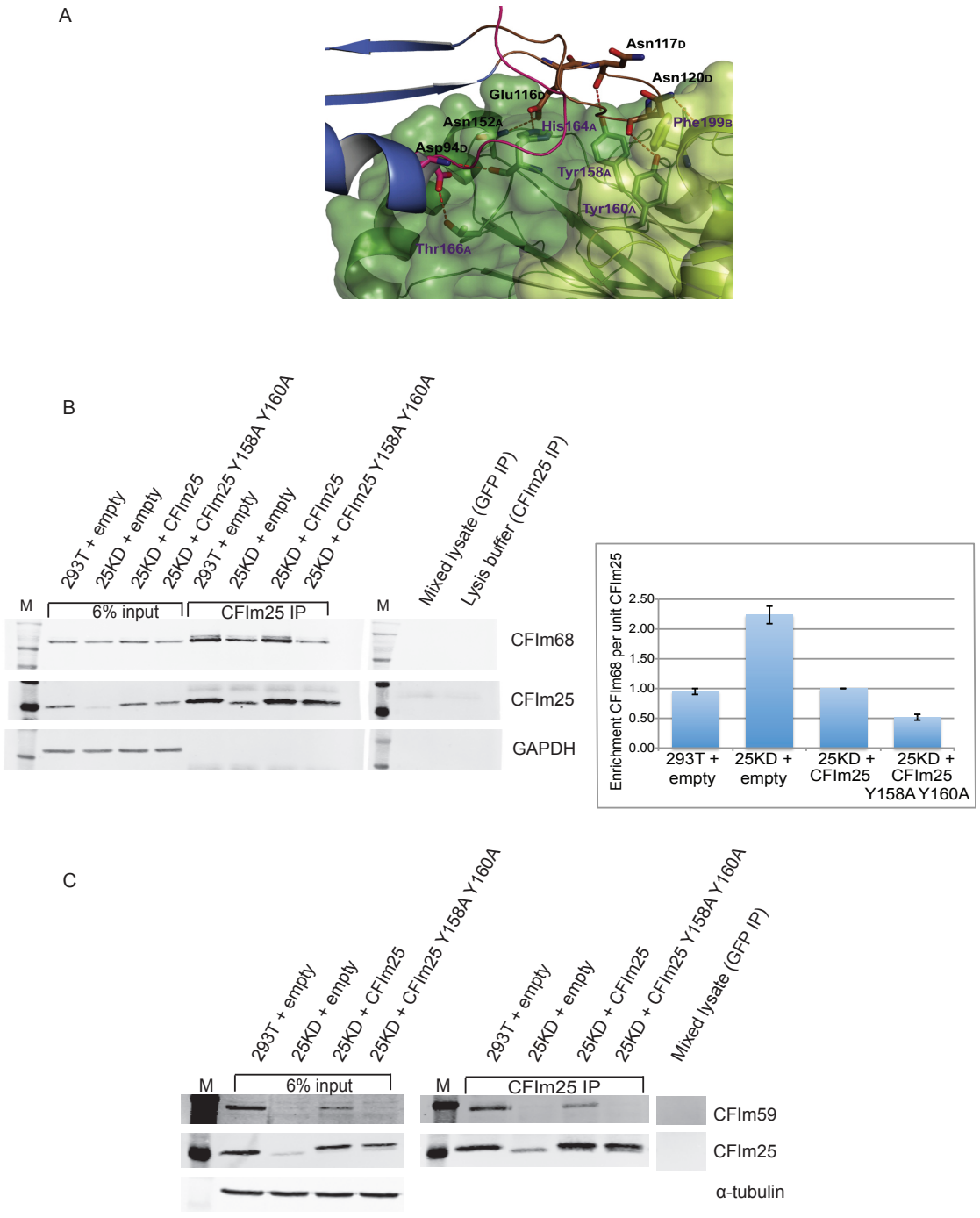


Figure 4.1: The CFIm25 Y158A Y160A mutant shows impaired interaction with CFIm68 and CFIm59. (A) Crystal structure showing interactions at the CFIm25-CFIm68 interface (CFIm25 in green; CFIm68 in blue/dark brown). Image reproduced with permission from: (Yang et al., 2011a) under a RightsLink licence. (B) Except for where otherwise stated, CFIm25 was immunoprecipitated from the indicated cells following a transfection experiment. Western blotting was performed with the antibodies indicated to quantify CFIm68 co-immunoprecipitation; GAPDH was used as a loading control for inputs. Mixed lysate GFP IP acted as an isotype control; the lysis buffer control confirmed the absence of CFIm25 antibody-specific light chain bands. M = marker lane. Graph - quantification of CFIm68 immunoprecipitation by CFIm25. CFIm68 IP band intensity was divided by CFIm25 IP band intensity, which was then further divided by the CFIm68 input band intensity (normalised to GAPDH). This yielded a value for enrichment of CFIm68 per unit of CFIm25 with reference to available input, and values were standardised to a value of 1 for the 25KD+CFIm25 sample. n=3, error bars represent ±SEM. (C) IP/western blot as described in (B), but probing for CFIm59 co-precipitation.

WT CFIm25 (Figure 4.1B). It might have been expected that this proportion would be much lower if Y158A Y160A were completely interaction-defective. However, this may be explained by apparent saturation of the CFIm25-binding capacity of CFIm68 for WT CFIm25, as evidenced by the greater than two-fold increase in CFIm68 enrichment per unit CFIm25 in the 25KD+empty sample, where CFIm25 is limiting. With CFIm68 saturation, for 25KD+empty and 25KD+CFIm25 Y158A Y160A, around half of the available CFIm68 would already be bound by the residual 'WT' CFIm25. In parental 293T cells and the 25KD+CFIm25 transfection, despite the four- to five-fold increase in immunoprecipitated WT CFIm25, there was only around a two-fold increase in CFIm68 pulled down as this represents the entire CFIm68 pool – thus, the increase in CFIm68 per unit CFIm25 is smaller than might be expected.

It was expected, given the high degree of sequence similarity between CFIm68 and CFIm59 and the similar overall architecture of CFIm25-CFIm59 and CFIm25-CFIm68 crystal structures (Yang et al., 2011b), that the CFIm25 Y158A Y160A mutation might also abolish the interaction with CFIm59. Although this was only tested in one experiment, it was readily observable that CFIm25 Y158A Y160A, by contrast to WT CFIm25, failed to immunoprecipitate CFIm59 detectably (Figure 4.1C). Interestingly, the loss of CFIm59 was also observed in the Y158A Y160A input sample by comparison with the WT input sample, suggesting that impaired incorporation of CFIm59 into CFIm may decrease its stability (see also Section 3.10). These data suggest, as anticipated based on structural data, that CFIm25 interacts with both large subunits via the same interface.

Overall, these results suggest that CFIm25 Y158A Y160A cannot interact with CFIm68 or CFIm59 in vivo, confirming in vitro predictions. However, this does not confirm that both Y158 and Y160 are necessary for the interaction – it may be that mutation of either Y158 or Y160 alone would have a similar effect. In addition, other residues are likely to make important contributions to the interaction - for example, crystal structure evidence shows H89 and F199 of CFIm25 forming contacts with the β_1/α_1 loop of CFIm68 (Li et al., 2011; Yang et al., 2011a), though only one of these studies showed

that H89A F199A mutagenesis abolished the CFIm25-CFIm68 interaction in vitro. It would therefore be of interest to assess the role of these residues in vivo.

4.3 The CFIm68 Δ116-122 deletion mutant is unable to interact with CFIm25 in vivo

Having confirmed the involvement of CFIm25 Y158/Y160 in the CFIm25-CFIm68 interaction, a similar approach was used to assess the contribution of the CFIm68 β_2/β_3 loop. In the crystal structure, this loop, between the 2nd and 3rd beta strands and comprising residues 116-122, forms several contacts with CFIm25. The main-chain carbonyl groups of N117 and N120 form hydrogen bonds with Y158 and Y160 respectively, while the side chain of E116 hydrogen bonds with H164 and N152 in CFIm25 (Yang et al., 2011a) – see Figure 4.1A. Based on these observations, site-directed mutagenesis was used to generate three CFIm68 internal deletion mutants using the pcDNA5/FRT-CFIm68 plasmid (see Section 3.9). One mutant, Δ116-122, lacked the whole β_2/β_3 loop, while the other two, ΔN117 and ΔN120, lacked only the single amino acids that participate in main-chain hydrogen bonding interactions with CFIm25. Cell lines stably expressing these mutant CFIm68 isoforms were generated as described in Section 3.9.

CFIm68 was immunoprecipitated from the stable cell lines expressing WT or mutant CFIm68 and western blotting was used to assess the interaction with CFIm25 (Figure 4.2A). CFIm68 was successfully expressed and immunoprecipitated from all stable cell lines and was absent from the control 68KD+empty IP as expected. It was readily apparent that the CFIm68 Δ116-122 loop deletion mutant was unable to interact with CFIm25, as the absence of a detectable CFIm25 band was comparable to the 68KD+empty line. Conversely, IP of the CFIm68 ΔN117 and ΔN120 mutants recovered CFIm25 to a similar extent as WT CFIm68. This suggests that while loss of the whole loop comprising amino acids 116-122 (or perhaps certain residues within this loop) prevents interaction

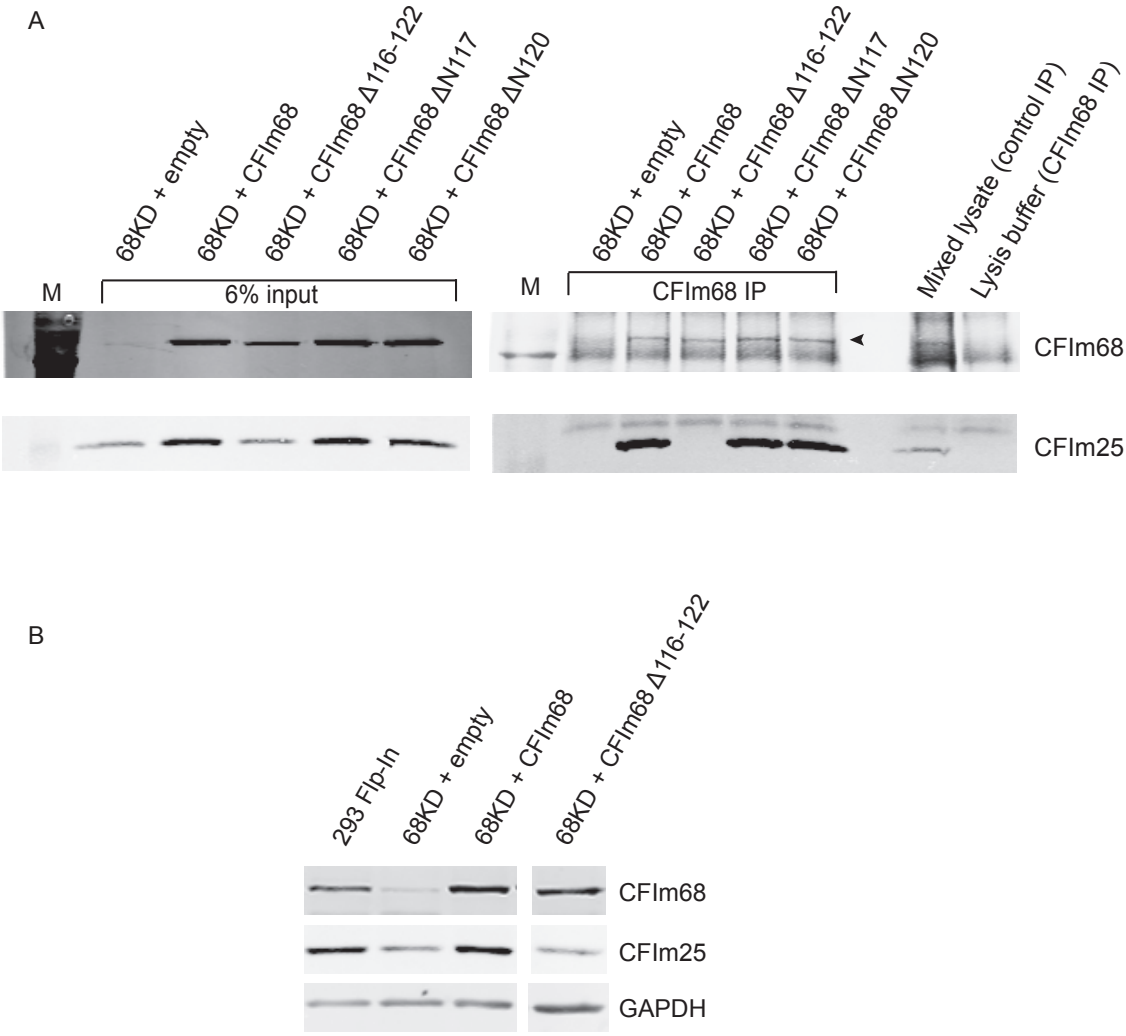


Figure 4.2: Deletion of residues 116-122 from CFIm68, but not of N117 or N120 alone, abolishes interaction with CFIm25. (A) Except for where otherwise stated, CFIm68 was immunoprecipitated from the given stable cell lines. Left panel: input samples; right panel: IP samples. Western blotting was performed with the given antibodies to quantify CFIm25 co-precipitation. The mixed lysate control IP used normal rabbit pre-immune IgG and acted as an isotype control; the lysis buffer control confirmed the absence of CFIm68 antibody-specific heavy chain bands. An arrowhead highlights the position of the CFIm68 band in the IP samples for clarity. n=2; this blot is representative of the observed results. Quantification was not performed due to the complete absence of a detectable CFIm25 band in the 68KD+empty and 68KD+CFIm68 Δ116-122 samples. (B) Western blotting of whole-cell lysate from the given cell lines, to allow a comparison of CFIm68 and CFIm25 expression levels with GAPDH loading control included.

with CFIm25, loss of N117 or N120 alone does not. It is likely that other residues compensate for loss of the main-chain N117/N120 hydrogen bonding interactions in this situation. Shortly after making these observations, a study was published reporting similar findings regarding the inability of CFIm68 Δ116-122 to interact with CFIm25 in vivo (Rasheedi et al., 2016).

It was interesting in itself that the 68KD+empty control IP showed no detectable CFIm68 or CFIm25, as it would be expected that the small amount of CFIm68 remaining in this cell line would be enriched by IP. This provides some indication that 68KD may in fact be functionally null, if the remaining protein is not available to interact with CFIm25. Equally, it was interesting that the low input level of CFIm25 in the Δ116-122 line was similar to that seen in the control line, despite the clear re-expression of CFIm68 (Figure 4.2A; Figure 4.2B – with loading control included). This suggests that where the CFIm25-CFIm68 interaction is impaired, re-expression of CFIm68 cannot reverse the concomitant reduction in CFIm25 levels, supporting the idea that complex formation promotes the stability of CFIm subunits (see also Section 3.10).

4.4 The CFIm25-CFIm68 interaction is necessary for normal regulation of APA by both subunits

Having generated and validated CFIm25 and CFIm68 mutants that fail to interact with each other, the complementation assay was then used to assess whether the CFIm25-CFIm68 interaction is necessary for these subunits to function normally in APA regulation.

First considering the CFIm25 Y158A Y160A mutant, transfection experiments were carried out in the same way as for the IP experiments described in Section 4.2 and RNA was analysed by northern blotting. The Y158A Y160A mutant was unable to complement the APA shift for either transcript, with the APA profile remaining similar to that seen with the 25KD+empty transfection (Figure 4.3A).

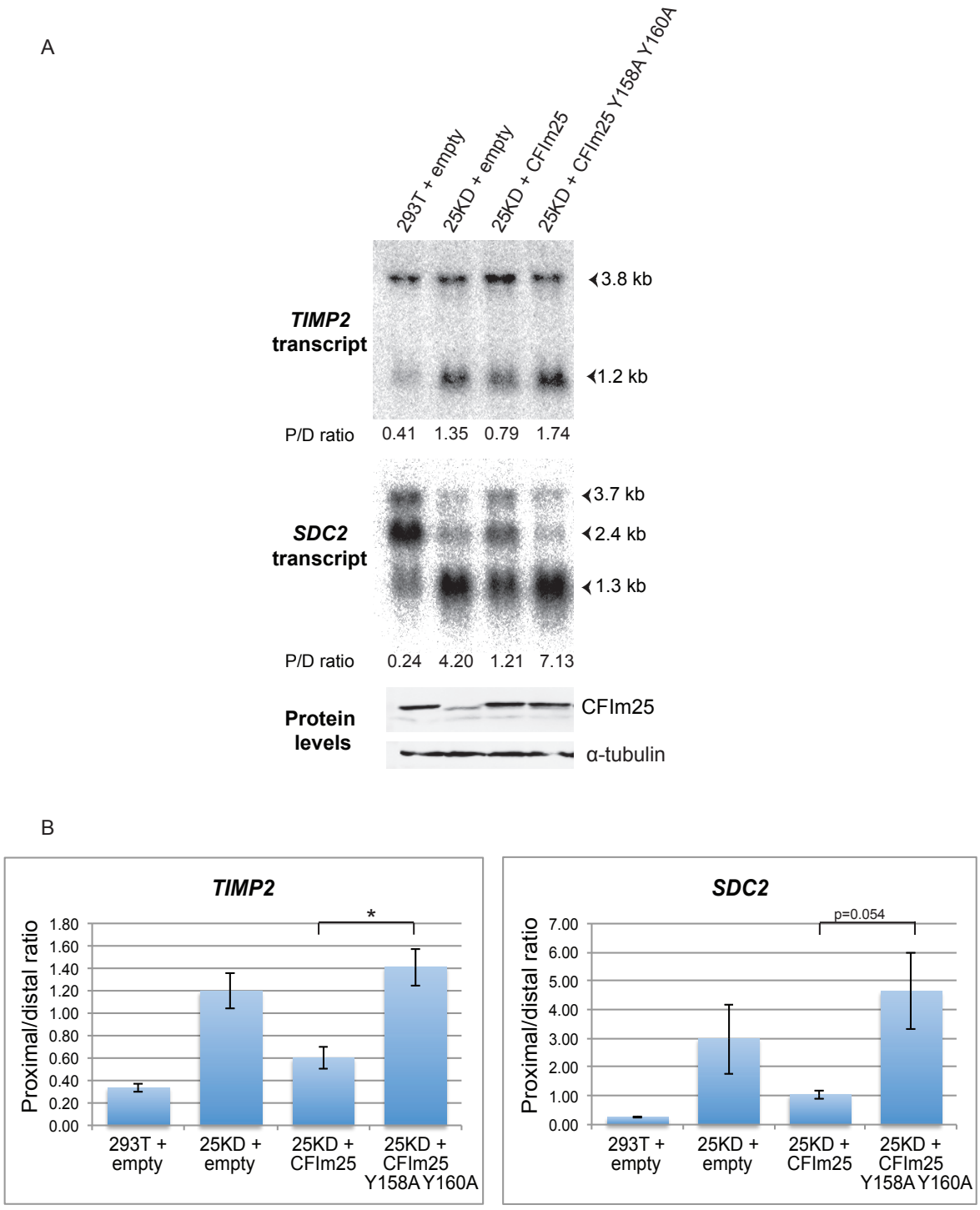
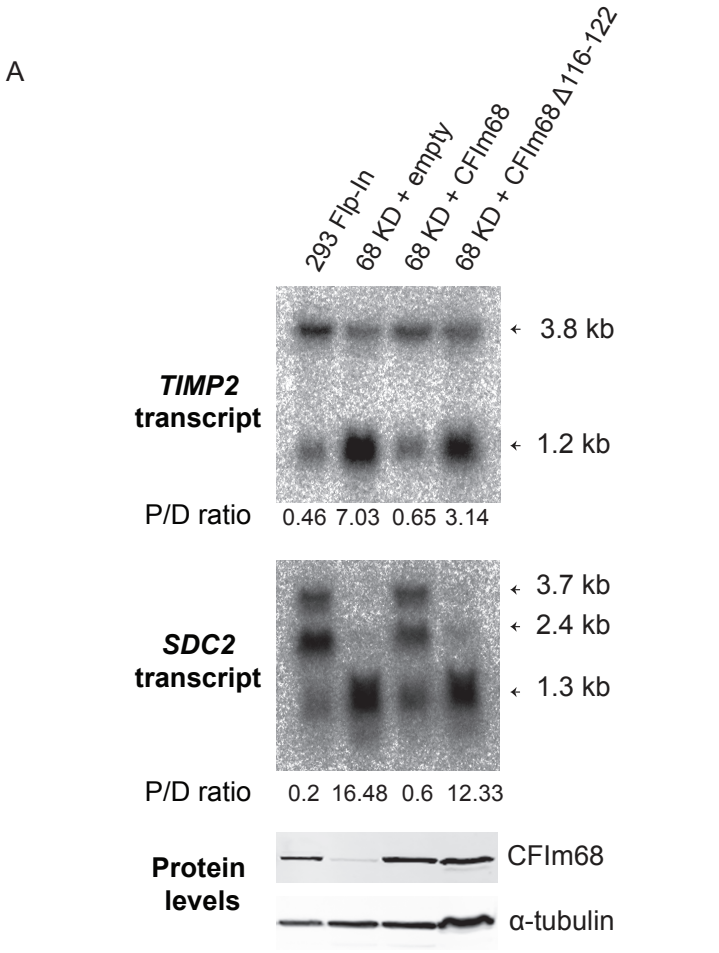


Figure 4.3: The CFIm25 Y158A Y160A mutant has an impaired ability to function in APA regulation. (A) Northern blotting and P/D ratio calculations were performed on the given samples following a transfection experiment as previously described (Figure 3.3). The corresponding western blot at the bottom shows CFIm25 expression in the transfected clones, with α-tubulin as a loading control. (B) Mean P/D ratios from repeats of this analysis (n=3). Error bars depict ±SEM and statistical significance of the different between WT and Y158A Y160A, calculated using a two-tailed Student's t-test, is illustrated with * symbols.

Quantification across three experiments (Figure 4.3B) confirmed that there was a significant difference between the *TIMP2* proximal/distal band ratio of the wild type and Y158A Y160A mutant ($p<0.05$), although for *SDC2* this comparison narrowly missed the significance threshold ($p=0.054$). While further repeats would be necessary to confirm that this mutant significantly differs from the wild type in APA function for *SDC2*, these results suggest that when CFIm25 is unable to interact with the large CFIm subunits, it is unable to function in APA regulation of these two transcripts. It is possible that this impairment results from abolished interaction with CFIm59, CFIm68 or both.

A similar complementation analysis was performed on the 68KD+CFIm68 Δ116-122 stable cell line, comparing it to the CFIm68 WT and negative control lines. This revealed a striking inability of the Δ116-122 mutant to complement the APA shift for both transcripts, with the profile remaining similar to that seen with the 68KD+empty line (Figure 4.4A). The difference here was more readily apparent than for the CFIm25 Y158A Y160A mutant, due to the more complete APA shift and complementation in this system. Quantification across two experiments confirmed that there was a significant difference between proximal/distal band ratios for the wild type and Δ116-122 mutant ($p <0.05$ for *TIMP2* and *SDC2*, Figure 4.4B). This suggests that if CFIm68 is unable to interact with CFIm25, it is unable to function in poly(A) site selection.

Overall, these results suggest that both CFIm25 and CFIm68 need to be incorporated into the CFIm complex in order to act in poly(A) site selection, and that they are unable to function independently in this process. It would be important to confirm the generality of these results at the transcriptome-wide level, which could be achieved using the QuantSeq approach as discussed in Section 3.5. However, it seems likely that these two APA-regulated transcripts are representative of a general requirement for CFIm complex formation. It is also worth noting that while the loss of interaction seems the most likely explanation for impaired function here, it cannot be ruled out that some other aspect of CFIm function aside from the CFIm25-CFIm68 interaction may have been impaired by these mutations. In addition, in the case of the CFIm25 Y158A Y160A mutation, the fact that interactions with both



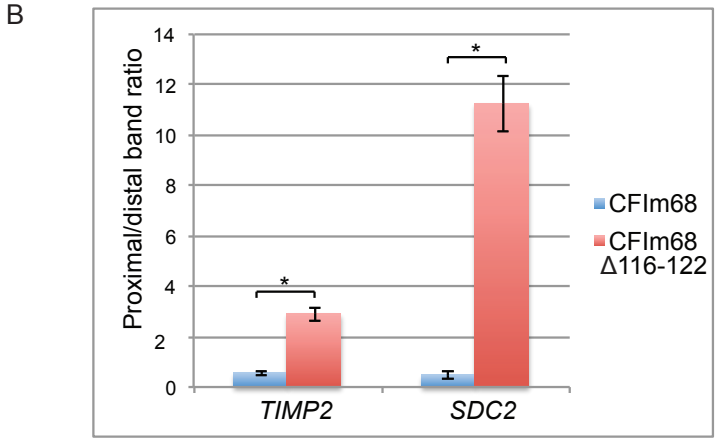


Figure 4.4: The CFIm68 Δ116-122 mutant has an impaired ability to function in APA regulation. (A) Northern blotting and P/D ratio calculations were performed on the given stable cell lines as previously described (Figure 3.3). The corresponding western blot shows CFIm68 expression in the cell lines, with α-tubulin as a loading control. (B) Mean P/D ratios (showing WT/mutant re-expression samples only) from biological repeats of this analysis (n=2). Error bars depict ±SEM and statistical significance, calculated using a two-tailed Student's t-test, is illustrated with * symbols.

CFIm59 and CFIm68 are abolished means that the loss of interaction with either or both of these subunits may explain the phenotype. However, the similar phenotype observed with the CFIm68 Δ116-122 mutant and the lack of a previously reported role for CFIm59 in APA suggests that the CFIm25-CFIm68 interaction is more likely to be responsible.

The finding that CFIm25 and CFIm68 appear to only function in APA when incorporated into the CFIm complex may seem intuitive, but had not been formally demonstrated before. Although CFIm has long been known to purify as a complex, it remained possible that CFIm25 and CFIm68 may be able to function independently of each other to regulate poly(A) site selection. This work therefore provides important clarification that CFIm25 and CFIm68 show interdependency, and co-operate with each other as part of the full CFIm complex to regulate APA.

4.5 Recognition of UGUA sequences by CFIm25 plays a key role in regulation of APA transcriptome-wide

As well as inter-subunit interaction within the complex, another aspect of CFIm likely to make a crucial contribution to its functionality is its RNA binding. CFIm25 was identified as the major RNA-binding subunit based on in vitro crosslinking and structural evidence, although CFIm68 may also make some contribution to binding efficiency (Dettwiler et al., 2004). SELEX studies identified UGUAN (N = A > U ≥ G/C) as the major sequence bound by CFIm25 (Brown and Gilmartin, 2003; Venkataraman et al., 2005), and structural evidence and mutagenesis studies have identified key residues of CFIm25 involved in the sequence-specific recognition of this motif (Yang et al., 2010) – see Figure 4.5.

Although the CFIm25-UGUA interaction is well defined, the importance and role of this in APA regulation in vivo has not been fully explored. UGUA sequences are enriched around 40-60 nt

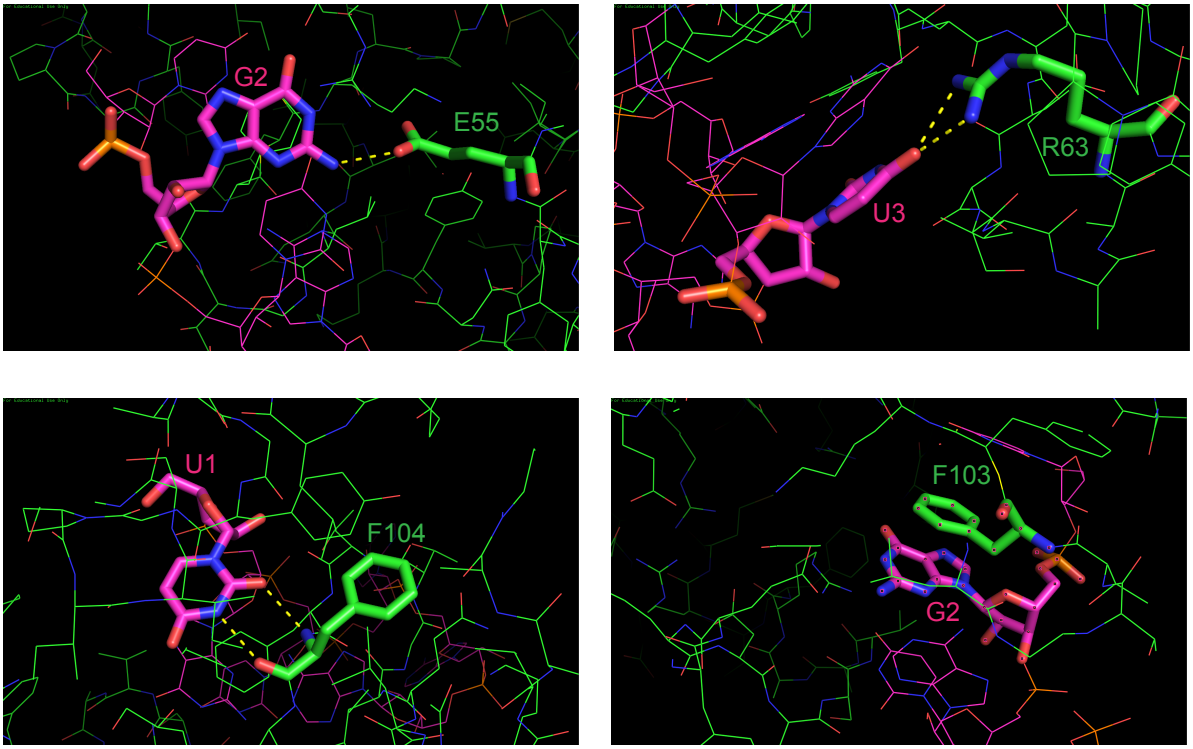


Figure 4.5: Visualisation of CFIm25-UGUA interactions in a solved crystal structure. The structural data from a solved CFIm25-UGUA RNA crystal structure (Yang et al., 2010a) was downloaded from the protein data bank (http://www.rcsb.org/pdb/). The relevant PDB ID is 3MDI, and the structure was visualised using the MacPyMOL programme. Residues of interest are highlighted here as thicker sticks for clearer visualisation - green sticks are CFIm25 residues as labelled and pink sticks are nucleotides within the UGUA sequence as labelled (U1,G2,U3,A4). Hydrogen bonding interactions were identified by the software and are shown as dotted yellow lines.

upstream of many dominant poly(A) sites, with co-incident peaks of CFIm binding at such sites, supporting a model whereby CFIm is recruited upstream of dominant (often distal) poly(A) sites by UGUA binding and supports their use (Li et al., 2015; Martin et al., 2012). However, CFIm can also be recruited at a similar position at other dominant cleavage sites that lack UGUA sequences, and some CFIm-regulated transcripts, such as *RNF11*, do not have a well-positioned UGUA at the distal site (Martin et al., 2012). Therefore, it appears that a more complex model, based on more than just the absence/presence of UGUA near different poly(A) sites, may be needed to fully understand CFIm RNA binding and function.

In order to investigate the contribution of UGUA recognition to CFIm25-regulated APA, site-directed mutagenesis was used to substitute key residues involved in this interaction. The main mutant studied was CFIm25 E55A R63A F104A, which was predicted to lose specific hydrogen bonding interactions with G2, U3 and U1 on the RNA respectively (Yang et al., 2010) – see Figure 4.5. While loss of RNA binding for this mutant was not confirmed here, previous EMSA studies with a UGUA-containing short RNA confirmed that mutation of E55 and R63 led to a decrease in RNA binding of 88% and 99% respectively (Yang et al., 2010). CFIm25 F103A was also studied for comparison. While F103 does not participate in direct UGUA recognition, it provides stacking interactions to stabilise the RNA (Figure 4.5), and F103A mutation led to a 99% loss in RNA binding as measured by EMSA (Yang et al., 2010). These mutants were generated, transfected and analysed in the complementation assay as previously described in Section 4.2.

The F103A and E55A R63A F104A mutants behaved similarly in the complementation assay (Figure 4.6A), supporting the idea that the observed phenotypes resulted from impairment of UGUA binding. However, clear conclusions can only be drawn for E55A R63A F104A, which was analysed across repeat experiments. For the *SDC2* transcript, the mutants showed a clear impaired complementation phenotype (Figure 4.6A), similar to that seen for the Y158A Y160A interaction mutant (Figure 4.3A). In contrast, for *TIMP2*, the mutants clearly showed some degree of complementation, which appeared

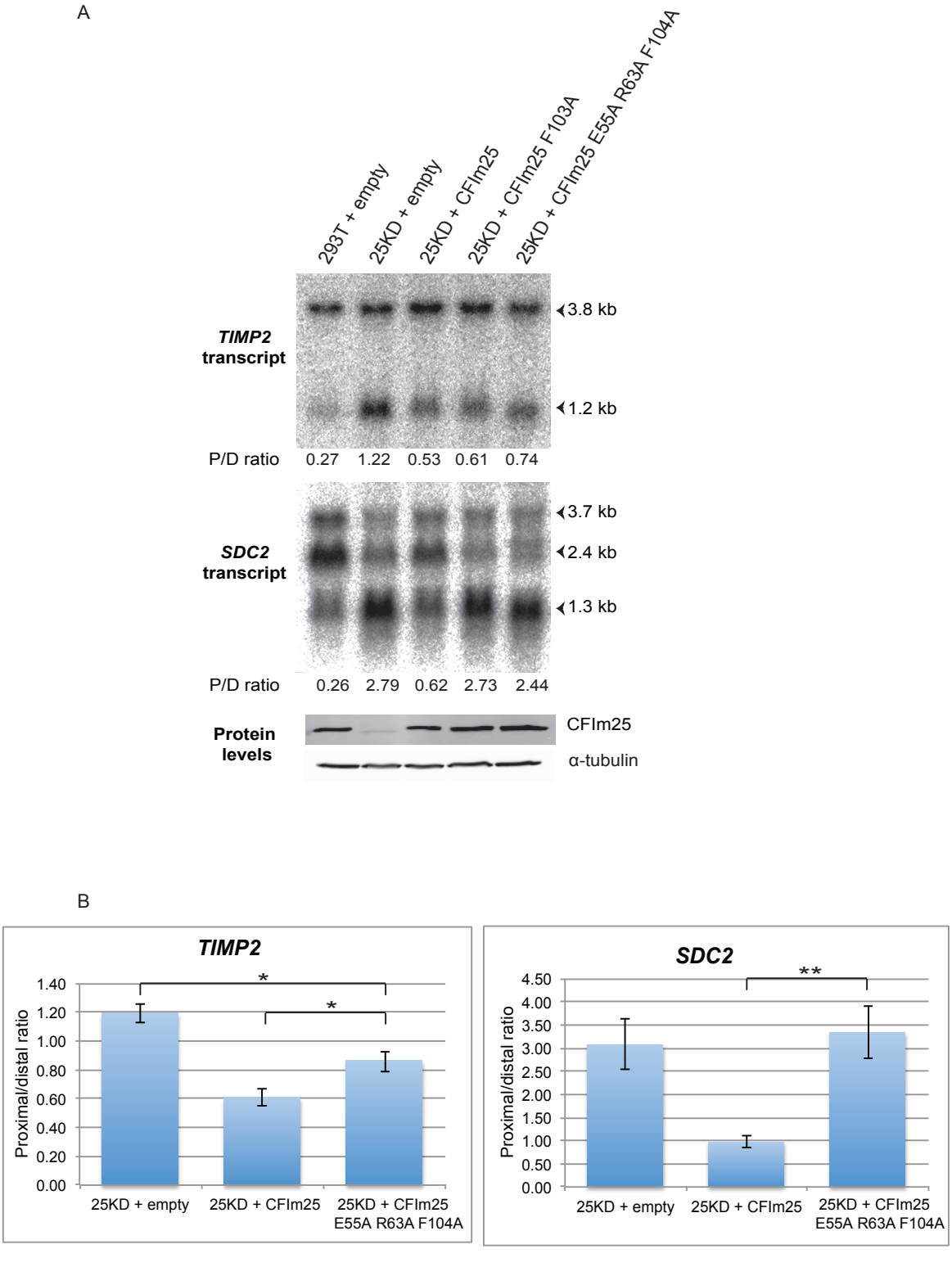


Figure 4.6: CFIm25 mutants unable to bind UGUA sequences show transcript-specific impaired ability to function in APA regulation. (A) Northern blotting and P/D ratio calculations were performed on the indicated samples following a transfection experiment as described in Figure 3.3. The corresponding western blot shows CFIm25 expression in the transfected clones, with α-tubulin as a loading control. (B) Mean P/D ratios from repeats of the analysis for the E55A R63A F104A mutant (n=5). Error bars depict ±SEM and statistical significance, calculated using a two-tailed Student's t-test, is illustrated with * symbols.

similar to WT complementation (Figure 4.6A). Quantification across five experiments for the E55A R63A F104A mutant confirmed the difference in response between the two transcripts (Figure 4.6B). For *SDC2*, 25KD+CFIm25 E55A R63A F104A gave a similar profile to 25KD+empty, and showed a significant impairment in complementation compared to WT CFIm25 ($p<0.01$). For *TIMP2*, however, E55A R63A F104A showed a significant difference to both 25KD+empty ($p<0.05$) and 25KD+CFIm25 ($p<0.05$), demonstrating an intermediate complementation profile. This suggests that UGUA binding is absolutely necessary for CFIm25 to regulate poly(A) site selection in the *SDC2* transcript, but only makes a partial contribution to the process for *TIMP2*.

To increase confidence that the observed results were indeed a read-out of impaired UGUA binding by the mutant, it was important to confirm that complex formation was not impaired. A CFIm25 IP was performed in cells transfected with CFIm25 WT or E55A R63A F104A, and western blotting was used to assess co-precipitation of the other subunits (Figure 4.7). This confirmed that CFIm25 E55A R63A F104A was able to co-precipitate CFIm59 and CFIm68 as efficiently as the wild type protein, suggesting that observed complementation defects could not be attributed to impaired complex formation and were thus more likely to reflect impaired UGUA recognition.

The intriguing transcript-specific difference in UGUA-dependency was explored further. One obvious factor that may explain such a result is differences in distribution of UGUA sequences between the two transcripts, so a search was carried out to identify UGUA sequences in the vicinity of the *TIMP2* and *SDC2* poly(A) sites. Based on the known binding position of CFIm around 40-60 nt upstream of poly(A) sites, the search was restricted to the 100 nt upstream of each poly(A) site (Figure 4.8). There were clear similarities in the distribution of UGUA sequences (TGTA at the DNA level) between the two transcripts – the most proximal poly(A) sites both lack upstream TGTA sequences, while the more distal poly(A) sites all have a TGTA around 50 nt upstream (Figure 4.8). These findings are consistent with previous reports of typical TGTA distributions for transcripts showing CFIm-regulated APA (Li et al., 2015), but do not readily explain the reason that *TIMP2* and *SDC2* showed different

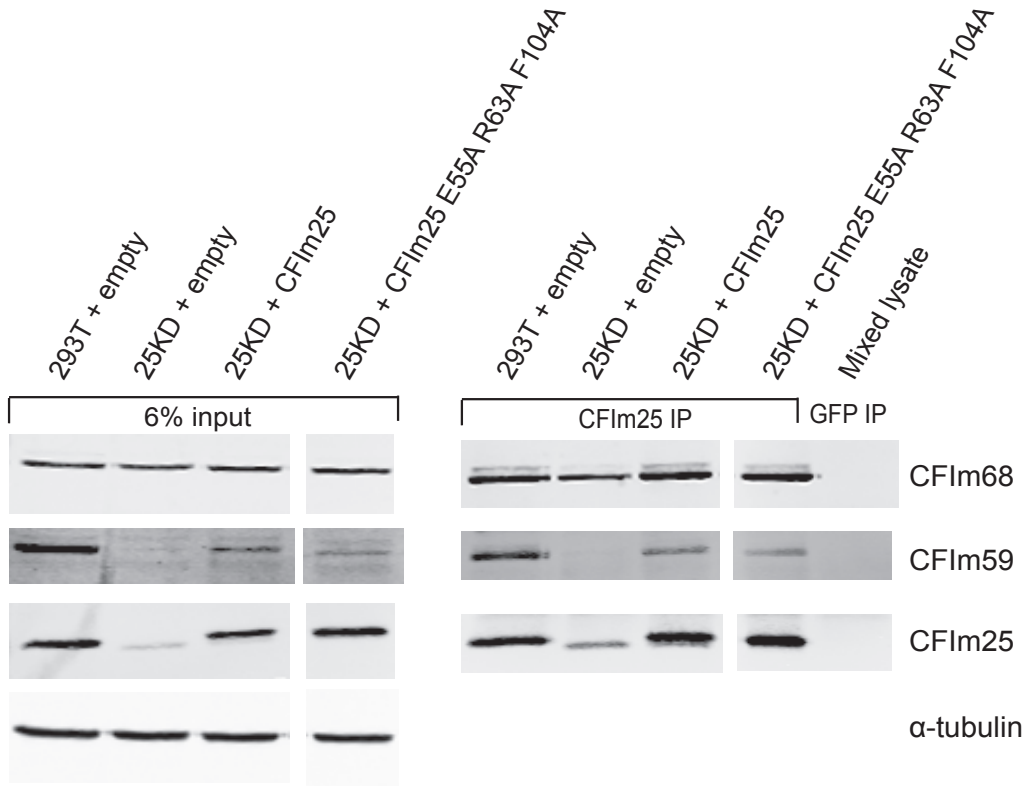


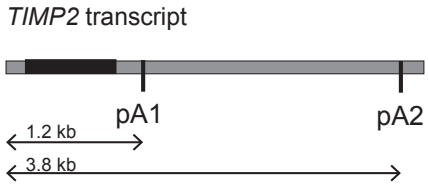
Figure 4.7: The CFIm25 E55A R63A F104A is able to interact with CFIm68 and CFIm59. Except for where otherwise stated, CFIm25 was immunoprecipitated from the given cells following a transfection experiment. Western blotting was performed with the given antibodies to analyse co-precipitation of other CFIm subunits. α-tubulin was used as a loading control for input samples. The mixed lysate GFP IP acts as an isotype control.

extents of UGUA-dependency. The only notable difference between the transcripts was an additional TGTA sequence at position -75 for pA2 of *SDC2*, which may perhaps reflect an increased reliance of this site on UGUA-mediated CFIm25 recruitment. However, the two most distal sites (*TIMP2* pA2 and *SDC2* pA3) show a near-identical positioning of a single TGTA relative to the PAS and poly(A) site, but while *TIMP2* is able to significantly recover use of this site in the presence of UGUA binding-defective CFIm25, *SDC2* is not. Overall, this analysis suggested that the number and position of UGUA sequences relative to poly(A) sites is not sufficient to predict the importance of UGUA-mediated CFIm25 recruitment.

This analysis only considered two transcripts, but highlighted an interesting difference in their reliance on the UGUA-binding capacity of CFIm25. This gave rise to a question concerning the generality of these findings – do the majority of CFIm25-dependent APA events rely on the UGUA-binding capacity of CFIm25? In order to assess this, the transcriptome-wide QuantSeq method, introduced in Section 3.5, was used to compare poly(A) site usage upon re-expression of WT CFIm25 and CFIm25 E55A R63A F104A (Figure 4.9). Two independent experiments were carried out, as part of the experiments previously presented in Figure 3.5 (Run 1 also included comparisons shown in 3.5A,B and C and Run 2 included the comparison in 3.5D).

By comparison to WT CFIm25, the CFIm25 E55A R63A F104A mutant showed a clear increased relative usage of proximal poly(A) sites, reflecting a widespread inability to complement (Figure 4.9A). On average, 86% of transcripts showing a significant difference in APA between the two conditions showed increased proximal site usage for the E55A R63A F104A mutant. This suggests that most transcripts subject to CFIm25-dependent APA are highly dependent on the UGUA-binding capacity of CFIm25 for this regulation, as was observed for *SDC2*. There was significant overlap between the identities of transcripts identified in the two runs (Figure 4.9B), suggesting a reasonable consistency in the response of this group of UGUA-dependent transcripts, with the lack of a more

A



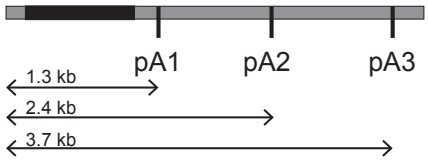
***TIMP2* pA1**

CTGTGGCCAACTGCAAAAAAAGCCTCCAAGGGTTTCGACTGGTCCAGCTCTGACATCCCTTCCTGG
AAACAGCATG**AATAAA**ACACTCATCCCATGGGTC

***TIMP2* pA2**

AGTTGTTGAAAGTTGACAAGCAGACTGCGCATGTCTCTGATGCTT**TGTA**TCATTCTTGAGCAATCG
CTCGGTCCGTGGAC**AATAAA**CAGTATTATCAAAG

B



***SDC2* pA1**

TATCAGAAGGCACCTACTAAGGAGTTTTATGCGTAAAACTCCAACTTAGTGTCTCTATTTATGAGATC
ACTGAACTTTTCAA**AATAAA**GCTTTTGCATAG

***SDC2* pA2**

CACAGACAATTCTAATGGTAACTTT**TGTA**GTCTTATGAATAGACATAAAT**TGTA**ATTTGGGAACATAAA
AACTACTG**AATAAA**TCATGTGGCCTAATATT

***SDC2* pA3**

ATCCGTGCCAGGAAAAGAAAAATTTCTGGCAAATATTTTGTCACTGC**TGTA**AAGCAAAATATTTGTGA
AAGTGCCAA**AATAAA**GTCTGTCATGCCAAAAG

Figure 4.8: Distribution of UGUA sequences upstream of *TIMP2* and *SDC2* poly(A) sites. The average cleavage positions for each poly(A) (pA) site in the *TIMP2* and *SDC2* transcripts were identified from the 3' end sequencing data and cross-referenced with the APASdb online database (http://genome.bucm.edu.cn/utr/). The 100 nucleotides of sequence upstream of each site was selected for analysis. In each case, the PAS is highlighted in blue and any UGUA sequences present (TGTA at the DNA level) are highlighted in red. Numbers refer to distance away from the pA site (sequence end). (A) Illustration of the *TIMP2* transcript (not to scale) and sequence upstream of its pA sites. (B) As in (A) for the *SDC2* transcript.

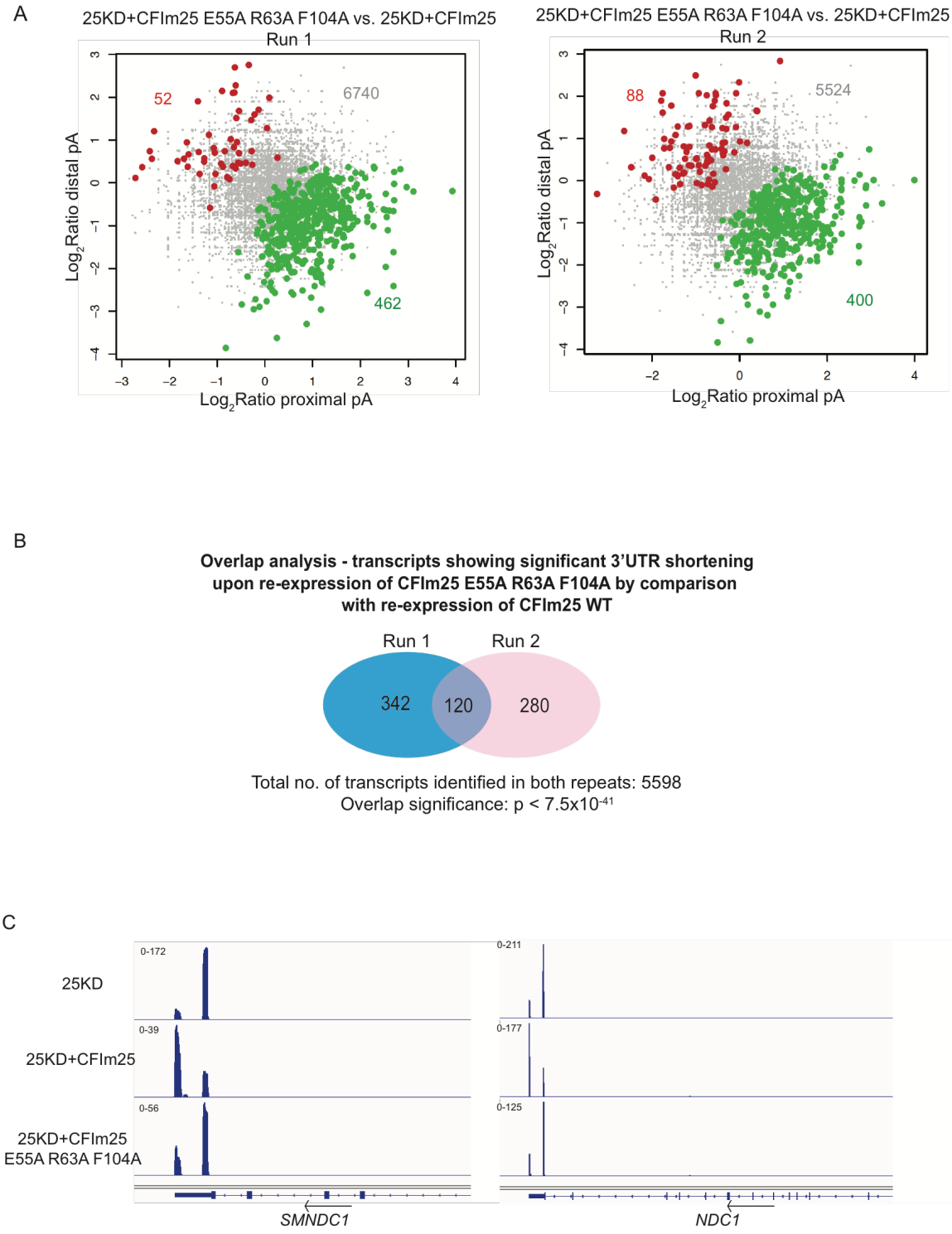


Figure 4.9: Transcriptome-wide analysis of APA upon re-expression of a CFIm25 UGUA-binding mutant. Total RNA was isolated from given samples and preparation of 3'end sequencing libraries and data analysis was carried out as described in Figure 3.5. (A) Results from two biological repeats (Run 1/Run2) of a transfection experiment and sequencing analysis with CFIm25 WT and E55A R63A F104A (UGUA-binding mutant). Plotting of transcript dots is as described in Figure 3.5. (B) Venn diagram of overlap between transcripts showing 3'UTR shortening upon CFIm25 E55AA R63A F104A re-expression by comparison with CFIm25 WT re-expression in the two runs (green dots in (A) and (B)). Overlap significance was calculated as described in Figure 3.5. (C) Screenshots from the Integrative Genomics Viewer showing poly(A) site usage (blue peaks representing reads) in two individual transcripts with significantly decreased relative use of distal poly(A) sites upon expression of E55A R63A F104A compared to expression of WT (green dots in both runs in (A)). Traces are from Run 2, but are representative of observations for both runs.

complete overlap perhaps resulting from randomness in the system and the significance cut-off used, as discussed in Section 3.5.

Inspection of poly(A) site usage for individual transcripts in the dataset allowed visualisation of this mode of regulation. *SMNDC1* and *NDC1* (Figure 4.9C) are two examples of transcripts that showed a significant and reproducible restoration of distal poly(A) site use upon re-expression of WT CFIm25 but not upon re-expression of CFIm25 E55A R63A F104A, and thus could be described as 'UGUA-dependent'.

Consistent with the northern blotting data (Figure 4.6), by comparison to WT CFIm25, E55A R63A F104A showed a significant inability to restore distal poly(A) site use for *SDC2* in both repeats (Figure 4.10A). The difference was much more readily observable in Run 2, where the complementation by CFIm25 WT was considerably more effective. Although the algorithm used here only considers the two most-used poly(A) sites, and therefore is not entirely appropriate for analysing *SDC2*, which has three major sites, the changing use of the two most-used sites was sufficient to highlight a significant difference in APA.

The data for *TIMP2*, however, did not agree well with the findings from northern blotting (Figure 4.10B). Different results were obtained between the two replicates, with Run 2 showing no significant APA difference between E55A R63A F104A and the CFIm25 WT or 25KD samples. Run 1 showed a significant complementation impairment for E55A R63A F104A compared to CFIm25 WT, but also found that there was a significant 3'UTR shortening in this mutant compared even to the 25KD sample (i.e. a complete failure to complement, as for *SDC2*). This was inconsistent with the northern blotting data, which indicated that expression of the E55A R63A F104A mutant led to an APA profile intermediate between those of 25KD and CFIm25 WT (Figure 4.6). It is not clear why these results did not concord, but this highlights the importance of performing repeat runs of this analysis, given the transient nature of re-expression achievable in this system.

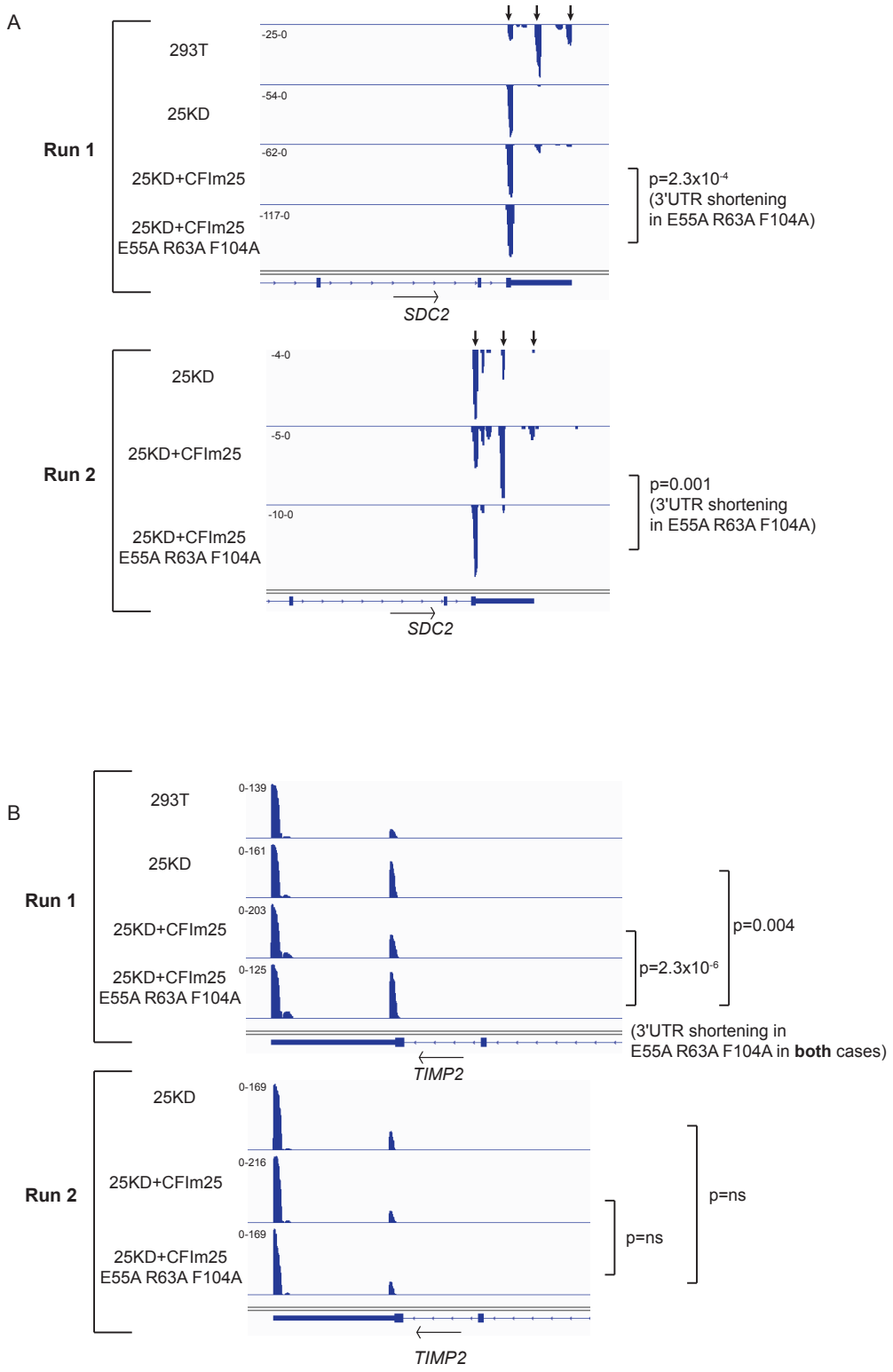
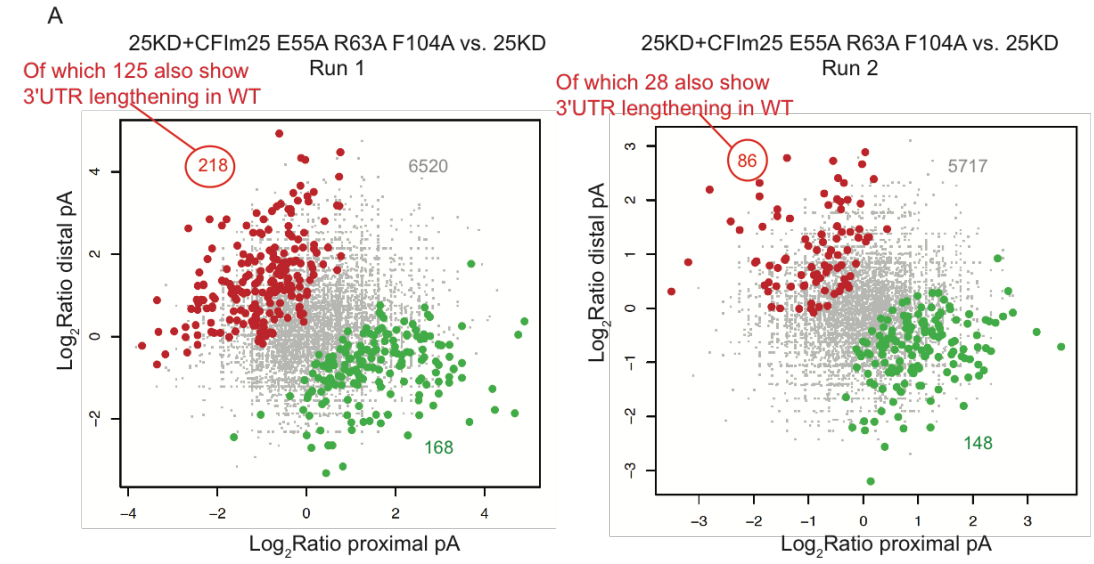


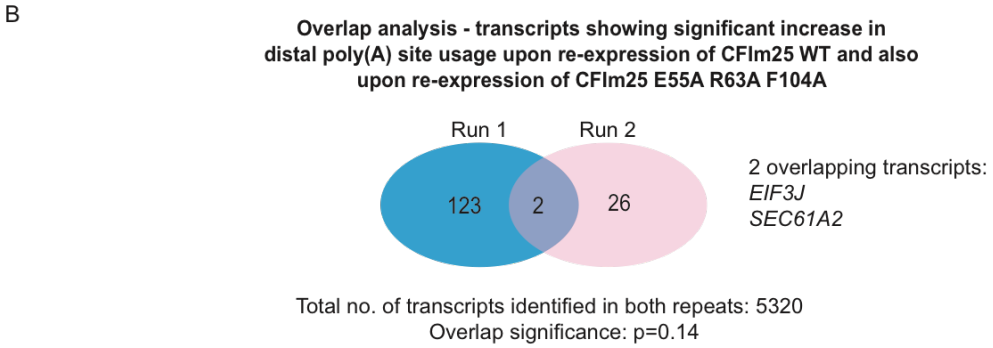
Figure 4.10: QuantSeq analysis of poly(A) site usage in the *SDC2* and *TIMP2* transcripts. Integrative Genomics Viewer (IGV) was used to view poly(A) site usage data for individual transcripts from the two QuantSeq experiments presented in Figures 3.5 and 4.9. (A) Traces for the *SDC2* transcript from Run 1 (included a 293T parental sample) and Run 2 (no 293T sample), with blue peaks representing mapped reads (reads per million, rpm). Arrows highlight the three poly(A) sites visible by northern blotting. Significance values are as calculated by Fisher's exact test in the 3'READS pipeline. (B) As in (A), for the *TIMP2* transcript.

While *TIMP2* did not behave as expected in the QuantSeq analysis, the potential of this dataset to reveal CFIm25-dependent transcripts which show incomplete or absent UGUA-dependence was still of interest. Identification and analysis of such transcripts could potentially improve understanding of how CFIm may be recruited to poly(A) sites in the absence of UGUA sequences, which may have important regulatory implications.

In both datasets, a group of apparently UGUA-independent transcripts were identified – specifically, those that showed significant 3'UTR lengthening both upon re-expression of CFIm25 WT and upon re-expression of CFIm25 E55A R63A F104A (125 transcripts for Run 1 and 28 transcripts for Run 2; Figure 4.11A). However, the overlap between transcript identities from these two sets was poor, with only two transcripts, *EIF3J* and *SEC61A2*, identified as UGUA-independent in both (Figure 4.11B). This suggests considerable randomness in UGUA-independent restoration of distal poly(A) site use, which may again result from the variability of the system as discussed in Section 3.5. The ability to detect a robust set of reproducibly UGUA-independent transcripts was probably further diminished by the small numbers of transcripts involved here. Increasing the significance cut-off to $p<0.05$ only increased the overlap to six transcripts, which still does not provide a large enough set to perform motif analysis or draw any generalised conclusions.

Manual inspection of the two transcripts identified as UGUA-independent in both sequencing runs, *EIF3J* and *SEC61A2*, confirmed the effective restoration of distal poly(A) site use by both CFIm25 WT and E55A R63A F104A (Figure 4.11C). Interestingly, both of these genes have neighbouring convergent genes on opposing strands, though whether this has any relevance to CFIm25-dependent regulation is unclear. A sequence analysis of the 3'UTR regions upstream of poly(A) sites in these two transcripts revealed some differences in UGUA distribution (Figure 4.12). *SEC61A2* has a TGTA positioned upstream of the distal poly(A) site (Figure 4.12A), similarly to the UGUA-dependent *SDC2* (Figure 4.8B), although this is further away from the pA site (-77 bp), which could potentially reflect a diminished role for CFIm-mediated cleavage factor recruitment at this site. *SEC61A2* also has a TGTA





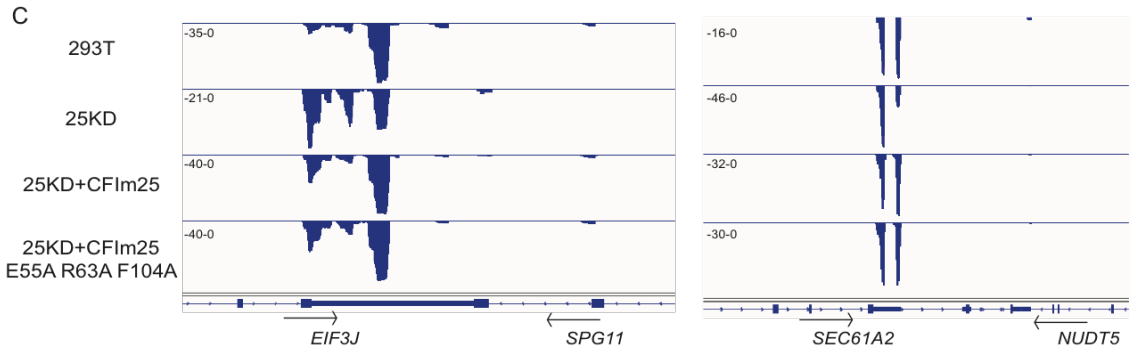


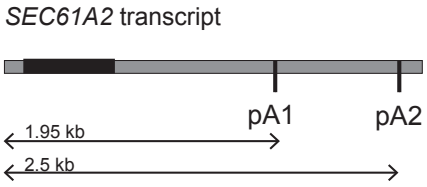
Figure 4.11: Identification of CFIm25-dependent transcripts showing reduced or absent UGUA-dependence. Total RNA was isolated from given samples and preparation of 3'end sequencing libraries and data analysis was carried out as described in Figure 3.5. (A) Results from two biological repeats (Run 1/Run2) of a transfection experiment and sequencing analysis with pcDNA3.1 vs. pcDNA3.1-CFIm25 E55A R63A F104A (UGUA-binding mutant). Plotting of transcript dots is as described in Figure 3.5. (B) Venn diagram of overlap between the two runs of transcripts showing 3'UTR lengthening upon CFIm25 WT and CFIm25 E55A R63A F104A re-expression by comparison with control empty plasmid (selected set of red dots in (A)). Overlap significance was calculated as described in Figure 3.5. (C) Screenshots from the Integrative Genomics Viewer showing poly(A) site usage (blue peaks representing (reads per million, rpm) in the two transcripts showing 'UGUA-independent CFIm25-dependence' in both runs. Traces are from Run 1, but are representative of observations for both runs.

immediately following the end of the PAS sequence upstream of the proximal poly(A) site, suggesting potential stimulatory or suppressive UGUA-mediated CFIm25 regulation (Figure 4.12A). For *EIF3J*, there was an interesting absence of UGUA sequences upstream of pA2 and pA3, which would be a convincing explanation for the UGUA-independence observed for this particular transcript (Figure 4.12B).

These results suggest that a lack of dependency of individual transcripts on the UGUA-binding capacity of CFIm25 may be partly predictable based on the absence of UGUA sequences or their increased distance away from distal poly(A) sites. However, only two transcripts have been analysed here, and the principles governing UGUA-dependence are likely to be more complex. Ideally, searches for other features common to UGUA-independent transcripts would be carried out, for example to identify motifs that may directly or indirectly aid UGUA-independent recruitment of CFIm, but too few high-confidence UGUA-independent transcripts were identified here to perform this analysis. Further sequencing repeats, perhaps coupled with the use of a more stable complementation system, would be required to perform such an analysis.

To summarise, this transcriptome-wide analysis of the CFIm25 UGUA-binding mutant (E55A R63A F104A) suggested that the majority of transcripts subject to CFIm-regulated APA are dependent on the UGUA-binding capacity of CFIm25 for this regulation. However, there are clear exceptions to this finding, and it is likely that a more robust approach may be able to define a larger group of 'UGUA-independent' transcripts. Analysis of UGUA distribution and other motifs within such a group could provide further insight into CFIm recruitment and mechanism of action. A particular area of interest would be potential regulation of the UGUA-binding capacity of CFIm25 in different physiological circumstances. For example, a scenario could be envisaged whereby a modification of CFIm25 masks UGUA-binding residues and impairs recognition of UGUA sequences, leading to 3'UTR shortening for the majority of CFIm-regulated transcripts but leaving UGUA-independent transcripts unaffected. This would be analogous, for example, to the selectively maintained translation of stress-response

A



***SEC61A2* pA1**

TCATGACAGTGAGACGGTGAGATGGATTCGTTTTGCACACAACATTCAAAACACTTCATATTGCCCCC
ACTTGTTGAAA**AATAAA****TGTA**GTTCAAATTGC

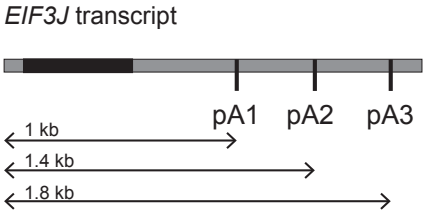
-15

***SEC61A2* pA2**

-100 **-77**

TGTACTCACAGCAGCAACATGAG**TGTA**AACAGTAGACAATAAACTTTTATTTAAGAAAACTGATTCAG
TTGTGTTGGAAAA**AATAAA**GAAATCTGATATT

B



***EIF3J* pA1**

-85

ACAATCCCTTGAACA**TGTA**GCACAACTTCCTTTCCTTTCAGTTCTGCCAAATGCTACAATCAGAAGTG
CAGTATCTTTTGTGCTGGTTATTTAACCCCTT **(No obvious PAS)**

***EIF3J* pA2**

AAAAGCAGTTTGCAGCAAGGGCATGTGGTGTGCACCTAGTATTAAAATTGCTTTGTCTTAAAATTGAA
CATGAGGAT**ATTAAA**AATACATTGTGAAGAAG

***EIF3J* pA3**

CCAGTTAACTACAGTTTGGTAAATTGTTATGTTAACAATTATGACATCTGCAATGTTTTATAAAGCAACT
AATTT**AATAAA**ATCACTGTTGTGAGGACTT

Figure 4.12: Distribution of UGUA sequences upstream of *SEC61A2* and *EIF3J* poly(A) sites. The average cleavage positions for each poly(A) (pA) site in the *SEC61A2* and *EIF3J* transcripts were identified from the sequencing data and cross-referenced with the APASdb online database (http://genome.bucm.edu.cn/utr/). The 100 nucleotides of sequence upstream of each site was selected for analysis. In each case, the PAS is highlighted in blue and any UGUA sequences present (TGTA at the DNA level) are highlighted in red. Numbers refer to distance away from the pA site (sequence end). (A) Illustration of the *SEC61A2* transcript (not to scale) and sequence upstream of its pA sites described (B) As in (A) for the *EIF3J* transcript.

genes in the face of generalised translation inhibition upon eIF2 alpha phosphorylation (Holcik and Sonenberg, 2005). This remains a speculative idea, but is touched on briefly through analysis of potential phosphorylation on T101 and T102, which are in close proximity to the UGUA-binding residues (Section 5.5).

4.6 The CFIm68 RS domain is necessary for normal regulation of APA

Work so far presented in this chapter has confirmed roles for CFIm inter-subunit interactions and UGUA recognition by CFIm25 in normal regulation of APA. To investigate other potential determinants of CFIm activity, attention was turned to the C-terminal RS domain of CFIm68 (Figure 4.13). As discussed in Section 1.4.3, this domain has been implicated in nuclear import and speckle localisation of CFIm68, is able to mediate interactions with SR splicing regulator proteins and may be involved in RNA binding (Cardinale et al., 2007; Dettwiler et al., 2004). However, any potential contribution of this domain to poly(A) site selection remained unexplored.

In order to investigate a potential role for the CFIm68 RS domain in poly(A) site selection, a range of CFIm68 RS domain deletion mutants were constructed (Figure 4.13). The two major mutants, ΔRS1 (lacking residues 481-506) and ΔRS2 (lacking residues 511-551) were designed such that in each case a portion of the domain was removed while leaving intact a putative nuclear localisation signal (pNLS; residues 507-510, RRHK; Figure 4.13). This strategy was used in an attempt to maintain nuclear localisation of CFIm68, in order to potentially highlight specific impairments in poly(A) site selection rather than indirect effects of mis-localisation. To probe whether the pNLS was indeed necessary and sufficient for nuclear import of CFIm68, a separate mutant with deletion of this sequence was generated (ΔpNLS; Figure 4.13). Stable cell lines expressing these CFIm68 mutant isoforms were generated as described in Section 3.9.

Western blotting confirmed that the ΔRS mutant isoforms were expressed at levels comparable to that of the wild type protein, with alterations in apparent molecular weight for ΔRS1 and ΔRS2 as

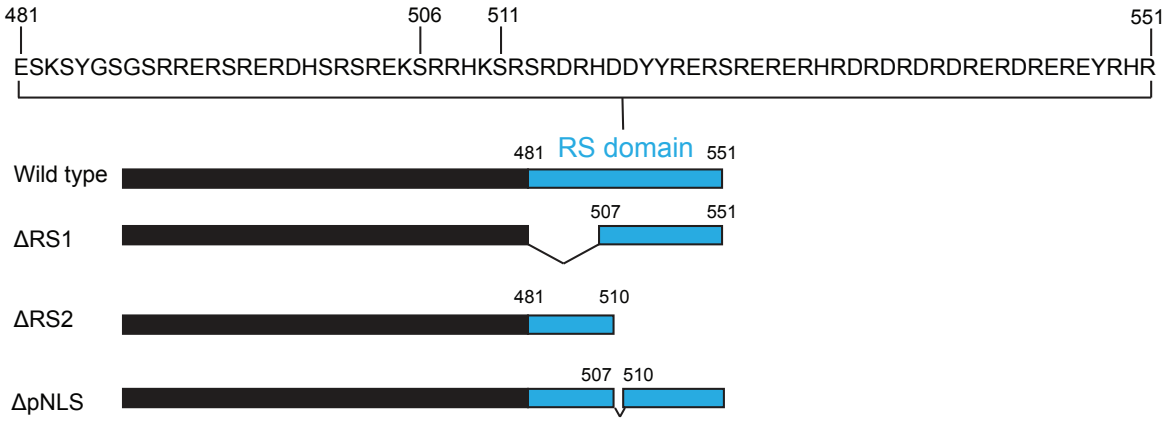


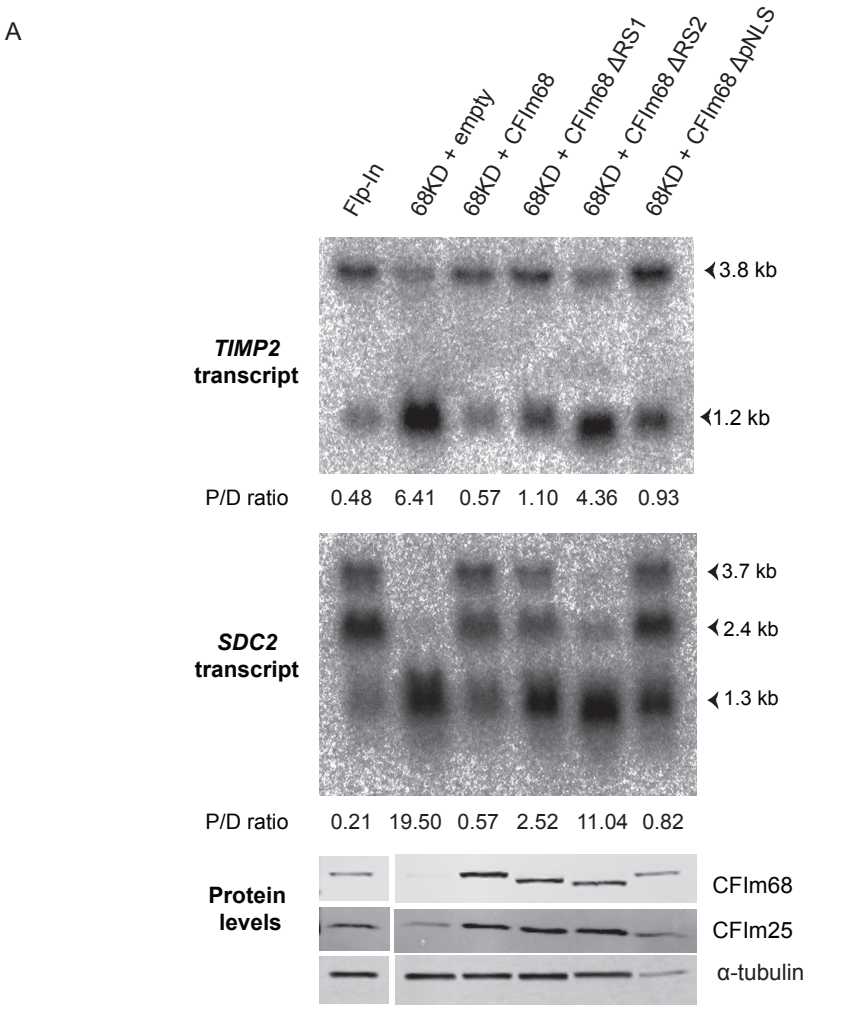
Figure 4.13: The CFIm68 RS domain and ΔRS mutants. An illustration showing the position and sequence composition of the CFIm68 RS domain and the portions missing in the different RS deletion mutants. The full RS domain is highlighted in light blue. Not to scale.

expected (Figure 4.14A). Analysis of RNA from these cell lines in the northern blotting complementation assay showed that both ΔRS1 and ΔRS2 had impaired complementation ability compared to CFIm68 WT for both *TIMP2* and *SDC2* (Figure 4.14A). ΔRS2 showed a more substantial impairment than ΔRS1, with a near-complete loss of complementation for *SDC2*. While the reduced complementation was visually less obvious for ΔRS1, band ratio quantification across three replicates confirmed that both ΔRS1 and ΔRS2 behaved significantly differently to CFIm68 WT (Figure 4.14B). The ΔpNLS mutant did not show any significant impairment in complementation (Figure 4.14A,B).

Overall, these results suggest that the CFIm68 RS domain, and in particular the RS2 region, makes an important contribution to poly(A) site selection. The pNLS, by contrast, does not appear to be necessary for normal APA regulation, suggesting it is unlikely to act as the sole determinant of CFIm68 nuclear localisation. Further investigations were required, however, to explore possible mechanistic explanations for the observed phenotypes.

4.7 CFIm68 partial RS domain deletion mutants are able to localise to the nucleus and interact with CFIm25

One of the simplest potential explanations for the impaired complementation ability of the ΔRS mutants was that loss of the RS domain may disrupt the incorporation of CFIm68 into the CFIm complex, as was observed for the Δ116-122 mutant (Sections 4.3 and 4.4). To examine this possibility, a CFIm25 IP was performed in the WT, ΔRS1 and ΔRS2 cell lines and western blotting was used to quantify co-precipitation of CFIm68 (Figure 4.15). The amount of co-precipitated CFIm68 was similar across all three cell lines, suggesting that the functional impairment of the ΔRS mutants is unlikely to be explained by a loss of interaction with CFIm25. This is in line with the available evidence that the RRM of CFIm68 is sufficient for interaction with CFIm25 (Dettwiler et al., 2004). It does, however, remain a possibility that the RS domain may be involved in mediating



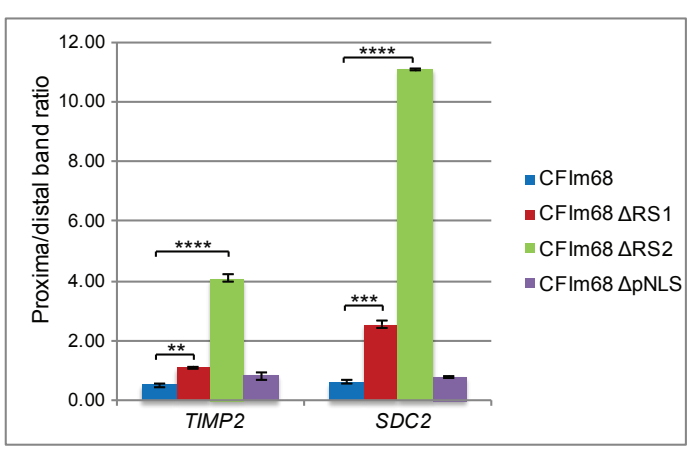


Figure 4.14: CFIm68 ΔRS mutants have an impaired ability to function in APA regulation. Northern blotting and P/D ratio calculations were performed on the given stable cell lines as described in Figure 3.3. The corresponding western blot shows CFIm68 expression in the cell lines, with α-tubulin as a loading control. (B) Mean P/D ratios from biological repeats of this analysis (n=3 for CFIm68, CFIm68 ΔRS1 and CFIm68 ΔRS2; n=2 for ΔpNLS). Error bars depict ±SEM and statistical significance, calculated using a two-tailed Student's t-test, is illustrated with * symbols.

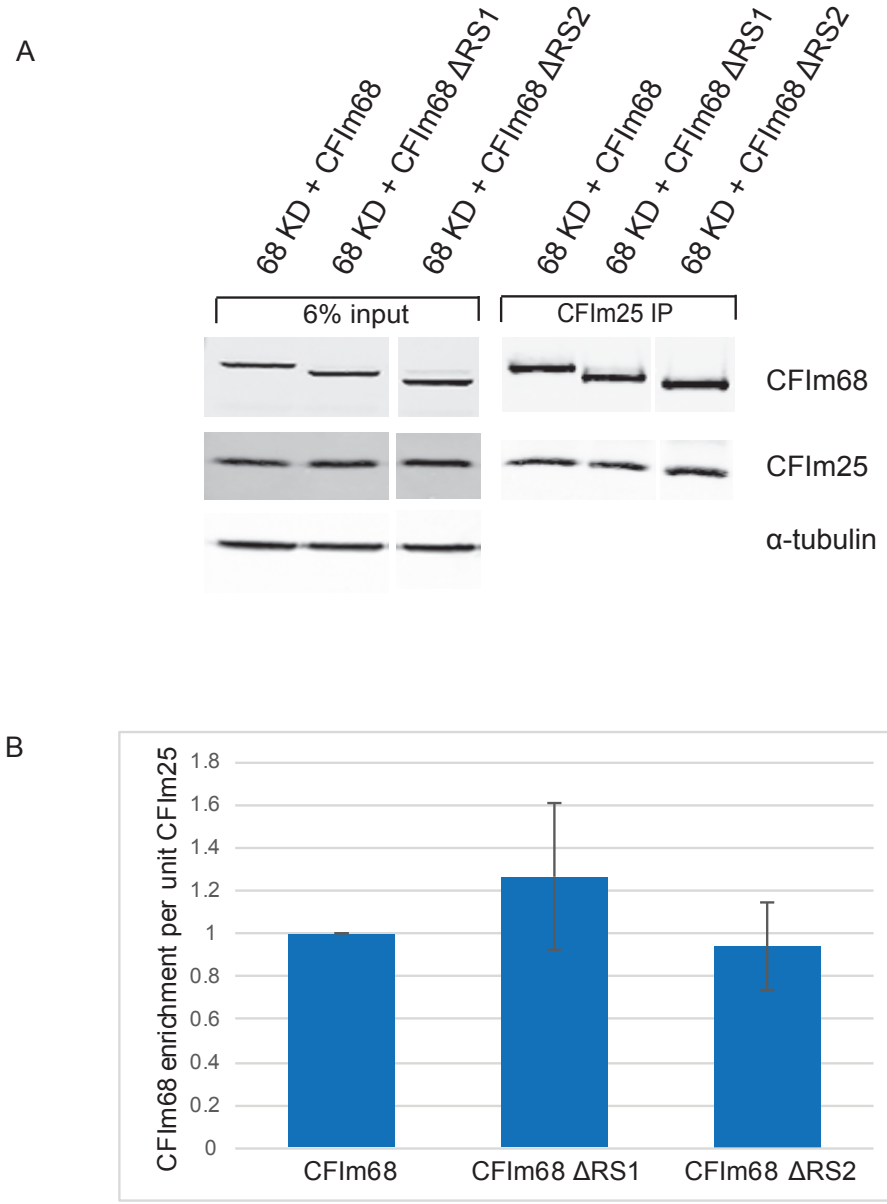


Figure 4.15: The CFIm68 ΔRS mutants retain their ability to interact with CFIm25. (A) CFIm25 was immunoprecipitated from the indicated stable cell lines. Western blotting was performed to investigate the degree to which CFIm68 was co-immunoprecipitated. α-tubulin was used as a loading control for input samples. (B) Quantification of CFIm68 immunoprecipitation by CFIm25. The CFIm68 IP band intensity was divided by the CFIm25 IP band intensity, which was then further divided by the CFIm68 input band intensity (normalised to α-tubulin). This yielded a value for enrichment of CFIm68 per unit of CFIm25 with reference to available input, and values were then standardised to a value of 1 for the 68KD+CFIm68 sample. n=2, error bars depict ±SEM.

interactions between large subunits (CFIm59-CFIm68/CFIm68-CFIm68) within the heterotetramer, as potential interactions of this sort have not been defined.

Another potential explanation for the observed phenotype was that ΔRS1 and ΔRS2 may be functionally inhibited due to reduced nuclear localisation. As discussed in Section 1.4.3, available evidence suggests that the CFIm68 RS domain is sufficient for nuclear localisation of the protein, and this is likely to reflect TNPO3-mediated import (Dettwiler et al., 2004; Maertens et al., 2014). In order to monitor the localisation of ΔRS1 and ΔRS2 by comparison with WT CFIm68, stable cell lines expressing N-terminal GFP-tagged versions of these proteins were generated. Successful expression of the GFP-tagged isoforms was confirmed by western blotting (Figure 4.16).

Before considering localisation, it was important to confirm that the GFP tag itself did not affect the function of CFIm68. This was assessed using the complementation assay, which revealed that GFP-CFIm68 had a complementation ability equivalent to non-tagged CFIm68, and that GFP-CFIm68 ΔRS1 and ΔRS2 also behaved similarly to their non-tagged counterparts (Figure 4.16, compare to Figure 4.14). This suggested that the GFP tag did not affect the function of CFIm68 in poly(A) site selection, and therefore the GFP-tagged proteins could be used as reliable surrogates to assess localisation.

Live cells were imaged using confocal microscopy to visualise the GFP signal, with a Hoechst 33342 counterstain added shortly before imaging to highlight nuclei (Figure 4.17). It was readily apparent that the GFP-tagged WT, ΔRS1, ΔRS2 and ΔpNLS proteins were all predominantly nuclear, showing co-localisation with the Hoechst stain (Figure 4.17A). This suggests that retention of either the RS1 or RS2 region alone is sufficient to promote nuclear localisation, but it is not clear whether this requires the pNLS, which is present in both ΔRS1 and ΔRS2 mutants. The fact that GFP-ΔpNLS retained nuclear localisation suggests that this RRHK sequence is dispensable for nuclear import, though it may still co-operate with the RS domain to increase the efficiency of the process. Consistent with this,

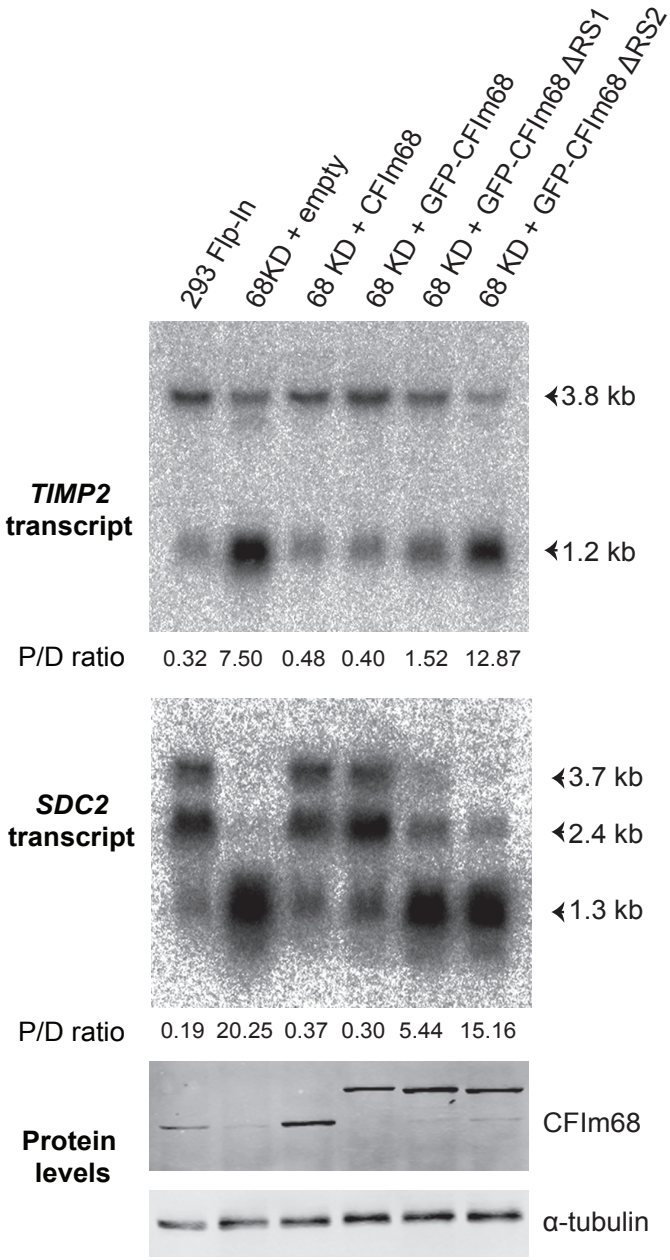


Figure 4.16: GFP-tagged CFIm68 functions similarly to non-tagged CFIm68 in APA regulation. Northern blotting and P/D ratio calculations were performed on the given stable cell lines as described in Figure 3.3. The corresponding western blot shows CFIm68 expression in the cell lines, with α-tubulin as a loading control.

a previous report found that a mutant similar to ΔRS2, but with the additional loss of the pNLS, showed a diffuse cytoplasmic/nuclear distribution (Dettwiler et al., 2004). Despite the use of identical imaging settings, the nuclear GFP signal intensity for ΔRS1, ΔRS2 and ΔpNLS appeared slightly lower than that of the WT (Figure 4.17A). This contrasted with the comparable levels of the GFP-tagged isoforms observed by western blotting (Figure 4.16), suggesting a difference in subcellular distribution rather than expression. While a more quantitative microscopy approach would be needed to confirm and measure this, it suggests that RS1, RS2 and pNLS may all be involved in promoting nuclear import to some extent, perhaps functioning co-operatively in the process.

It could be postulated that the functional impairments observed for the ΔRS mutants in the complementation assay resulted from this observed slight reduction in nuclear localisation. However, this seems unlikely, as there were clear differences in the extent of impaired complementation between ΔRS1, ΔRS2 and ΔpNLS (Figure 4.14), despite the apparent similarity in GFP intensity (Figure 4.17A). Additionally, it is likely that the CFIm68 overexpression achieved in these Flp-In cell lines may result in a quantity of nuclear CFIm68 comparable to that in parental Flp-In cells, even if nuclear localisation is slightly impaired.

In Figure 4.17B, data from a repeat microscopy experiment have been presented slightly differently to highlight further features. Firstly, the larger field of view provides confirmation of the uniformity of nuclear localisation, particularly for GFP-ΔRS2, for which only a few cells were included in Figure 4.17A. In addition, images in Figure 4.17B were adjusted to bring the GFP signal to a similar intensity between WT, ΔRS1 and ΔRS2, so intensities cannot be directly compared as an indicator of protein quantity. This was done to observe and highlight differences in speckle patterns between the variants. For GFP-CFIm68, bright foci, or ‘speckles’, were clearly visible in most cells in combination with the diffuse nuclear staining pattern (Figure 4.17B). This was consistent with previous reports, which found that CFIm68 is targeted to nuclear speckles and paraspeckles (Cardinale et al., 2007). Three

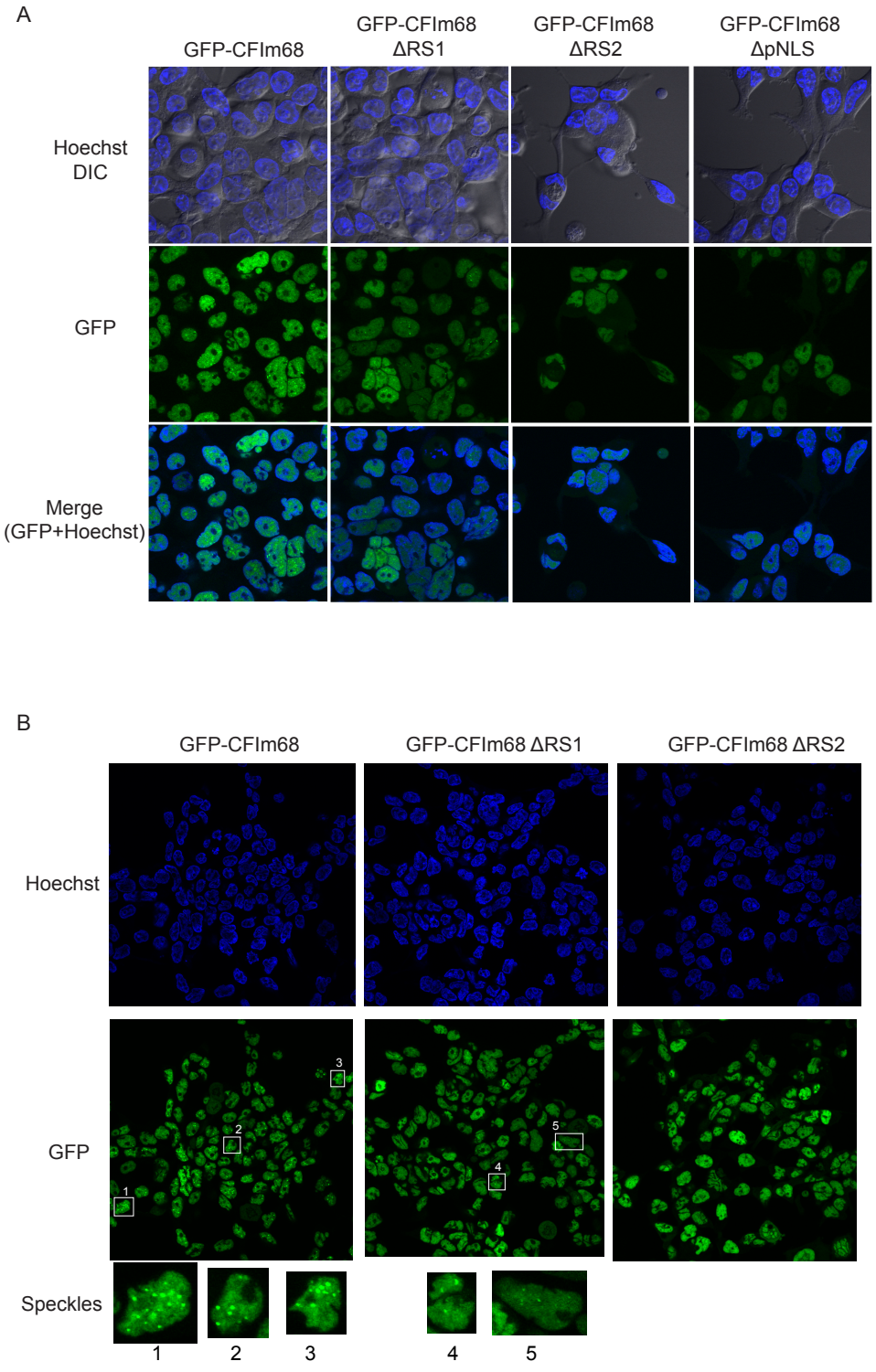


Figure 4.17: The pNLS and the RS1 and RS2 domains are individually dispensable for nuclear localisation of CFIm68. Cell lines expressing the given GFP-tagged CFIm68 variants were observed with live cell confocal microscopy to assess CFIm68 localisation. Hoechst stain was added to the medium shortly before imaging to allow the position of nuclei to be visualised. (A) Separate images were captured and overlaid using Adobe Photoshop. Images were captured at the same intensity and manipulated concordantly to allow direct comparison of signal intensity. (B) Images from a different microscopy session, presented at lower magnification to show a larger number of cells. GFP signals were adjusted individually to achieve comparable intensity and highlight differences in speckling patterns. White numbered boxes and corresponding enlarged views (bottom panel, 'Speckles') highlight examples of cells with nuclear CFIm68 speckles.

examples of GFP-CFIm68 cells with speckles have been highlighted (white boxes), and enlarged for clearer visualisation (bottom panel). For GFP-ΔRS1, the speckled pattern was less pervasive and was observable in only a few cells, two of which have been highlighted and enlarged (bottom panel). For GFP-ΔRS2, no obvious speckling pattern could be observed.

These results suggest that, while the RS1 and RS2 domains may be individually dispensable for nuclear localisation, they appear to be important for subnuclear localisation of CFIm68 within speckles. A more quantitative approach would be needed to confirm these differences in speckle localisation, but this would be an interesting area to follow up. These findings are consistent with the reported contribution of the RS domain to nuclear speckle targeting, both in CFIm68 and other SR proteins, as discussed in Section 1.5.3 (Caceres et al., 1997; Cardinale et al., 2007). The extent of loss of the speckled appearance in ΔRS1 and ΔRS2 was perhaps surprising, given that this pattern is thought to result from a combination of RS-mediated targeting to nuclear speckles and targeting to paraspeckles via the RRM, which remains present in these ΔRS mutants (Cardinale et al., 2007). Localisation of the CFIm68 RRM alone, however, has never been investigated, so it is possible that the RS domain in fact co-operates with the RRM to promote targeting to paraspeckles as well as nuclear speckles.

Overall, these results suggest that the differing complementation phenotypes observed for ΔRS1 and ΔRS2 mutants by comparison with WT CFIm68 cannot be attributed to an inability to interact with CFIm25, nor to failed nuclear import. This suggests that the RS domain of CFIm68 may make a specific contribution to selection of distal poly(A) sites, and further work should focus on elucidating specific mechanistic roles of this domain. The preliminary observations surrounding impaired speckle targeting for ΔRS1 and ΔRS2 highlight the possibility that the contribution of the RS domain to subnuclear localisation plays a key role in regulation of poly(A) site selection.

4.8 Transcriptional consequences of CFIm depletion

Work so far presented here has described the contributions of various domains and functional features of CFIm to poly(A) site selection. However, a full understanding of the mechanism underlying the relative preference of CFIm for distal poly(A) site selection remains elusive. In particular, as discussed in Section 1.4.7, it is unclear whether CFIm acts to stimulate the use of distal poly(A) sites, suppress the use of proximal poly(A) sites, or both, as there is evidence supporting all of these possibilities. To distinguish between these possibilities, two models were proposed, with associated predictions that could be tested using changes in transcriptional profiles as a read-out (Figure 4.18).

In the first model (Figure 4.18A), CFIm25/CFIm68-containing complexes bind at proximal cleavage sites and repress their use. Upon CFIm depletion, it would therefore be expected that these sites are used on a 'first-come, first-served' basis shortly after they are transcribed.

In the second model (Figure 4.18B), CFIm25/CFIm68-containing complexes are considered essential for stimulating the use of distal poly(A) sites. Thus, upon CFIm25/CFIm68 knockdown or inhibition, these distal sites cannot be used, and eventually, CFIm-independent cleavage occurs at the often suboptimal proximal poly(A) site. A prediction of this model would be that in the absence of CFIm, failed cleavage at the distal poly(A) site would lead to delayed transcription termination before eventual recovery of cleavage at the proximal site, resulting in read-through transcription. It was therefore possible to test this model by measuring the presence of transcribing pol II downstream of CFIm-regulated genes in the 25KD and 68KD cell lines.

To this end, a chromatin immunoprecipitation (ChIP) experiment was carried out with a polyclonal antibody against pol II in 25KD and 68KD cells and their parental counterparts, alongside a parallel normal IgG negative control IP. DNA from the ChIP experiment was first analysed by quantitative PCR (qPCR). Oligonucleotides were designed to amplify a region around 1.2 kb downstream of pA3

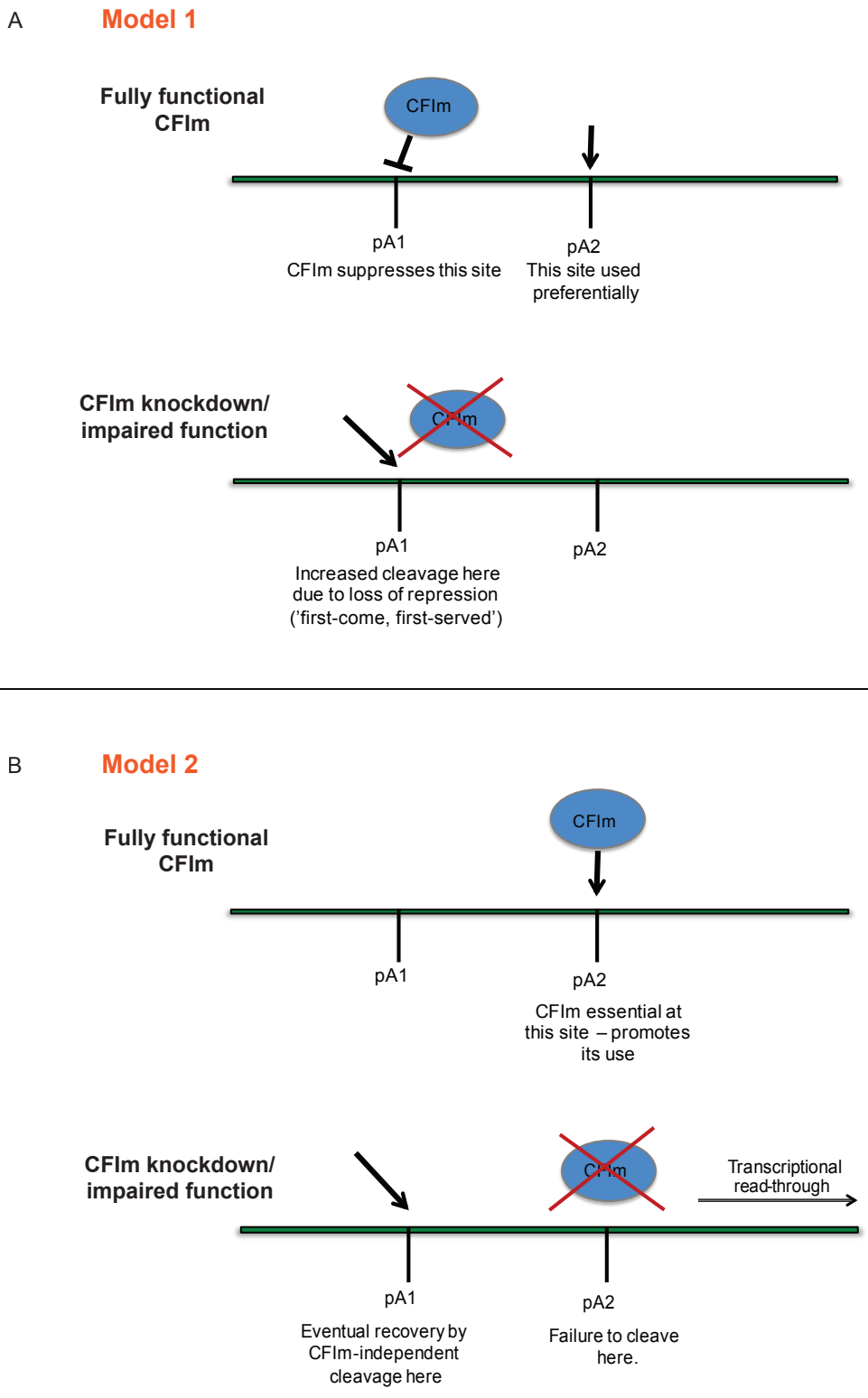


Figure 4.18: Two alternative models for the contribution of CFIm to poly(A) site selection. (A) A model suggesting a suppressive role for CFIm at proximal poly(A) sites. (B) A model suggesting a stimulatory role for CFIm at distal poly(A) sites.

in the *SDC2* gene, one of the model transcripts shown to have a clear APA response upon CFIm25/CFIm68 knockdown (Figure 3.3). Another pair of oligonucleotides around the TSS was tested as a positive control.

Results across two qPCR repeats on the same ChIP samples suggested that there was no detectable pol II binding in the region 1.2 kb downstream of the *SDC2* gene tested in either the parental, 25KD or 68KD cells (Figure 4.19). Values obtained as a % of input were very close to zero, even dipping below zero for 25KD and 68KD, suggesting that this represents background signal. At the *SDC2* TSS, there was detectable pol II binding across all samples, as would be expected due to pol II pausing (Adelman and Lis, 2012). Results suggested that there was a loss of around 75% of the pol II signal at this site in 25KD compared to 293T, but that by contrast there was a small increase in pol II signal here in 68KD compared to Flp-In cells (Figure 4.19). One experiment was performed with a pair of oligonucleotides slightly less downstream (300 bp) of the *SDC2* pA3 site, in order to probe for potential 'short-lived' transcriptional read-through, but again the signal was not above background in any cell line (Figure 4.19). Overall, these results suggested that pol II occupancy is not increased downstream of the *SDC2* gene in response to CFIm25 or CFIm68 depletion, and therefore provide no support for the model of CFIm function presented in Figure 4.18B. However, they revealed potentially interesting impacts of CFIm depletion on transcription initiation/elongation.

The ChIP-qPCR analysis only considered one gene, and while *SDC2* provides a good example of a CFIm-regulated APA event, it may not be representative of the general mechanism of CFIm action. To investigate the global transcriptional consequences of CFIm depletion, ChIP samples from a repeat experiment were submitted for whole-genome deep sequencing,

To look for potential CFIm-associated global changes in transcription termination, metaprofiles of pol II distribution across non-overlapping protein-coding genes were constructed, centred on the poly(A) site (Figure 4.20A,B). The poly(A) sites were as defined by GENCODE, which annotates the most

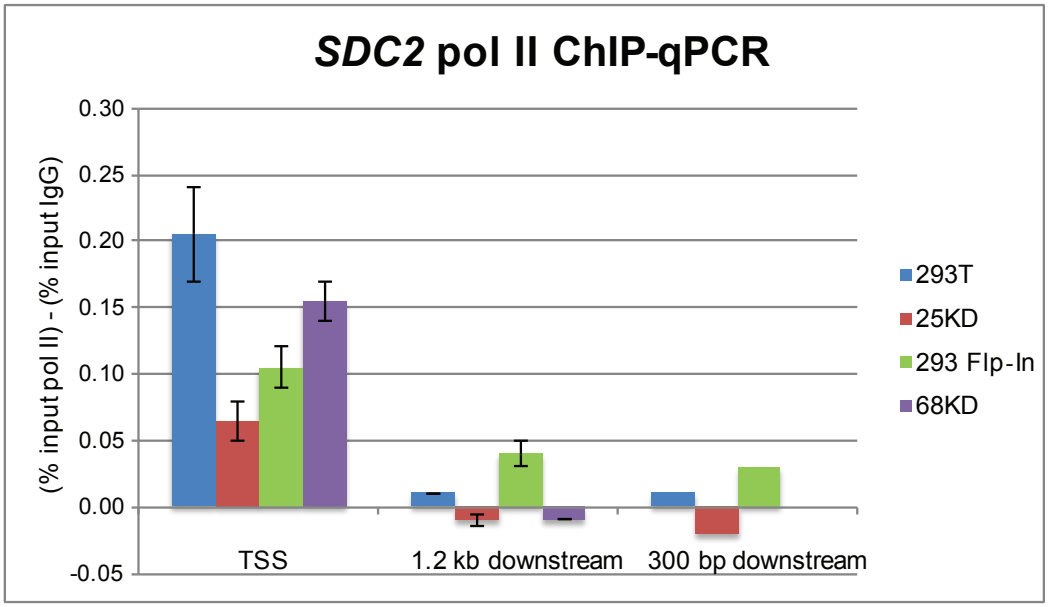


Figure 4.19: ChIP-qPCR analysis of RNA polymerase II occupancy at the TSS and 3'end of the *SDC2* gene. A ChIP experiment was performed on the given cell lines with pol II and normal IgG control antibodies. The immunoprecipitated DNA was analysed in a qPCR experiment with oligonucleotides spanning the TSS or given regions downstream of the third poly(A) site (pA3, see Figure 3.3) of *SDC2.* A five-fold dilution series of input samples for each cell line was analysed in parallel and used to generate a standard curve, which was then used to quantify the IP signal as a percentage of input (% input). The % input values for control IgG were subtracted from % input values for pol II to remove background. TSS and 1.2 kb downstream were analysed across two experiments, and 300 bp downstream only in one. Error bars represent ±SEM.

distal site. The metaprofiles revealed a very small decrease in pol II occupancy around these sites in the 25KD line compared to 293T, but this was present in both the gene body and after the poly(A) site, suggesting no poly(A) site-associated effect (Figure 4.20A). There was a much more noticeable loss of pol II signal in this region in 68KD compared to 293 Flp-In cells, but again the signal loss was observable to a similar extent both before and after the poly(A) site (Figure 4.20B). Overall, these results suggested that depletion of CFIm25 or CFIm68 does not cause delayed termination after the most distal poly(A) site, and in fact leads to decreased pol II occupancy in this region. The observation of similar changes in pol II occupancy both preceding and following the poly(A) site suggests a generalised decrease in overall transcription in 25KD and 68KD cells, which will be discussed further below.

Pairwise comparison of pol II occupancy in individual genes (293T vs. 25KD or 293 Flp-In vs. 68KD) somewhat supported these conclusions (Figure 4.20C). For 68KD, there was a significant decrease in pol II occupancy in both the gene body and downstream of the poly(A) site compared to the parental line, consistent with the appearance of the metaprofile. For 25KD, however, there was a significant decrease in the gene body, but no significant change after the poly(A) site. This discrepancy is likely to be a result of different gene selection criteria in these two analyses, but nonetheless supports the general finding that depletion of CFIm25 or CFIm68 does not lead to any obvious transcriptional read-through downstream of distal poly(A) sites. The genome browser screenshot for the 3' end of *SDC2* (Figure 4.20D), the gene analysed by qPCR (Figure 4.19), confirmed the general lack of transcriptional read-through following the poly(A) sites and the similarity in pol II occupancy in the different cell lines (for reference, compare to TSS read values in Figure 4.21D).

With the absence of transcriptional read-through resulting in a lack of support for model 2 (Figure 4.18B), further data analysis was performed to test the assumptions of model 1. If loss of CFIm-mediated repression at proximal poly(A) sites leads to increased use of these sites upon their

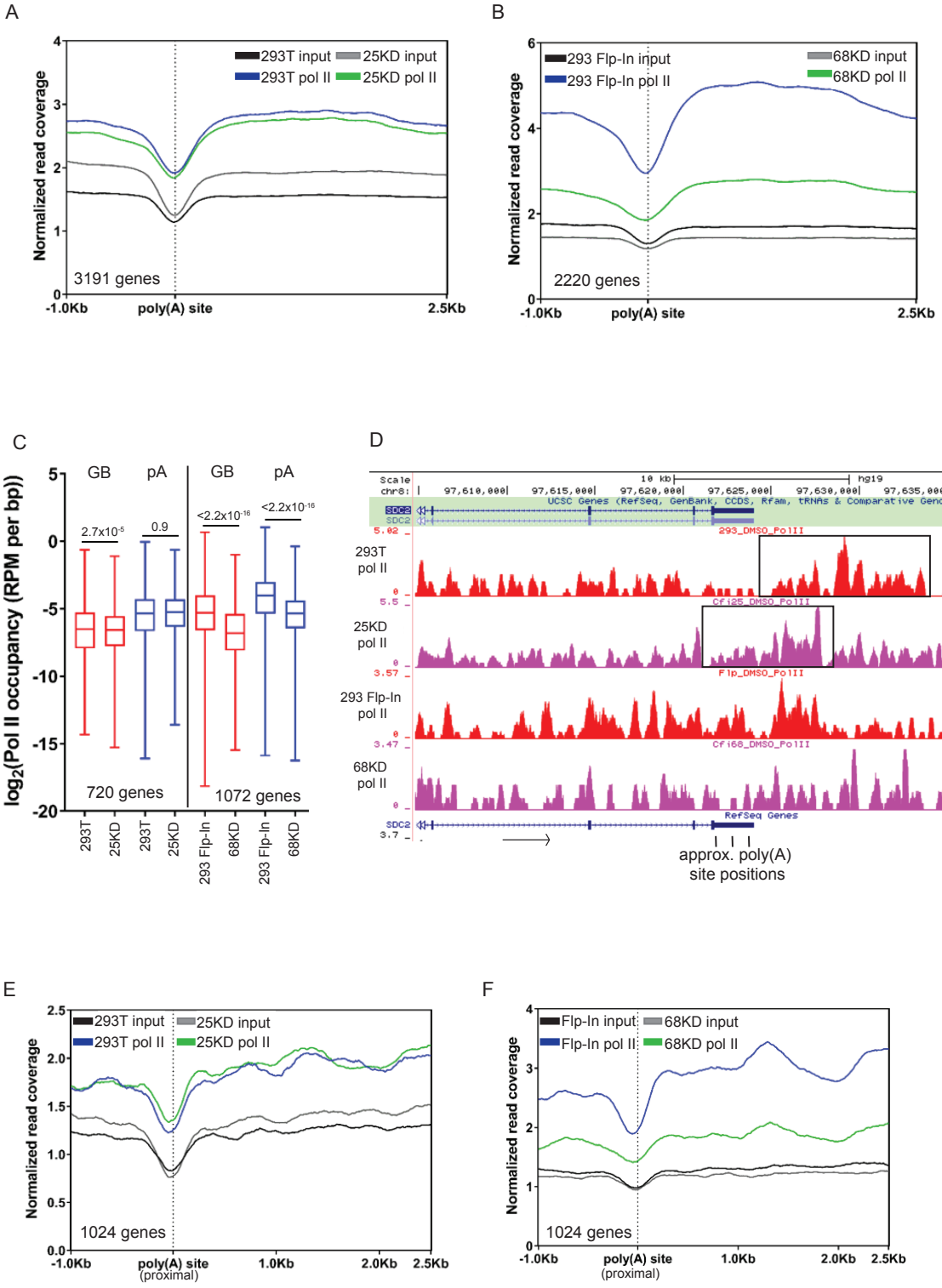


Figure 4.20: ChIP-sequencing analysis of genome-wide RNA polymerase II occupancy at 3' ends of genes in CFIm-depleted cells. A ChIP experiment was performed on the given cell lines with pol II and normal IgG control antibodies. The immunoprecipitated DNA was analysed by deep sequencing, alongside input samples for reference. Figures are all adapted from originals by Dr Michael Tellier. (A-B) Metaprofiles of non-overlapping protein-coding genes centred on the GENCODE-defined poly(A) site for 293T vs 25KD (A) and 293 Flp-In vs 68KD (B). (C) Box-and-whisker plots of pol II occupancy in the gene body (GB) and poly(A) site (pA). RPM = reads per million. Significance of differences between sample pairs was calculated using the Wilcoxon matched-paired signed rank test. (D) Genome browser screenshot of pol II occupancy at the 3' end of the *SDC2* gene. Boxes highlight possible 'termination peaks' as discussed in the text. (E-F) Metaprofiles as in (A/B), but for genes with >20% increased use of proximal pA sites in 25KD condition based on 3'seq data, and centred on proximal poly(A) sites. For 293T vs 25KD (E) or 293 Flp-In vs 68KD (F).

transcription, it may be expected that there would be an increased pol II pausing just downstream of these sites ('termination peak'). To search for this, metaprofiles were constructed of genes that showed >20% increased use of the proximal poly(A) site in 25KD cells based on the 3'seq data (see Section 3.5), and were centred on the proximal poly(A) sites (Figure 4.20E,F). These profiles revealed no obvious change in pol II occupancy around the proximal poly(A) site between 293T and 25KD (Figure 4.20E). For 68KD, there was again a general drop in overall amount of pol II around these sites as observed at the distal poly(A) sites, but the shape of the profiles was similar between 293 Flp-In and 68KD (Figure 4.20F). These results suggest that CFIm depletion does not lead to a global upstream shift in transcription termination to the proximal poly(A) site, and therefore do not provide any support for the 'first-come, first-served' model (Figure 4.18A). Considering *SDC2* again (Figure 4.20D), however, a possible upstream shift in the 'termination peak' was visible in the 25KD cell line compared to 293T, as highlighted by boxes. This may represent a gene-specific example of this type of mechanism, but would require confirmation by qPCR and repeat analysis. In addition, the same pattern was not visible for the 293 Flp-In/68KD comparison, with the peak already appearing at this upstream position in the Flp-In cells (Figure 4.20D), making interpretation more complex.

While there was no clear effect of CFIm depletion on transcription termination, the ChIP-seq data did highlight a striking global effect at TSSs (Figure 4.21). For the 68KD line in particular, by comparison with the parental line, metagene analysis and pairwise gene comparisons highlighted a substantial drop in pol II occupancy at TSSs genome-wide (Figure 4.21B,C). This drop was maintained into the gene body, as discussed above, leading to a general decrease in pol II occupancy throughout genes. A much smaller loss of TSS occupancy was observed for 25KD compared to 293T, which only became visible when normalising to input values (Figure 4.21A,C). This may reflect the lesser degree of CFIm depletion in 25KD compared to 68KD, or possibly a mechanistic difference. The pol II profile in *SDC2* was consistent with the general findings, showing a decreased TSS peak in 25KD and 68KD cells compared to parental counterparts (Figure 4.21D). Interestingly, when normalised to inputs, there was actually a bigger pol II signal loss in 25KD than in 68KD for this gene – this is somewhat

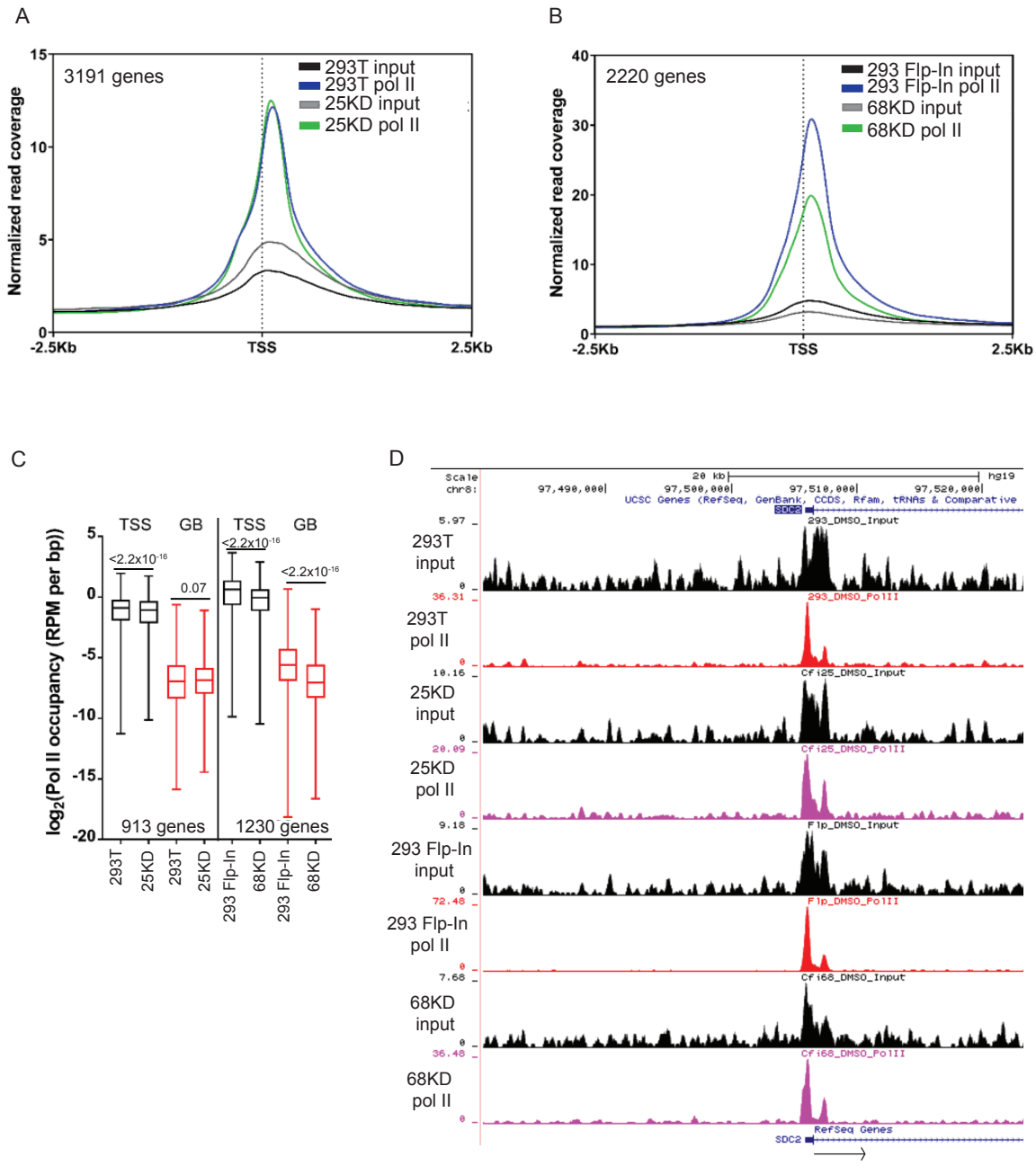


Figure 4.21: ChIP-sequencing analysis of genome-wide RNA polymerase II occupancy at transcription start sites of genes in CFIm-depleted cells. A ChIP experiment was performed on the given cell lines with pol II and normal IgG control antibodies. The immunoprecipitated DNA was analysed by deep sequencing, alongside input samples for reference. Figures are all adapted from originals by Dr Michael Tellier. (A-B) Metaprofiles of non-overlapping protein-coding genes centred on the TSS as defined by GENCODE for 293T vs 25KD (A) and 293 Flp-In vs 68KD (B). (C) Box-and-whisker plots of pol II occupancy at the TSS and in the gene body (GB). RPM = reads per million. Significance of differences between sample pairs was calculated using the Wilcoxon matched-paired signed rank test. (D) Genome browser screenshot of pol II occupancy at the TSS of the *SDC2* gene.

consistent with the qPCR results, from a different ChIP experiment, which actually suggested an increase in pol II signal in the 68KD line and a large decrease in the 25KD line (Figure 4.19).

These intriguing observations of CFIm-related changes in pol II occupancy at the TSS and in the gene body suggest that CFIm68, and to a lesser extent CFIm25, plays an important role in productive transcription initiation and elongation. Whether this decreased pol II occupancy throughout genes plays a causal role in the observed changes in APA upon CFIm depletion is unclear, and would require further investigation – ideally through phenocopying the initiation defect and analysing the APA phenotype.

Overall, these results did not provide conclusive support for either of the CFIm mechanistic models presented in Figure 4.18, and it remains unclear how CFIm depletion brings about a directional shift in poly(A) site choice. The lack of any clear change in pol II occupancy at the 3' end was surprising, given the known APA changes in these CFIm-depleted cell lines and the intimate links between CPA and transcription termination. It may be that the technique used here was not sensitive enough to detect any transcriptional read-through or subtle changes in pol II pausing after proximal poly(A) sites, and that any small effects were potentially masked by the dramatic changes at the TSS and in gene bodies. Further analyses may need to return to the individual gene level, with perhaps more detailed dissection of pol II and CFIm occupancy across whole genes and artificial tethering of CFIm to different sites, to gain a clearer mechanistic picture.

Chapter 5
Post-translational modification of CFIm

5.1 Introduction

Before commencing this project, a number of post-translational modification (PTM) sites had already been identified on all three subunits of CFIm. Identified CFIm PTMs include numerous sites of phosphorylation, acetylation and methylation, which are summarised in the online PhosphoSite Plus (PSP) database (Hornbeck et al., 2015). The majority of these PTMs were identified in proteome-wide mass spectrometry (MS) studies rather than CFIm-focused studies, and the supporting evidence for many of them is lacking or low-quality. With two exceptions, namely acetylation of K23 in CFIm25 (Shimazu et al., 2007) and methylation of the GAR motif in CFIm68 (Martin et al., 2010), CFIm PTMs have not been the subject of specific investigation. As a result, the extent to which such PTMs may regulate CFIm activity remains unclear, as do the identities of enzymes and pathways involved. Furthermore, it is likely that other PTMs on the complex may have escaped detection due to the lack of focused investigation.

To expand current knowledge of CFIm PTMs and their significance, experiments were performed to identify new PTM sites, and to screen these and known sites for any effects on CFIm-regulated APA. There was particular interest in the hypothesis that changes in the CFIm PTM profile during the transition from quiescence into proliferation may drive changes in APA, in the absence of changes in CFIm expression level (Appendix 1).

MS and western blot-based analysis of immunoprecipitated CFIm allowed identification of several PTM sites, some of which were novel. In particular, extensive serine phosphorylation was detected in the CFIm68 RS domain, and site-directed mutagenesis allowed identification of some of the phosphorylated residues.

Using site-directed mutagenesis coupled with the APA complementation assay (Chapter 3), newly identified and known PTM sites on CFIm25 and CFIm68 were screened for effects on APA. The majority of PTM site mutants studied showed no major difference in poly(A) site selection to WT CFIm25/CFIm68. However, a mutant with a number of phosphomimetic serine-to-aspartate mutations in the CFIm68 RS1 domain showed a complementation defect, raising the possibility that phosphorylation within this domain may contribute to APA regulation.

5.2 Selection of known PTM sites for analysis

Alongside attempts to identify CFIm PTM sites de novo, sites of interest were selected for analysis from publicly available databases and literature. The PSP database (Hornbeck et al., 2015) provides a continuously curated resource that collates low- and high-throughput data from PTM studies. The online platform, at phosphosite.org, allows searches to be carried out by protein name, yielding a list of curated PTMs (acetylation, phosphorylation, methylation, ubiquitination) and the studies in which they were identified.

The PSP database was used to identify PTMs of interest on CFIm25 and CFIm68 for screening in the APA complementation assay. CFIm25 and CFIm68 had 18 and 42 PTMs listed respectively, with varying sources of supporting evidence. Due to the labour-intensive

nature of screening and the likelihood of false positives in these data, several criteria were used to prioritise PTMs for screening (Table 5.1).

Firstly, sites that had been extensively validated in a publication were selected based on this alone. Where such direct evidence was not available, repeated identification in diverse studies was considered a good indicator of confidence in the assignment. Therefore, sites found in at least five studies, with at least one that could be manually validated, were selected. Sites which were found in several studies all of the same type, for example repeated data sets derived from Jurkat cells treated with pervanadate, were not selected based on this criterion alone.

Some sites with weak or missing evidence were still included on the basis of potential biological interest. For example, S165 in CFIm68 was found phosphorylated in only one study, but an in vitro demonstration that phosphomimetic mutation of S165 may increase RNA binding affinity (Yang et al., 2011b) prompted its inclusion.

Based on these criteria, six CFIm25 modification sites and eight CFIm68 modification sites were selected for screening (Table 5.1).

Another factor considered was whether these sites were conserved across diverse vertebrates, including mammals, birds and fish, as an indicator of potential functional importance (Appendix 3; Table 5.1). This was of limited utility for CFIm25, which has a very highly conserved sequence overall, showing 100% conservation with *Bos taurus* (cow), only one amino acid difference with *Mus musculus* (mouse), and even 98% sequence conservation with *Danio rerio* (zebrafish). Thus, while all CFIm25 PTM sites were fully conserved across the

vertebrates studied, this was unsurprising and did not necessarily increase confidence in their functional importance. The analysis was slightly more useful for CFIm68, which shows less overall sequence conservation (e.g. 84% with *Danio rerio*). This revealed that all PTMs selected were conserved across mammals, but the fact that T407 was also conserved in birds and fish while T404 was mammal-specific suggested that it may be more likely to represent an important phosphorylation site, for example (Table 5.1).

Table 5.1: Selection of known CFIm25 and CFIm68 PTMs for investigation. The PSP database (Hornbeck et al., 2015; phosphosite.org) was used to search for PTMs on CFIm25 and CFIm68. PTMs of interest were selected based on criteria listed here and as discussed in the text. Protein sequence conservation was measured using the T-Coffee alignment programme (EMBL-EBI) across the following vertebrates: *Homo sapiens, Pan troglodytes, Bos taurus, Mus musculus, Gallus gallus, Python bivittatus, Xenopus tropicalis, Latimeria chalumnae, Danio rerio, Callorhinchus milii.*

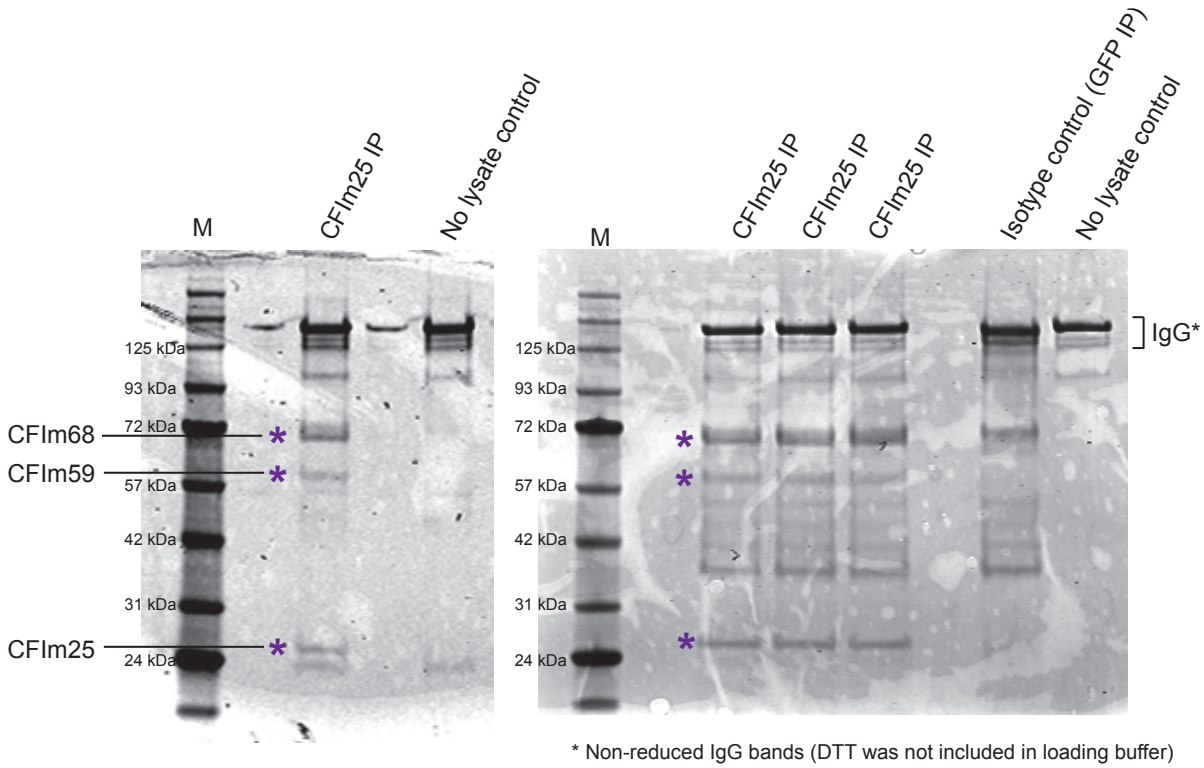
Subunit	Modification	Selection criteria	Residue conserved?
CFIm25	R15 methylation	5 or more identifications; at least 1 manually validated.	Yes
CFIm25	K23 acetylation	Extensively validated in a publication (Shimazu et al., 2007)	Yes
CFIm25	Y24 phosphorylation	5 or more identifications; biological interest – neighbours acetylated K23.	Yes
CFIm25	Y40 phosphorylation	5 or more identifications; at least 1 manually validated.	Yes
CFIm25	T101 phosphorylation	Biological interest – close to UGUA binding residues.	Yes
CFIm25	T102 phosphorylation	Biological interest – close to UGUA binding residues.	Yes
CFIm68	Y74 phosphorylation	5 or more identifications; at least 1 manually validated.	Yes
CFIm68	Y76 phosphorylation	5 or more identifications; potential ambiguity with Y74.	Yes
CFIm68	S165 phosphorylation	Biological interest – affects RNA binding in vitro.	Yes
CFIm68	R202 methylation	Extensively validated in a publication (Martin et al., 2010)	Yes, except for *Danio Rerio*
CFIm68	R204 methylation	Extensively validated in a publication (Martin et al., 2010)	Yes, except for *Danio Rerio*
CFIm68	R206 methylation	Extensively validated in a publication (Martin et al., 2010)	Yes
CFIm68	T404 phosphorylation	5 or more identifications; at least 1 manually validated.	In mammals only
CFIm68	T407 phosphorylation	5 or more identifications; at least 1 manually validated.	Yes, except for *Xenopus tropicalis*

5.3 Identification of new CFIm PTM sites by mass spectrometry

In an attempt to validate previously described CFIm PTM sites and identify new ones, a directed MS analysis was carried out on immunopurified CFIm. The most efficient IP of the CFIm complex was achieved using a monoclonal antibody against CFIm25, which interacts with both larger subunits. This IP approach was applied to whole cell lysates from asynchronously growing BJ-5ta fibroblasts, with a view to later use of this method to track changing PTM profiles during synchronised cell cycle progression in this cell type (see Appendix 1).

Following IP, LDS-eluted samples were run on a polyacrylamide gel, and the gel was stained with a Coomassie-based reagent to visualise proteins (Figure 5.1). Bands corresponding to the expected sizes of the three CFIm subunits were identifiable in the IP and absent from the 'no lysate' control, confirming that they were not antibody-derived. A repeat of this procedure (Figure 5.1, right-hand side) using a new batch of Protein G/sepharose beads and antibody gave rise to more bands, but inclusion of an isotype control IP suggested most additional bands were non-specific (e.g. from bead binding) rather than CFIm interaction partners. In summary, the successful identification of CFIm subunit bands following IP suggested that there was enough material obtained from this approach to allow analysis by MS.

For MS analysis, the three CFIm bands (Figure 5.1) were excised and subjected to trypsin digestion to generate peptides, and samples were then analysed by LC-MS/MS. Data analysis of three repeat experiments using the Mascot server showed that CFIm25, CFIm59 and CFIm68 were consistently detected in the top five protein hits for their respective bands, coming only occasionally below keratin and heat shock proteins (common contaminants in



* Non-reduced IgG bands (DTT was not included in loading buffer)

Figure 5.1: Immunoprecipitation of CFIm for mass spectrometry analysis. Anti-CFIm25 immunoprecipitation from BJ-5ta fibroblast whole-cell lysate was carried out and the resulting sample eluted from beads and resolved by SDS-PAGE. This image shows two repeats - that on the right was performed with new beads and antibody, which may explain the higher number of bands. The no lysate control (CFIm25 antibody and lysis buffer only) allowed antibody-derived bands to be identified, while the isotype control (right-hand side only) allowed identification of non-specific binders in the lysate to beads or common antibody regions. Protein was visualised using a Coomassie-based stain, and the stained gel was imaged using the LI-COR Odyssey scanner in the 700 nm channel. M = marker lane, purple asterisks indicate the CFIm subunit bands as labelled (with the same subunits in the same order on the right-hand side).

MS experiments). The high Mascot scores (Table 5.2) assigned to the subunits reflect a large number of high-confidence peptides identified, and confirmed that the bands seen on gel staining corresponded to CFIm subunits.

The main interest in this experiment was whether any PTMs could be identified on CFIm, and this was investigated by applying a Mascot variable modification search to highlight peptide masses indicating potential phosphorylation (serine, threonine, tyrosine), acetylation (lysine) or methylation (arginine, lysine). Mascot peptide scores and inspection of spectra were used to determine which PTMs could be considered high-confidence, and the results from this analysis across all experiments are summarised in Table 5.2.

Table 5.2: PTMs identified on CFIm subunits by MS. Data from three repeat MS experiments following trypsin digestion of CFIm gel bands was analysed using Mascot software (Matrix Science). Modifications listed were deemed to be likely true positives based on identification in at least two experiments, E values of <0.05 and manual validation of the individual peptide spectra. The best Mascot peptide score and associated E value is reported for each PTM site. The average Mascot protein score reported for the subunits represents the average of the sum of peptide scores from that subunit. Novel PTMs (not reported on PSP) are highlighted in green. Residue conservation is across selected vertebrates as discussed in the text.

Subunit and modification	Best Mascot peptide score	Residue conserved?
CFIm25 (average Mascot protein score - 7383)		
R15 monomethylation	45, E = 6.0×10^{-4}	Yes
R15 dimethylation	63, E = 2.9×10^{-5}	Yes
K23 acetylation	55, E = 1.2×10^{-5}	Yes
K29 acetylation	34, E = 7.8×10^{-4}	Yes
K173 acetylation	67, E = 2.8×10^{-6}	Yes
CFIm59 (average Mascot protein score - 5109)		
T203 phosphorylation	62, E = 2.1×10^{-6}	
S46/S47/S48 phosphorylation (site unclear)	49, E = 1.6×10^{-3}	
CFIm68 (average Mascot protein score - 4979)		
K139 acetylation	24, E = 0.025	Yes
K161 acetylation	16, E = 0.046	Yes
T404/T407 phosphorylation (site unclear)	81, E = 3.4×10^{-8}	See Table 5.1

Most of the identified PTMs on CFIm25 were lysine acetylation sites. The previously characterised K23 acetylation site (Shimazu et al., 2007) was identified, and K29 acetylation, previously identified on PSP, was also detected. CFIm25 K173 acetylation was identified as a novel PTM. Inspection of a CFIm25-CFIm68 tetrameric crystal structure, with the PDB ID of 3Q2T (Yang et al., 2011a), revealed that K173 sits at the CFIm25 dimer interface, suggesting the potential for acetylation to have an interesting regulatory role. As well as acetylation, methylation events were identified on R15, which was subject to both mono- and di-methylation. This had already been reported in several studies on PSP.

On CFIm68, two acetylation sites were identified, although with slightly lower confidence than those in CFIm25. One of these sites, K139, was a novel identification, while K161 is listed on PSP. Both of these residues were fully conserved across all vertebrates studied. As well as these acetylation events, one phosphorylation site was identified, but it was not possible to distinguish between T404 and T407 as the site responsible – both of these have been previously reported on PSP. CFIm59 was also studied here, despite the lack of a functional complementation assay. For this subunit, two phosphorylation sites were detected: T203, and S46, S47 or S48.

Several alternative MS sample preparation methods were tried, including two solution-based digest protocols in which gel electrophoresis was omitted to reduce sample manipulation steps and potentially better preserve modifications. These involved acidic glycine elution of protein from the beads, followed by either direct trypsin digestion or trypsin digestion following filter-aided sample preparation - FASP (Wisniewski et al., 2009). While these approaches did identify CFIm subunits, identification of PTMs was less effective than for the in-gel digestion approach, and data from these experiments were not used.

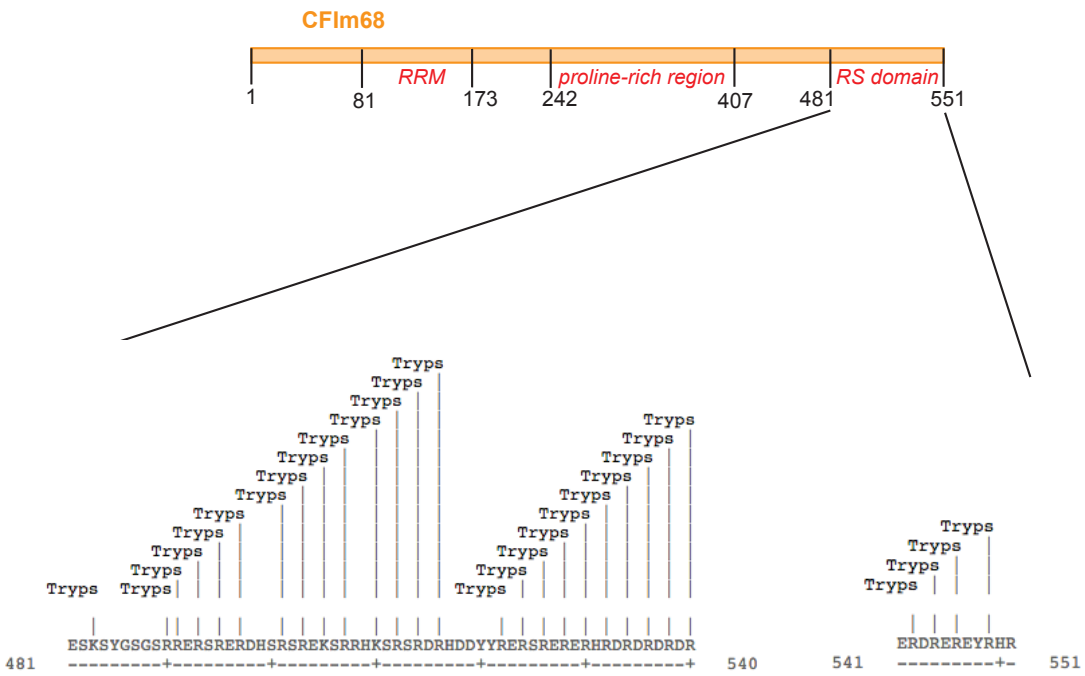
In summary, CFIm-directed MS analysis provided supporting evidence for several previously identified PTM sites and identified two novel acetylation sites (Table 5.2), which were added to the list of PTMs for functional screening.

5.4 Identification of serine phosphorylation sites in the CFIm68 RS domain

While MS can be a powerful method for detection of new PTMs on proteins of interest, it also presents certain disadvantages. The 'gold standard' method of trypsin digestion to generate peptides gives good coverage of proteins where arginine and lysine residues are relatively regularly spaced and infrequent, as is the case in CFIm25. For CFIm59 and CFIm68, however, the C-terminal RS domain is heavily digested by trypsin due to the high arginine density, producing very short peptides, which do not have sufficiently distinctive masses to be identified unambiguously (Figure 5.2A). There was therefore no coverage of the CFIm68 RS domain in the MS data, meaning that potential phosphorylation events involving the multiple serines and tyrosines present in this region (Figure 5.2B) may have been missed.

In an attempt to overcome this pitfall, an MS experiment with the alternative protease AspN, which was predicted to yield peptides of 15-20 amino acids in length, was attempted. However, for unknown reasons, there was still no coverage of the RS domain in the results (not shown). Ultimately, a non-MS-based method for phosphorylation detection was therefore employed to analyse the RS domain. CFIm was immunoprecipitated and subjected to gel electrophoresis in the same way as for MS, but subsequently analysed using a western

A



B

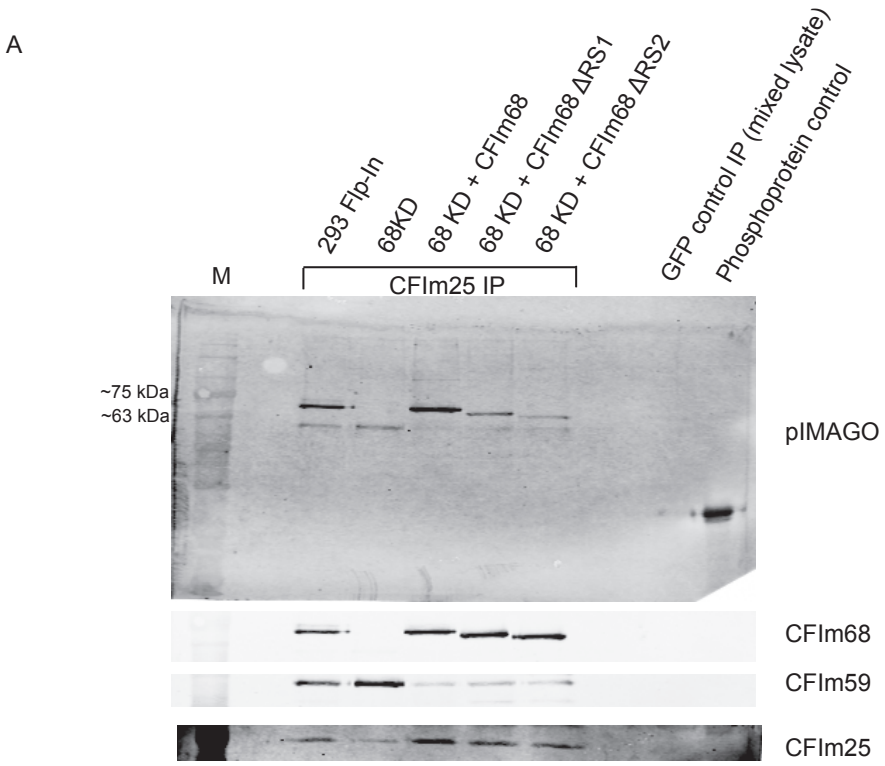
481 ESKSYGSGSRRERSRERDHSRSREKSRRHK 510
511 SRSRDRHDDYYRERSRERERHRDRDRDRD 539
540 RERDRERERYRHR 551

Figure 5.2: Trypsin digestion pattern and potential phosphoacceptor sites in the CFIm68 RS domain. (A) The ExPASy PeptideCutter tool (http://web.expasy.org/peptide_cutter/) was used to visualise the sites of trypsin digest within the RS domain (amino acids 481-551) of CFIm68. The domain organisation is depicted at the top and sites of trypsin digestion in the RS domain (C-terminal side of arginines and lysines) are highlighted (Tryps). (B) Amino acid sequence of the RS domain (residues 481-551) with all potential phosphoacceptor sites (serines and tyrosines) highlighted in red.

blotting-based approach. The pIMAGO kit, which utilises a titanium-based nanopolymer able to detect phosphorylation in a context-independent manner (Iliuk et al., 2012), was used to identify phosphorylated proteins on the blot. Initial testing of the approach in 293T cells revealed bands suspected to correspond to CFIm68 and CFIm59 (not shown). Following this, a further IP experiment was performed to compare 293 Flp-In, 68KD and 68KD+CFIm68 (WT/mutant) cell lines (Figure 5.3A).

In the Flp-In cell line, two bands were clearly visible upon pIMAGO analysis, and corresponded to the expected sizes of CFIm68 and CFIm59 (Figure 5.3A). This suggested that CFIm68 and CFIm59 are detectably phosphorylated in these cells. Further probing with CFIm68 and CFIm59 antibodies confirmed that these bands were at exactly the same position with the same appearance, providing support for the presumed identities. In the 68KD cell line, the upper band was lost, confirming that this band did indeed correspond to CFIm68. Upon stable re-expression of WT CFIm68 (68KD+CFIm68), the phosphorylated CFIm68 band re-appeared. This was associated with a loss of the lower phosphorylated band, which mirrored the decreased expression of CFIm59 (a result of CFIm68 overexpression - see Section 3.10). This supported the conclusion that the lower band represents phosphorylated CFIm59. Interestingly, there was no detectable phosphorylation of CFIm25.

To assess whether the RS domain may be partially responsible for the CFIm68 phosphorylation signal, the stable lines expressing RS domain mutants lacking the first or second half of the RS domain (ΔRS1 and ΔRS2, see Section 4.6, Figure 4.13) were also tested. It was readily apparent that both of these ΔRS mutant lines (68KD+CFIm68 ΔRS1 and 68KD+CFIm68 ΔRS2) had a clear decrease in the phosphorylation signal, despite similar total CFIm68 expression level (Figure 5.3A). Quantification across three experiments, with



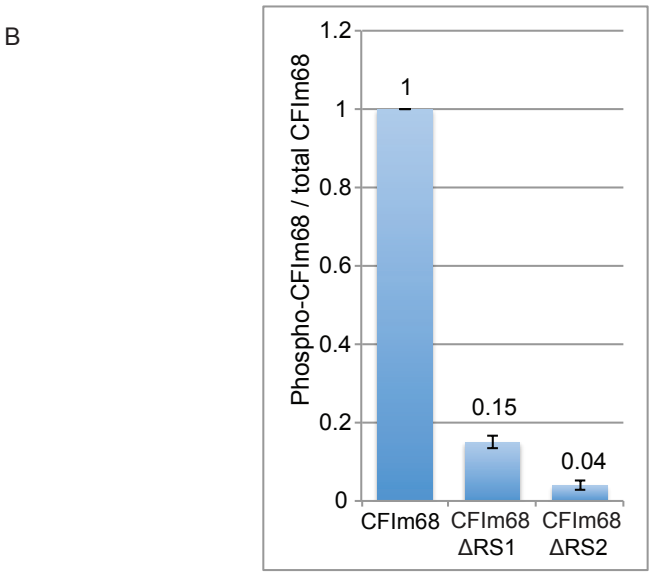


Figure 5.3: Identification of phosphorylation in the CFIm68 RS domain and CFIm59. (A) CFIm was immunoprecipitated from the cell lines indicated and samples were analysed by western blotting. Anti-GFP IP was used as a negative isotype control and phosphorylated beta-casein as a positive control. The nanoparticle pIMAGO reagent was used to detect phosphorylated proteins, followed by standard antibody-mediated detection of CFIm subunits. M = marker lane. (B) Band intensities for phosphorylated and total CFIm68 were quantified for the final three lanes (CFIm68 re-expressing stable cell lines) and results were expressed as a ratio of phospho-CFIm68/total CFIm68, standardised to a value of 1 in the 68KD + CFIm68 cell line. n=3, error bars depict ±SEM.

normalisation to total CFIm68, confirmed that there was a substantial and reproducible loss of phosphorylation for both mutants compared to WT: 85% loss for ΔRS1 and 96% loss for ΔRS2 (Figure 5.3B). This suggests that most CFIm68 phosphorylation occurs within the RS domain.

In order to identify the site(s) responsible for the CFIm68 RS phosphorylation signal, site-directed mutagenesis was carried out, with initial efforts focused on serine residues. The RS2 region was dissected in detail, as deletion of this yielded the largest drop in signal. All three serines were mutated to non-phosphorylatable alanine individually and in various combinations (S511A, S513A, S525A, S511A+S513A, S511A+S513A+S525A). To study the RS1 region, which appeared to make a smaller contribution to overall phosphorylation and had more potential sites, all eight serines were studied simultaneously. The approach here was slightly different, and involved annealing several oligonucleotides to generate a multi-site serine-to-alanine mutant (RS1 S-to-A – see Figure 5.2B for the eight residue identities). Interestingly, attempts to express a full RS domain S-to-A mutant with all 11 serines mutated simultaneously were unsuccessful, as for the earlier attempt to generate a full ΔRS mutant (Section 4.6), suggesting that such a completely non-phosphorylatable isoform may have a dominant negative lethal effect. There are also four tyrosines in the RS domain, which were analysed separately from serine by mutating to the non-phosphorylatable analogue phenylalanine (Y485F, Y520F+Y521F, Y548F). Stable cell lines were generated expressing all of these different CFIm68 mutant isoforms and analysed using the pIMAGO protocol as above.

None of the tyrosine mutants showed loss of phosphorylation, and so were not analysed further (Figure 5.4B). However, all serine mutants showed some loss of phosphorylation,

confirming that there is extensive serine phosphorylation at multiple sites within the RS domain (Figure 5.4A). Quantification across three repeats showed that the RS1 S-to-A mutant had a 45% decrease in signal, while the RS2 S-to-A (S511A+S513A+S525A) mutant had a 64% decrease in signal (Figure 5.4A). This is consistent with the finding that RS2 deletion had a bigger impact on phosphorylation loss than RS1 deletion (Figure 5.3).

While further, single-site mutagenesis studies would be required to identify the phosphorylated serines within RS1, the more detailed analysis of RS2 allowed conclusions to be drawn about the relative contribution of the three serines to phosphorylation in this region (Figure 5.4A). S511 and S513 appeared to represent the major phosphorylation sites, showing a similar extent of signal loss upon S-to-A mutagenesis (~36%), while S525A had a smaller impact (~14%). The effect of simultaneous S511A S513A mutagenesis was not additive, with only a small additional loss of signal compared to mutation of S511 or S513 alone, suggesting some degree of co-operativity in phosphorylation of these two residues. This would be consistent with the observation that other SR proteins are often subject to processive, multi-site phosphorylation of serines in close proximity by SRPKs (Ghosh and Adams, 2011). The additional signal loss upon additional S525A mutagenesis, however, appeared consistent with an additive effect.

It is interesting to note that the drop in phosphorylation signal observed in the full RS1 and RS2 S-to-A mutants did not fully recapitulate the signal loss observed upon deletion of the domains (45% vs. 85% for RS1 and 64% vs. 96% for RS2). This suggests that the RS1 and RS2 regions, as well as being direct phosphorylation targets, may also make an indirect contribution to phosphorylation elsewhere within CFIm68. For example, altered protein conformation upon deletion of the RS1 domain may impair site accessibility or kinase

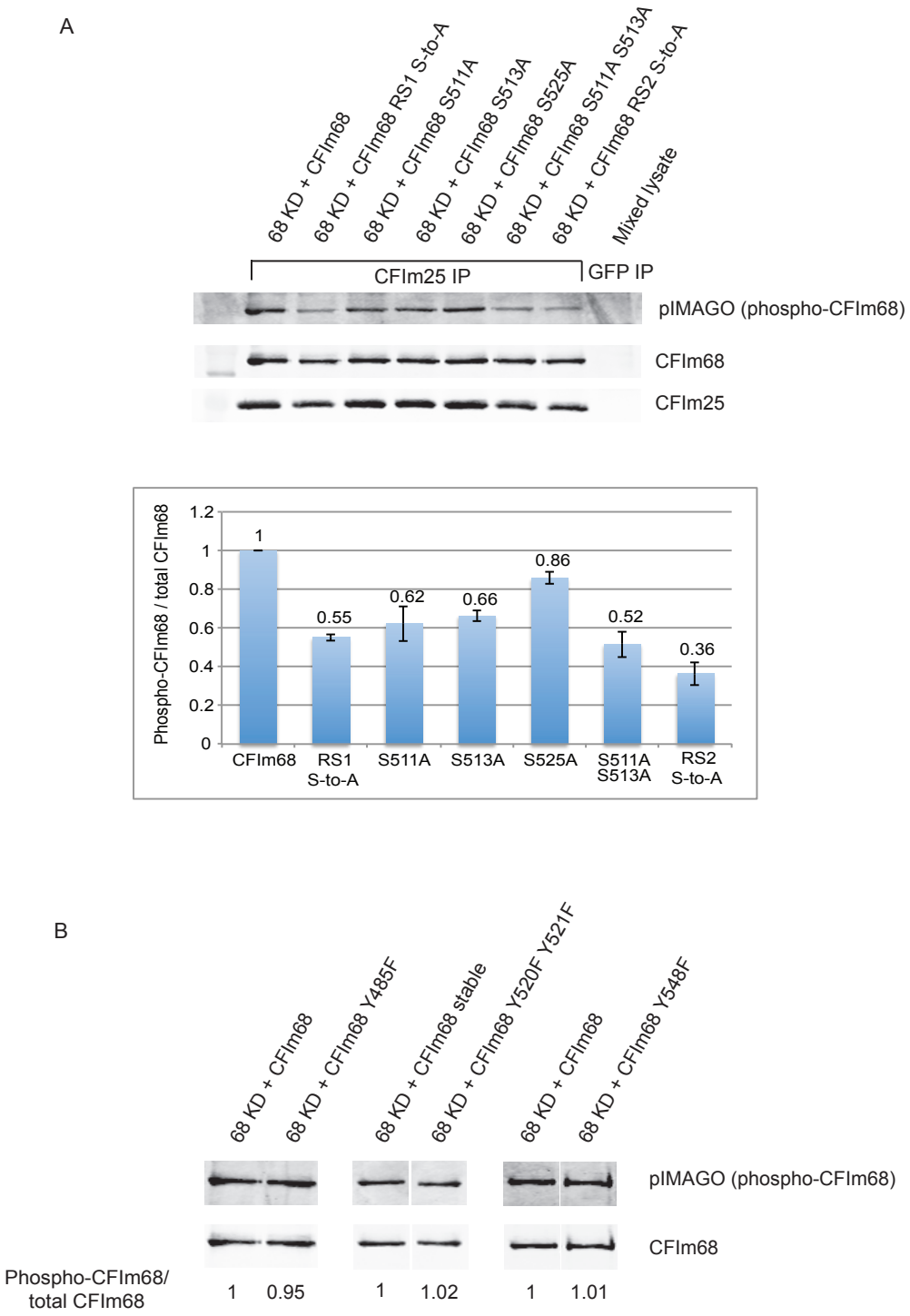


Figure 5.4: The CFIm68 RS domain is extensively serine-phosphorylated. CFIm was immunoprecipitated from the given cell lines and analysed by western blotting. The pIMAGO reagent was used to detect phosphorylated proteins, followed by standard antibody-mediated detection of CFIm subunits. (A) A representative blot showing phospho-CFIm68 levels for CFIm68 serine-to-alanine mutants. RS1 S-to-A specifies mutagenesis of all eight serines in the RS1 region; RS2 S-to-A specifies mutagenesis of all three in the RS2 region (S511, S513, S525). Anti-GFP IP was used as a negative isotype control. The graph shows mean quantification across repeats. Band intensities for phosphorylated and total CFIm68 were quantified and results were expressed as a ratio of phospho-CFIm68/total CFIm68, standardised to a value of 1 in the 68KD+CFIm68 cell line. n=3, error bars depict ±SEM. (B) As in (A), for CFIm68 tyrosine-to-phenylalanine mutants in the RS domain (each Y-to-F mutant analysed on a separate blot). Quantification is for the blots shown only and was performed as described in (A), standardised to a value of 1 for 68KD+CFIm68 in each case.

recruitment within the RS2 domain, such that deletion of RS1 leads to a concomitant drop in RS2 phosphorylation.

Efforts were next turned to identification of the kinase(s) responsible for the identified serine phosphorylation. The SRPK and CLK families, discussed in Section 1.5.4, represent good candidates, as they are involved in catalysing serine phosphorylation within RS domains of SR splicing proteins (Zhou and Fu, 2013). Experiments were therefore performed to investigate whether treatment of 293 Flp-In cells with inhibitors of these kinases would lead to loss of CFIm68 RS domain phosphorylation. Cells were treated with SRPIN340, an inhibitor of SRPK1 and 2 (Fukuhara et al., 2006), or separately with TG003, an inhibitor of CLK1, CLK2 and CLK4 (Muraki et al., 2004). Neither compound had any clear effect on CFIm68 phosphorylation levels as assessed by pIMAGO analysis, over a range of cell types, treatment concentrations and treatment times (Figure 5.5A, shown only for SRPIN340). However, analysis of whole cell lysates from these experiments by western blotting with the 1H4 broadly reactive anti-phospho-SR antibody (Neugebauer and Roth, 1997) suggested that there was no appreciable loss in phosphorylation even for known SRPK/CLK targets following treatment with the inhibitors (Figure 5.5B – shown for SRPIN340 treatment only; one example of many conditions trialled). As the inhibitors did not appear to be functioning as expected, the results surrounding CFIm68 were inconclusive – SRPKs or CLKs may or may not be involved in CFIm68 RS domain phosphorylation, and further experimental optimisation or alternative approaches would be needed to clarify this.

In summary, pIMAGO phosphorylation analysis coupled with site-directed mutagenesis allowed identification of multiple serine phosphorylation sites in the CFIm68 RS domain.

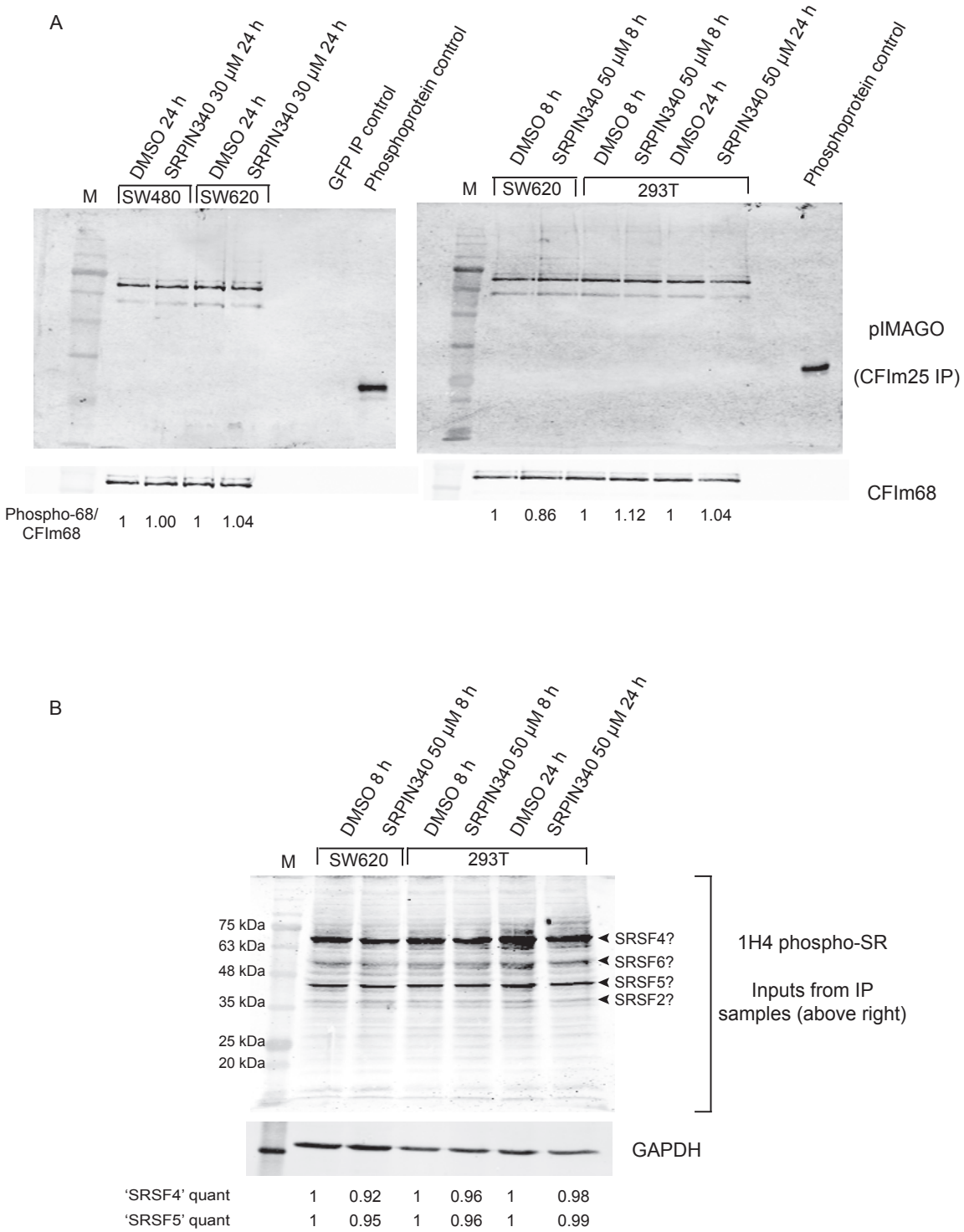


Figure 5.5: Use of the SRPK inhibitor SRPIN340 is inconclusive regarding potential roles of SRPKs in CFIm68 phosphorylation. Cells (SW480/SW620 colon cancer lines or 293T) were treated with the SRPIN340 inhibitor or DMSO control at the given concentrations for the given times, following which a CFIm25 IP and pIMAGO western blot was performed as described in Figure 5.3. (A) pIMAGO analysis of CFIm68/CFIm59 phosphorylation in the IP samples following treatments. The two blots represent two independent experiments. (B) 1H4 phospho-SR antibody probing of input samples (from IP experiment, top-right) was performed to assess phosphorylation of known SRPK targets for each treatment condition. SRSF protein identities were assumed based on protein size and comparison with (Neugebauer and Roth, 1997). Intensities of the two strongest bands were quantified (SRSF4/SRSF5 quant) and normalised to GAPDH loading control. M = marker lane.

This leaves open opportunities for further study of the functional roles and modification pathways associated with these phosphorylation events.

5.5 Analysis of CFIm25 PTM site mutants in the APA complementation assay

The combination of the above-described approaches (public database searches and MS/pIMAGO analysis of CFIm) yielded a list of CFIm PTMs deemed to be of potential biological interest. These include the PTMs summarised across Tables 5.1 and 5.2, as well as the CFIm68 RS domain serine phosphorylation sites. Having generated this list, the CFIm25 and CFIm68 complementation assays were used to investigate the possibility that one or more of these PTMs may be involved in regulating APA.

For CFIm25, site-directed mutagenesis was used to generate point mutations at the sites of interest using the pcDNA3.1-CFIm25 plasmid. Modification-ablating mutations (K-to-R, Y-to-F, T-to-A, R-to-A) were tested for each site, and for one acetylation site a K-to-E mutation was also studied, as discussed below. The CFIm25 mutants were expressed by transient transfection as previously outlined, with three repeat experiments carried out for each PTM, and analysed in the complementation assay.

None of the modification-null mutants studied showed any clear difference from WT CFIm25 in complementation ability, suggesting that, at steady state, these PTMs do not play a major role in regulating poly(A) site choice by CFIm25 (Figures 5.6, 5.7 and 5.8). Two example northern blots, encompassing four PTM site mutants, are shown in Figures 5.6 and 5.7, and full APA quantification for all mutants studied is given in Figure 5.8.

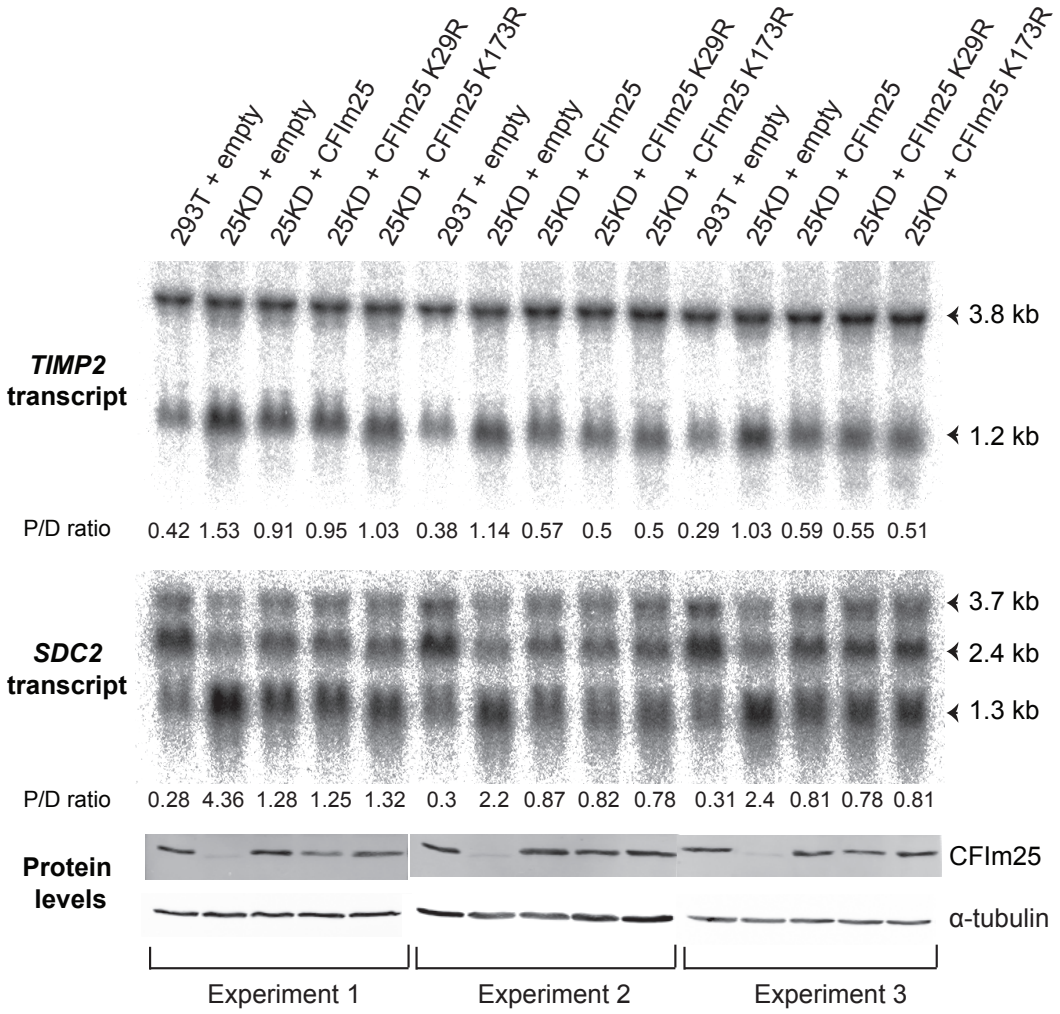


Figure 5.6: Acetylation-ablating mutagenesis of lysine 29 or lysine 173 on CFIm25 does not alter APA regulation in the *TIMP2* and *SDC2* genes. Northern blotting and P/D ratio calculations were performed on the given samples following transfection experiments (three biological repeats - experiments 1-3) as described in Figure 3.3. The corresponding western blot at the bottom shows CFIm25 expression in the transfected cells, with α-tubulin as a loading control.

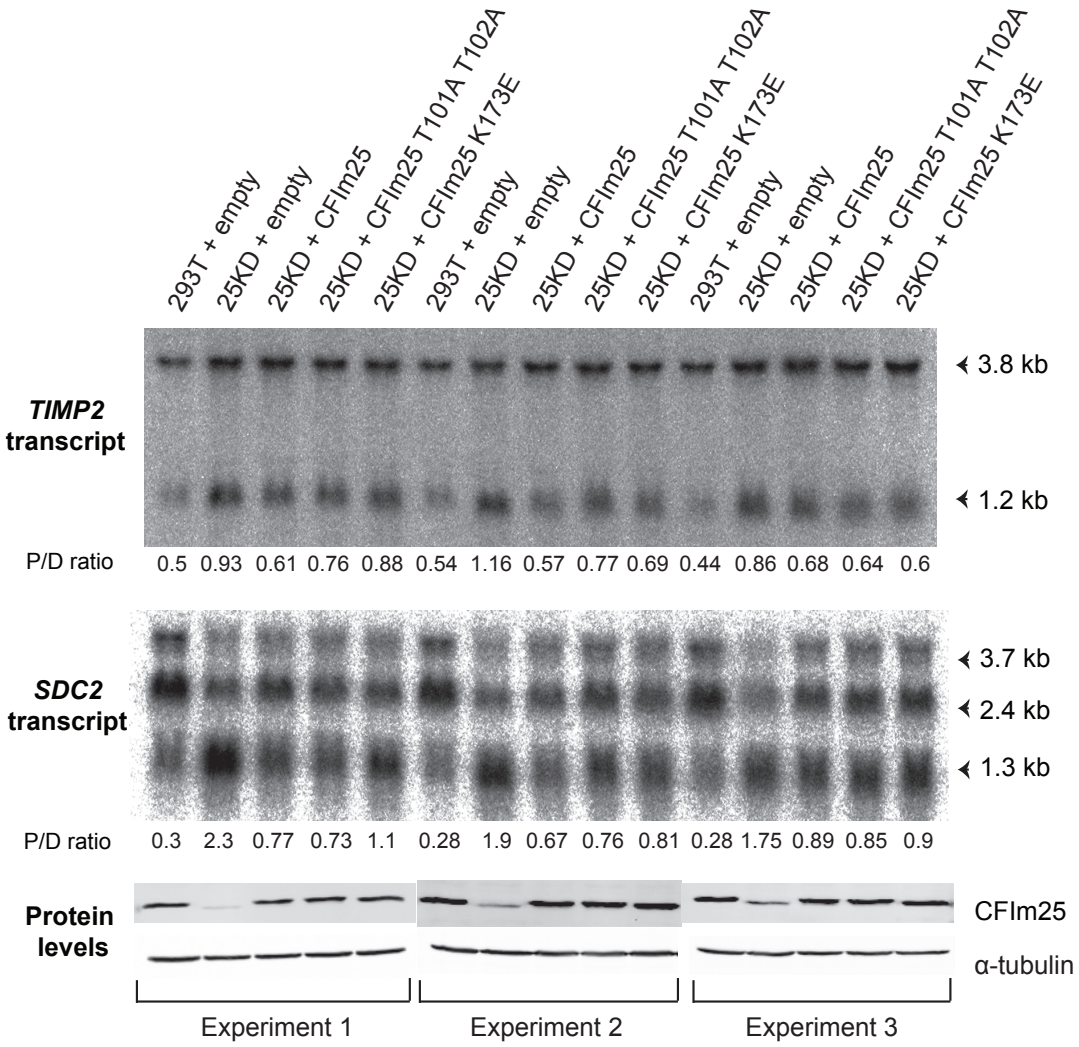


Figure 5.7: Phosphorylation-ablating mutagenesis of threonines 101 and 102 or glutamate substitution at lysine 173 on CFIm25 does not alter APA regulation in the *TIMP2* and *SDC2* genes. Northern blotting and P/D ratio calculations were performed on the given samples following transfection experiments (three biological repeats - experiments 1-3) as described in Figure 3.3. The corresponding western blot at the bottom shows CFIm25 expression in the transfected cells, with α-tubulin as a loading control.

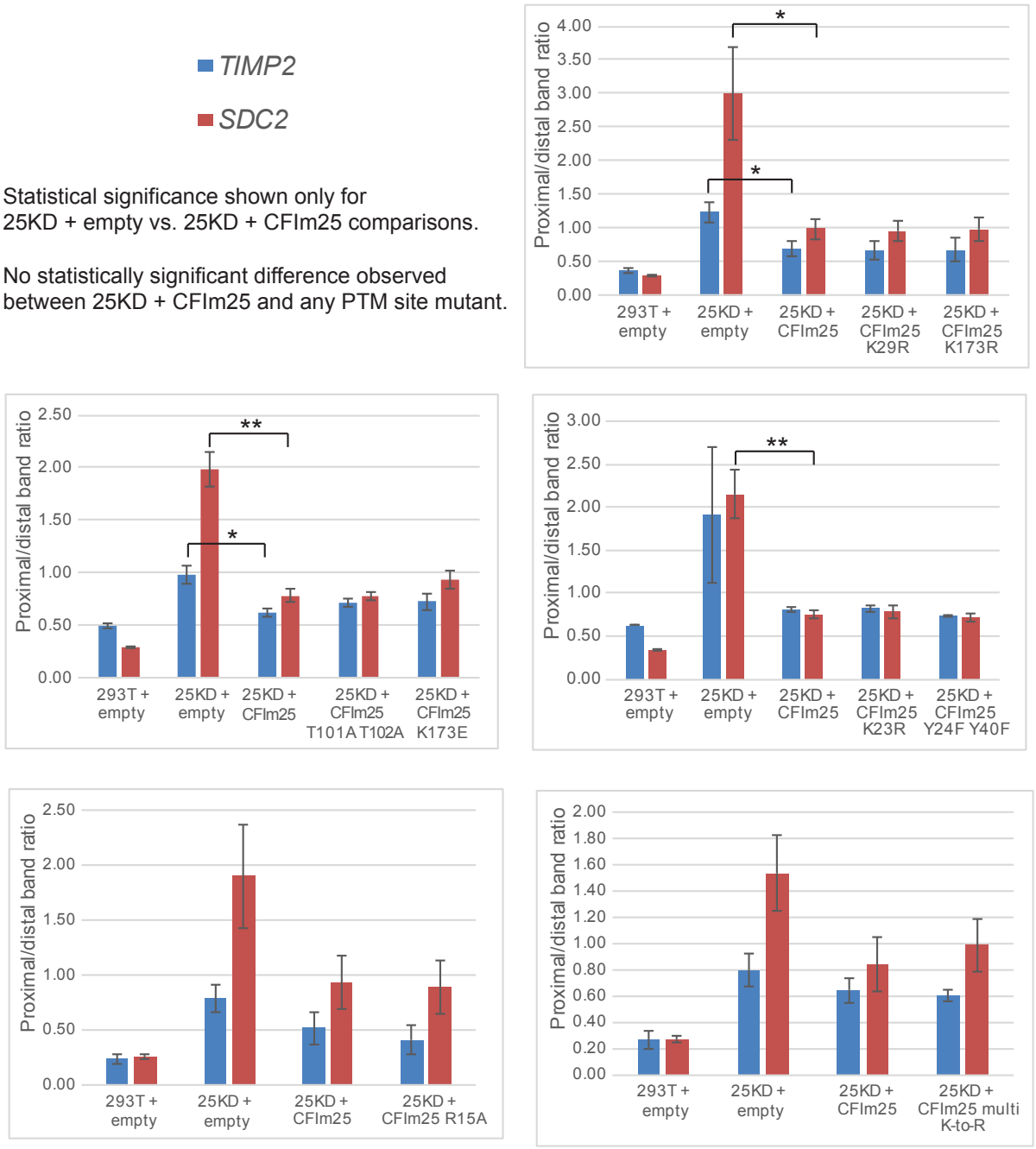


Figure 5.8: Mutagenesis at a range of PTM sites on CFIm25 does not alter APA regulation. Northern blotting and P/D ratio calculations were performed on the given samples following transfection experiments as described in Figure 3.3. Multi K-to-R includes K23R, K29R, K56R, K73R, K122R, K173R, K189R. Here, each graph represents a single northern blot encompassing three biological repeats (exception - n=2 for bottom-right, multi K-to-R), with the mean proximal/distal ratio plotted for each sample. For example, for the top-right graph, one transfection experiment (of three total) involved five samples, with comparison of K29R and K173R simultaneously to the same 293T/25KD+empty/25KD+CFIm25 samples (see Figure 5.6). Error bars depict ±SEM. Statistical significance was calculated using a two-tailed Student's t-test, only for comparisons between (25KD+empty vs. 25KD+CFIm25), and (25KD+CFIm25 WT vs. 25KD+CFIm25 PTM mutant).

The novel acetylation site identified at K173 on CFIm25 was of particular interest due to the positioning of K173 close to the dimer interface of the CFIm25 homodimer, and hence the possibility that acetylation may regulate dimer formation. In a CFIm25-CFIm68 crystal structure (Yang et al., 2011a), K173 can be seen to form a salt bridge with E117 within the same CFIm25 monomer, which acetylation would be expected to disrupt due to loss of the positive charge. Therefore, despite the lack of APA phenotype upon acetyl-ablating K173R mutagenesis (Figure 5.6, 5.8), a K173E mutant was also studied to mimic the effect of acetylation. The negatively charged E173 would be expected to repel E117 and abolish the salt bridge, in a manner similar to that expected to result from acetylation. Study of this mutant, however, showed that K173E mutation also had no effect on CFIm25 complementation ability (Figure 5.7, 5.8), suggesting acetylation at this site is not involved in APA regulation.

The double T101A T102A mutant (Figure 5.7) was of interest due to the proximity of these threonine residues to phenylalanines 103 and 104, both of which have a well-established role in UGUA sequence recognition by CFIm25 (see Section 4.5). Despite only limited evidence for phosphorylation at these sites (Olsen et al., 2010), the potential for phosphorylation to influence hydrogen bonding/stacking interactions of F103 and F104, or to make an additional contribution to UGUA recognition, may have been expected to influence poly(A) site choice at one or both of these transcripts. However, the T101A T102A mutant behaved indistinguishably from WT CFIm25 in the complementation assay (Figure 5.7, 5.8), suggesting that abrogating these potential phosphorylation events does not impact upon APA.

Overall, no CFIm25 PTM site mutants studied were deemed to show any functional difference to WT CFIm25 in the complementation assay. Specifically, this conclusion could

confidently be drawn where 25KD+CFIm25 showed a significantly different P/D ratio to 25KD+empty (i.e. successful complementation), but no significant difference in P/D ratio to the PTM site mutant studied (i.e. mutant complements equivalently to WT; Figure 5.8). This was the case for K29R, K173R, T101A T102A, K173E (both transcripts), and K23R and Y24F Y40F (*SDC2* only). Although the 25KD+CFIm25 complementation was not significant for *TIMP2* in the K23R/Y24F Y40F experiment, nor for either transcript in the R15A and multi K-to-R experiments, these mutants were still deemed indistinguishable from WT CFIm25 and not studied further. The lack of statistical significance likely resulted from large P/D ratio variation in the 25KD+empty sample; the striking similarity in the P/D ratios between WT and mutants in these cases, particularly for *SDC2*, raised confidence that the PTM site mutations had no effect on CFIm25 function (Figure 5.8).

There are a number of possible explanations for the observed lack of functional effect of these PTM site mutations. The simplest is that these modifications do not regulate the activity of CFIm25 in poly(A) site selection, and perhaps have some other role. For example, although loss of K23 acetylation in CFIm25 might be expected to enhance the CFIm25-PAP interaction (Shimazu et al., 2007), this would not necessarily alter poly(A) site choice by CFIm. Another explanation would be that some CFIm25 PTMs exert transcript-specific or cell type/condition-specific effects, which are missed by studying only *TIMP2* and *SDC2* in HEK293 cells under standard culture conditions.

Furthermore, it may be that these modifications are not present at high steady-state levels in these cells, and that mutations which mimic, rather than ablate, the modification would reveal a phenotype. This was not the case for K173E, and it can be difficult to identify a suitable 'modification-mimicking' substitution (e.g. for tyrosine phosphorylation), but it would be of

interest, for example, to check whether phospho-mimetic T101D T102D may reveal an effect where T101A T101A did not. Indeed, without prior mechanistic knowledge, it is equally possible that modification or loss of modification may drive changes in poly(A) site choice.

Finally, there remains the possibility that altering only one or two PTM sites at a time does not alter CFIm25 activity, but that combinatorial changes may. This was partially explored through construction of a 'multi K-to-R' mutant, with all reported acetylation sites from my studies and the PSP database mutated (K23R,K29R,K56R,K73R,K122R,K173R,K189R). Despite this mutant being expressed at levels lower than WT CFIm25 (not shown), possibly due to protein instability, preliminary indications were that it was still able to complement to a similar extent as WT CFIm25 (Figure 5.8, bottom right). More repeats would be needed to assess statistical significance here, but this suggests that even complete loss of CFIm25 acetylation does not lead to significantly altered function in APA.

5.6 Analysis of CFIm68 PTM site mutants in the APA complementation assay

A similar approach was used to assess CFIm68 PTM sites of interest, this time using the stable cell line system for expression of the mutants in the 68KD background as previously described. In most cases only a single RNA sample from a single cell line was analysed here, due to the more stable and robust complementation achievable with this system. Northern blots for the full range of PTM site mutants studied are presented in Figure 5.9 (non-SR domain mutants) and Figures 5.10 and 5.11 (SR domain serine mutants). As for CFIm25, no

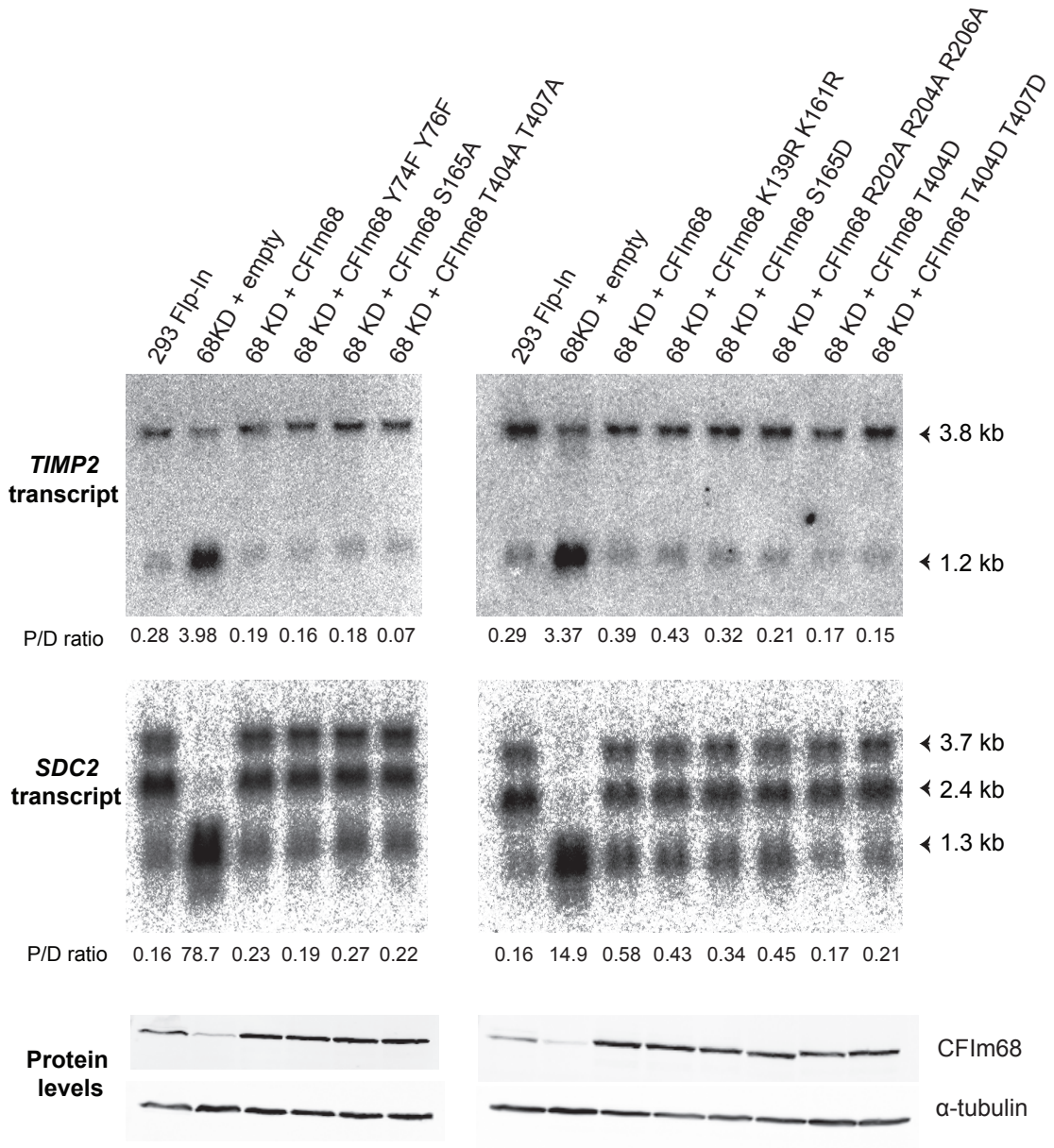


Figure 5.9: Mutagenesis at a range of PTM sites in CFIm68 does not alter APA regulation in the *TIMP2* and *SDC2* genes. Northern blotting and P/D ratio calculations were performed as described in Figure 3.3, on total RNA isolated from the given cell lines. The corresponding western blot at the bottom shows CFIm68 expression in the cells, with α-tubulin as a loading control.

modification-null mutants showed any obvious alteration in complementation compared to WT CFIm68, with all able to fully reverse the 3'UTR shortening induced by CFIm68 depletion. For example, triple R-to-A mutagenesis in the GAR motif, which is subject to methylation (Martin et al., 2010), did not alter the APA profile (Figure 5.9), consistent with a previous study that could not identify a clear role for this methylation. In some cases, phosphomimetic mutants were also analysed – for example, S165D was studied as well as S165A due to reports that the phosphomimetic mutation enhances RNA binding affinity (Yang et al., 2011b). However, this phosphomimetic mutant did not show any obvious difference in function to WT CFIm68 (Figure 5.9).

While the majority of phosphomimetic and phospho-ablating RS domain serine mutations similarly did not show any obvious change in complementation (Figure 5.10), a mutant with serine-to-aspartate conversion at all eight serines in the RS1 region (RS1 S-to-D; see Figures 4.13 and 5.2 for serine identities) did show a striking complementation defect for both transcripts studied (Figure 5.11). The P/D ratio was higher than that observed for the ΔRS1 mutant (Figure 4.14), suggesting that phosphomimetic mutation in this region impairs distal poly(A) site selection even more than deletion of the whole RS1 region. This finding raises the possibility that phosphorylation within the CFIm68 RS1 domain may physiologically regulate the activity of CFIm, and thus poly(A) site selection, in vivo. The equivalent phospho-ablating mutant, RS1 S-to-A, showed a small decrease in P/D ratio compared to both WT re-expressed CFIm68 and the parental Flp-In line. (Figure 5.10, 5.11). While further repeats would be required to assess whether this subtle change in P/D ratio is significant, the fact that phospho-ablating mutagenesis yielded the reverse phenotype to phospho-mimetic mutagenesis strongly supports the suggestion that this reflects a functional impact of phosphorylation.

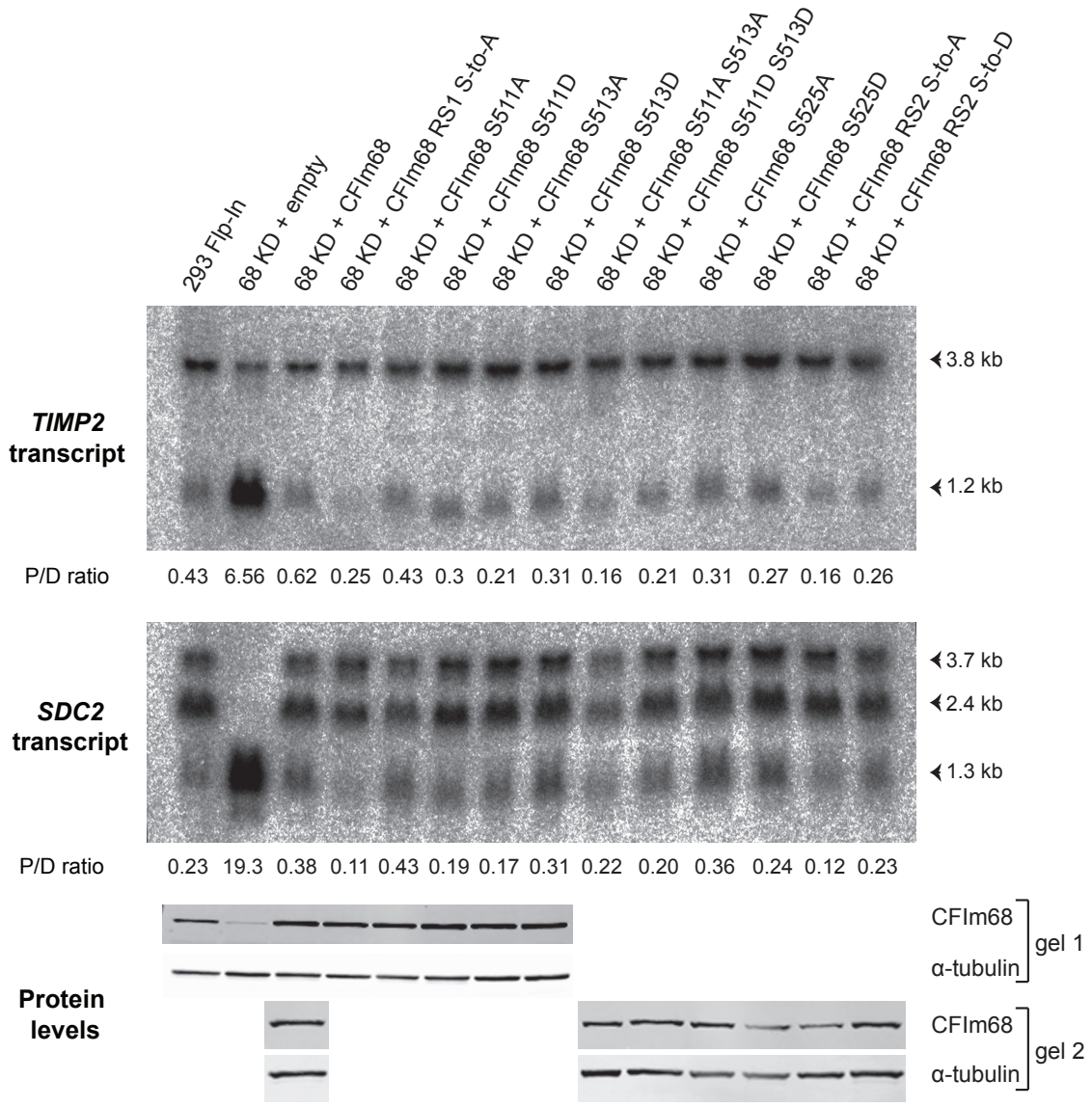


Figure 5.10: Phospho-ablating mutagenesis in the CFIm68 RS1/RS2 domains and phospho-mimetic mutagenesis in the RS2 domain does not appreciably alter APA regulation in the *TIMP2* and *SDC2* genes. Northern blotting and P/D ratio calculations were performed as described in Figure 3.3, on total RNA isolated from the given cell lines. The corresponding western blots at the bottom shows CFIm68 expression in the cells, with α-tubulin as a loading control. Samples were run across two western blots as indicated, and wild type (68KD+CFIm68) sample was included on both gels to allow direct comparison with all mutants.

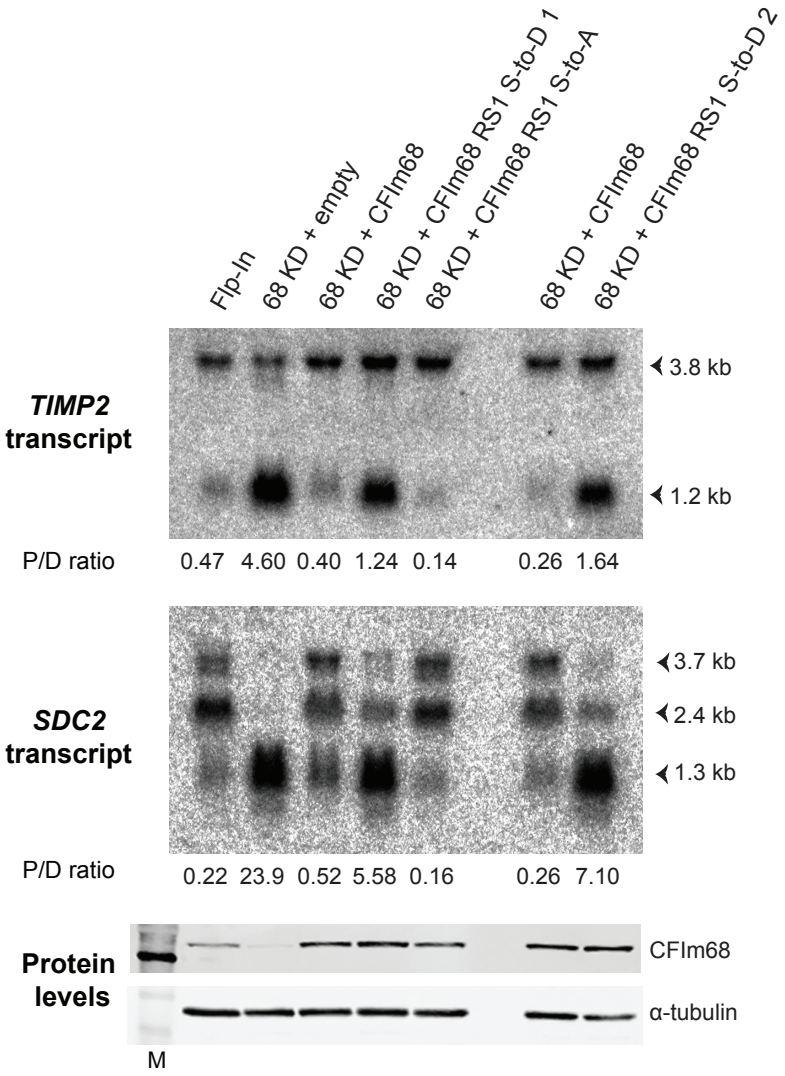


Figure 5.11: Phospho-mimetic mutagenesis of serines in the CFIm68 RS1 alters APA regulation in the *TIMP2* and *SDC2* genes. Northern blotting and P/D ratio calculations were performed as described in Figure 3.3, on total RNA isolated from the given cell lines. RS1 S-to-D 1 and 2 represent two biological repeats (distinct RNA samples). The corresponding western blots at the bottom shows CFIm68 expression in the cells, with α-tubulin as a loading control.

Similar mutagenesis in the RS2 region, with conversion of all three serines to alanine or aspartate, did not reveal any obvious effects on complementation (Figure 5.10), despite the fact that deletion of this domain yielded a more dramatic APA effect than deletion of the RS1 domain (Figure 4.14). This suggests that the observed effect does not arise from multi-site phosphorylation within the RS domain in general, but rather from phosphorylation of specific serines within the RS1 region. An important next step would be to analyse the eight serines in this region individually and in combination, to discern which phosphorylation events are responsible for the altered poly(A) site selection. In addition, a number of mechanistic explanations for the observed effect could be explored, such as potential altered protein-protein interactions or subcellular localisation associated with this CFIm68 phosphorylation.

In summary, this analysis found that many of the PTMs studied do not play any major role in regulating poly(A) site selection by CFIm25 or CFIm68 under the conditions studied (with the caveats as discussed earlier, such as potential transcript-specific effects or combinatorial effects of groups of PTMs not studied simultaneously here). Study of these PTMs at the transcriptome-wide level, in diverse cellular conditions, and perhaps with alternative functional assays, may be needed to elucidate what contribution, if any, they make to CFIm function. The finding that mimicry of phosphorylation in the RS1 domain does impact upon poly(A) site selection by CFIm68, however, leaves open avenues for further study of the potential role of phosphorylation at these sites. Validation of the observed effects at the transcriptome-wide level and mechanistic investigations, followed by dissection of the cellular conditions and upstream signalling pathways associated with this phosphorylation, may provide further insights into the physiological relevance of the findings.

Chapter 6
Discussion

This study set out to characterise more fully the contribution of CFIm to the regulation of APA in human cells. Here, the main findings of the work will be discussed in the context of existing knowledge, with a consideration of possible future directions for research into CFIm function.

6.1 Roles of individual CFIm subunits and the interplay between them

The generation of the 25KD and 68KD CFIm 'knockdown' cell lines as the basis for a complementation assay in itself provided a number of insights into the interplay between subunits within the CFIm complex and the function of the CFIm25 and CFIm68 subunits. Firstly, the lack of success in generating cell lines with complete loss of CFIm25 or CFIm68 expression suggested that these may be essential proteins. This is particularly likely for CFIm25, where no knockout was identified despite very efficient *NUDT21* gene editing and extensive clone screening. Consistently, *NUDT21* has been identified as a likely essential gene in genome-wide CRISPR screens (Hart et al., 2015; Wang et al., 2015). The suggestion that CFIm68 is essential, on the other hand, is inconsistent with a study that reported generation of *CPSF6* knockout cell lines, though it cannot be ruled out that a low residual level of CFIm68 was below the detection limit of the western blot technique used in that study (Sowd et al., 2016). The observed slow growth rate of the 25KD and 68KD cell lines, which was particularly apparent for 68KD, supports the view that both subunits play an important role in cell fitness, whether or not they are essential, which is consistent with their reported fundamental role in CPA (Takagaki et al., 1989). It is possible that some complete knockout clones were generated, but grew poorly and were perhaps therefore not picked up at the cloning stage.

The observed 3'UTR shortening in two known APA-regulated transcripts in both the 25KD and 68KD cell lines confirmed that the reported effects of transient CFIm25/CFIm68 depletion (Gruber et al., 2012; Kim et al., 2010; Kubo et al., 2006) can persist upon longer-term depletion of these subunits. However, while for CFIm68 there was a clear dose-dependence between the degree of CFIm68 depletion and 3'UTR shortening, with the two lines with >90% depletion both showing the same dramatic poly(A) site shift, the effects of CFIm25 depletion were more variable. It was very surprising that only one CFIm25-depleted cell line showed the expected 3'UTR shortening, while others with a similar extent of CFIm25 loss showed no obvious change in poly(A) site use. The reason for this difference between the range of CFIm25/CFIm68 depletion phenotypes is unclear, but perhaps suggests a compensatory mechanism can operate as a long-term response to CFIm25 depletion to return APA patterns to the unperturbed parental state. This may involve, for example, decreased expression of other CPA factors such as CstF-64, which is known to promote proximal poly(A) site use (Yao et al., 2012). Such a compensatory mechanism may have shown clone-specific inactivation in the 25KD line, and may not be capable of overcoming the effects of CFIm68 depletion. Analysis of the expression levels of other CPA factors in the various cell lines would be a useful next step in exploring these possibilities. Despite these unusual observations, the fact that the APA shift in 25KD could be complemented by CFIm25 re-expression confirmed a CFIm25-dependent effect and allowed this cell line to be used as a tool to explore CFIm25 function.

The observed effects of CFIm subunit depletion on the expression levels of other CFIm subunits also gave insights into subunit interdependency within the complex as a whole, as discussed in Section 3.10. Briefly, the two large subunits showed very different responses to CFIm25 depletion, with CFIm59 showing almost complete co-depletion while there was no change in the level of CFIm68. In addition, CFIm68 over-expression led to decreased levels of CFIm59. This would be consistent with a model whereby CFIm68 is preferentially incorporated into the CFIm complex, and any unincorporated CFIm59, either resulting from decreased CFIm25 expression or increased CFIm68 incorporation, is unstable and subject to rapid proteolysis. The observation that CFIm68 levels did not decrease upon CFIm25 depletion suggests that it may not be as unstable as CFIm59, and in addition that CFIm25

may be in excess, such that the CFIm25-binding capacity of CFIm68 is saturated and CFIm68 is limiting. This latter suggestion is clearly supported by findings from CFIm25 IP experiments, in which only a small increase in CFIm68 pull-down was achieved with a large increase in CFIm25 WT expression (Section 4.2). There was no increase in CFIm59 abundance in response to CFIm68 depletion, and CFIm25 was also co-depleted upon CFIm68 depletion, further supporting the conclusion that CFIm68 and CFIm59 are functionally distinct and cannot compensate for each other. These findings relating to CFIm stoichiometry are largely consistent with previous observations from transient CFIm knockdown studies, other than a finding from one study that CFIm25 depletion led to a concomitant decrease in CFIm68 levels – this may have reflected the higher degree of CFIm25 knockdown in that study, which may pass the threshold saturation point of CFIm68 (Gruber et al., 2012; Kim et al., 2010).

The homeostatic regulation between CFIm subunits discussed here seems to be most convincingly explained by stability effects, whereby CFIm monomers that are not complex-incorporated, particularly CFIm25 and CFIm59, are vulnerable to proteolytic degradation. The best supporting evidence for this is the observation that expression of the CFIm68 Δ116-122 mutant, which cannot interact with CFIm25, did not restore the CFIm25 level in 68KD cells to that observed in the parental cell line, suggesting that CFIm25 levels remain low due to a lack of complex incorporation (Section 4.3). Nonetheless, homeostatic regulation between the CFIm subunits via post-transcriptional RNA-based mechanisms cannot be ruled out as an additional or alternative explanation for the findings. For example, CFIm transcripts have APA sites, with associated CFIm binding sites, and the potential of CFIm to regulate APA within the *CPSF6* 3'UTR has been demonstrated (Brown and Gilmartin, 2003; Sartini et al., 2008). Measurement of CFIm RNA levels, APA patterns and translational efficiency in the KD cell lines would be an important next step in exploring this further.

The importance of interplay between CFIm subunits was more directly explored through abolition of the CFIm25-CFIm68 interaction. The observations that CFIm68 can direct HIV-1 integration to active genes independently of its interaction with CFIm25 (Rasheedi et al., 2016), and that CFIm68 functions

independently in histone 3' end processing (Ruepp et al., 2010), had raised the possibility that these two subunits may have separable functions in poly(A) site selection. However, the failure of the CFIm25 Y158A Y160A and CFIm68 Δ116-122 mutants to restore distal poly(A) site use in the complementation assay confirmed that the CFIm25-CFIm68 interaction is essential for the function of CFIm in poly(A) site selection. So, while CFIm68 may have the potential to be recruited to RNA and perform certain functions as an individual subunit, this does not extend to its contribution to CPA in protein-coding genes.

The CFIm59 subunit was not extensively studied here, due to the lack of any reported role of this subunit in poly(A) site selection. However, future generation of a 59KD cell line would be useful to complete the analysis of CFIm subunit interdependency and confirm previous observations regarding the lack of a role for CFIm59 in APA. Furthermore, for completeness, ideally more APA analyses would be extended to the transcriptome-wide level – while the generality of the 25KD-induced 3'UTR shortening was confirmed globally, it would be interesting to compare this to the 68KD cell line. A previous study found a high degree of correlation between the identities of CFIm25- and CFIm68-regulated transcripts in C2C12 myoblasts (Li et al., 2015), consistent with the idea that these subunits function together in poly(A) site selection, and it would be of interest to see if this holds true in HEK293 cells.

6.2 Mechanistic insights into regulation of poly(A) site selection by CFIm

While the APA phenotype resulting from CFIm depletion had been well established before this project began, the key question of how CFIm preferentially promotes selection of distal poly(A) sites remained unanswered. More generally, the contribution of CFIm to the CPA reaction, and in particular the role of protein-protein interactions, had not been defined in vivo. The KD cell lines and complementation assays based on CFIm25 and CFIm68 function allowed certain aspects of these questions to be addressed.

In particular, work presented in Section 4.5 provided the first in vivo demonstration that the UGUA-binding capacity of CFIm25 makes a crucial contribution to poly(A) site selection for the majority of CFIm-regulated transcripts. Numerous studies had already suggested an important role for UGUA recognition in CFIm function. Following identification of UGUA as the preferred CFIm binding motif and the demonstration that mutation of this sequence abolishes CPA in vitro (Brown and Gilmartin, 2003; Venkataraman et al., 2005), genome-wide studies found enrichment of UGUA sequences upstream of dominant, CFIm-regulated poly(A) sites and identified enrichment of the UGUA motif in CFIm CLIP binding sites (Li et al., 2015; Martin et al., 2012). While these studies suggested that CFIm binds at UGUA sequences and generally stimulates cleavage at associated distal poly(A) sites, observations were correlative, and none of them directly assessed the effect of abolishing UGUA binding in vivo. This study therefore adds important confirmation that the observed UGUA-binding function of CFIm25 is essential for poly(A) site selection for many transcripts in vivo.

It was hoped that the transcriptome-wide datasets for the UGUA-binding mutant would allow a more complex analysis of how UGUA binding promotes distal poly(A) site selection, but limitations of the technique prevented this. The transient nature of CFIm25 re-expression and partial complementation probably contributed to the limited reproducibility of sequencing results, and the outcome was that transcript groups confidently identified in both experiments were too small to perform powerful motif analyses. For example, enrichment of UGUA sequences upstream of distal poly(A) sites could only be identified when comparing CFIm25-regulated transcripts to shuffled sequences, and not to non-CFIm25-regulated controls, and the group of UGUA-independent transcripts identified was too small to use as a control group for comparison with UGUA-dependent transcripts. Therefore, while consideration of these data in conjunction with previous studies still favours a model in which CFIm is mainly recruited to distal sites, generally via UGUA binding, and stimulates their use, the data presented here do not allow confirmation of this.

Another avenue of interest was how CFIm may be recruited in the absence of UGUA binding, as had been demonstrated by previous CLIP studies (Martin et al., 2012). Indeed, while the group of

seemingly UGUA-independent CFIm-regulated transcripts identified in this study was again too small for group analysis, individual validation and further study of such transcripts could be informative. In particular, it would be interesting to identify other motifs or proteins that are important for CFIm recruitment where UGUA sequences are lacking. For future investigations, sequencing analysis would ideally be coupled with a more stable complementation system for CFIm25 function, for example using the Flp-In approach as for the 68KD cell line; this may remove some of the inherent variability in the data and allow more detailed mechanistic conclusions to be drawn.

In an alternative approach to the dissection of CFIm function, ChIP-seq analysis of genome-wide pol II distribution in the 25KD and 68KD cell lines was performed (Section 4.8). The lack of transcriptional read-through observed did not support the hypothesis that failed cleavage at the distal site in the absence of CFIm is the reason for the observed 3'UTR shortening, and in general the lack of any obvious transcriptional phenotype at the 3' end was surprising given the known contribution of CFIm to CPA. It may be that the ChIP-seq technique was not sensitive enough to detect small changes in pausing or read-through, particularly against the background of larger changes in pol II occupancy at the TSS and gene body. More direct approaches focused on nascent RNA, such as ChrRNA-seq or the single-nucleotide resolution mNET-seq may be useful for future studies, and indeed have revealed termination defects upon depletion of other CPA factors such as CPSF-73 and CstF-64 (Nojima et al., 2015). As they stand, these data do not help to distinguish between possible models for CFIm stimulatory/repressive function at APA sites.

The observed decrease of pol II occupancy at the TSS and gene body upon CFIm depletion, however, particularly in 68KD cells, was an unexpected and intriguing observation. Recruitment of CFIm and other cleavage factors at the TSS has been previously reported, with THOC5 playing an important role in the recruitment of CFIm68 at some genes (Katahira et al., 2013; Nojima et al., 2015; Venkataraman et al., 2005). However, no link has previously been made between the presence of CFIm near the TSS and recruitment, initiation, pausing or elongation of pol II. Single gene validation of the observed phenomenon would be an important next step, following which the cause and effect of the decreased

pol II occupancy should be explored. Analysis of APA patterns at appropriate model genes after phenocopying this altered occupancy, perhaps through alpha-amanitin or triptolide treatment (Bensaude, 2011), would be of particular interest to establish whether decreased pol II levels throughout genes may have a causal role in the increased proximal poly(A) site use associated with CFIm depletion.

The final contribution of this thesis to understanding of CFIm mechanism was the novel demonstration that the CFIm68 RS domain contributes to poly(A) site selection, and these findings will be discussed further below.

6.3 A role for the CFIm68 RS domain in poly(A) site selection

The novel observation that CFIm68 mutants lacking portions of their RS domain cannot fully restore distal poly(A) site use was conclusive for the two transcripts studied (Section 4.6), but awaits transcriptome-wide validation. The major outstanding question is how the RS domain contributes to CFIm function. The work presented here suggests that the contribution of the domain to poly(A) site selection goes beyond a role in nuclear import of CFIm68, as both the ΔRS1 and ΔRS2 mutants showed predominant nuclear localization. The slight discrepancy with previous reports that a ΔRS2 mutant with only additional loss of an RRHK sequence showed diffuse nuclear/cytoplasmic distribution (Dettwiler et al., 2004) was intriguing, and clearly the determinants of CFIm68 nuclear import need to be defined more precisely. The role of the TNPO3-CFIm68 RS interaction in nuclear import is well established, but the CFIm68 residues involved have not been defined as they have for the TNPO3-SRSF1 interaction (Maertens et al., 2014). Although unlikely for reasons previously discussed, there remains the possibility that the small decrease in nuclear signal intensity observed for the GFP-ΔRS mutants could partially explain the impaired complementation phenotype. Clarification on this matter could be gained using a more quantitative microscopy approach and by analysis of CFIm68 ΔRS1/ΔRS2 controls bearing an artificial NLS.

Assuming the function of the CFIm68 RS domain in nuclear localization is not responsible for the observed phenotypes, other possible explanations include an important role of RS domain-mediated speckle targeting of CFIm68 and/or RS domain-mediated protein-protein interactions. A previous report that the RS domain is involved in partial targeting of CFIm68 to nuclear speckles (Cardinale et al., 2007) was supported by the loss of speckling observed for the ΔRS mutants in this work. Nuclear speckles are thought to act as storage/assembly compartments for splicing factors, and recruitment of other SR proteins from speckles to sites of active transcription has been observed (Misteli et al., 1997; Spector and Lamond, 2011). Drawing parallels with this, it is possible that RS domain-mediated targeting of CFIm68 to speckles is crucial for its subsequent efficient recruitment to transcription sites and normal function in poly(A) site selection. Speckle localization may be impaired due to loss of RS domain-mediated protein-protein interactions, perhaps with some of the SR protein interaction partners previously identified (Dettwiler et al., 2004). Moreover, RS domain-mediated protein-protein interactions with these or other proteins may be critical for CFIm68 function more generally. Despite the in vitro demonstration that CFIm stabilizes recruitment of a CPSF-containing complex to RNA (Ruegsegger et al., 1996), the underlying mechanism has not been fully defined. CFIm was shown to interact with the hFip1 subunit of CPSF (Venkataraman et al., 2005) and the Clp1 subunit of CFIIm (de Vries et al., 2000), but it was not established which CFIm subunit or domain is involved. It is entirely possible that the CFIm68 RS domain is needed to mediate interactions with hFip1, Clp1 or other CPA factors, and that this is central to poly(A) site selection. Comparative mass spectrometry studies or more focused IP/western approaches with CFIm68 WT and ΔRS mutants would go some way towards exploring this possibility. Finally, binding of the RS domain to RNA may explain its essential role – this has already been demonstrated in vitro (Dettwiler et al., 2004), and RNA binding by RS domains of other SR proteins is known to play important functional roles (Shen and Green, 2004; Shen et al., 2004). Comparative CFIm ChIP/CLIP studies with the WT and ΔRS mutants would be informative in this respect, and it would be particularly interesting to establish whether the RS domain plays a key role in CFIm recruitment for UGUA-independent CFIm-regulated transcripts.

The discovery of a role for the CFIm68 RS domain in APA raised questions about the potential regulatory importance of phosphorylation within this domain, particularly as key functional roles for phosphorylation have been defined in a number of core SR proteins (Zhou and Fu, 2013). The novel identification of serine phosphorylation sites throughout the CFIm68 RS domain here, employing the new approach of pIMAGO western blot-based detection, represented a key step forward in defining the PTM landscape of CFIm. Furthermore, the demonstration that phosphomimetic serine mutagenesis of the eight serines in the CFIm68 RS1 domain led to impaired complementation in the APA assay suggests that this phosphorylation may have a key regulatory role. Following dissection of the contribution of individual phosphorylation events to this phenotype and validation at the transcriptome-wide level, a clear priority would be to investigate the mechanistic basis underpinning this observation. There are a number of examples of phosphorylation-dependent changes in protein-protein interactions of other SR proteins (Cho et al., 2011; Maertens et al., 2014), and this may provide a convincing explanation for the observed alterations in poly(A) site selection upon CFIm68 phosphorylation. For example, it is known that CFIm can interact with the hFip1 subunit of CPSF (Venkataraman et al., 2005), and may stabilise CPSF recruitment to stimulate poly(A) site use (Ruegsegger et al., 1996), but the CFIm subunit involved has not been identified – one possibility to explore would be that CFIm68 RS domain phosphorylation could impair such an interaction, or indeed interactions with other cleavage factors or RBPs, leading to reduced distal poly(A) site selection. Another key avenue to explore, based on the observation that hypophosphorylated SRSF1 binds more efficiently to ESEs than the hyperphosphorylated form (Cho et al., 2011), would be the possibility that CFIm68 RS phosphorylation impairs RNA binding of the complex around distal poly(A) sites.

Clearly, another important next step in characterising phosphorylation-mediated regulation of CFIm68 RS domain function will be identification of the kinases involved, which was not achieved in this study. SRPKs and CLKs represent the major candidates, and future work may need to employ different treatment conditions with small-molecule inhibitors, knockdown studies and perhaps analysis of panels of other kinases to characterize the signalling pathways involved. Understanding the

modification pathways may provide insights into the timing, cellular conditions and subcellular location of phosphorylation events, which will aid analysis of their potential roles.

6.4 Identification and functional analysis of CFIm PTMs

As well as the identified CFIm68 RS domain phosphorylation sites discussed above, MS analysis of CFIm identified two novel acetylation sites, at K173 on CFIm25 and K139 on CFIm68, and also confirmed a handful of other previously identified acetylation and phosphorylation sites. However, functional analysis of all newly identified PTMs, as well as a selection of known PTMs, surprisingly revealed that most of them had no obvious role in poly(A) site selection. There are of course limitations of the approaches used, which may have prevented identification of important PTMs. In particular, phosphorylation sites are notoriously hard to detect with conventional MS, though the pIMAGO analysis did suggest that there was no detectable phosphorylation on CFIm25, and that the majority on CFIm68 was accounted for by the characterised RS phosphorylation.

The simplest explanation for the observed results is that the majority of PTMs studied do not regulate CFIm function, and are functionally silent. There is a growing view that there may be a certain amount of 'noise' amongst PTMs such as phosphorylation, with certain phosphosites resulting from chance matches to kinase consensus sequences and not having undergone negative selection, for example (Lienhard, 2008). Indeed, algorithms have been generated in an attempt to predict the likely functionality of phosphosites, which interestingly predict that no CFIm sites have functional importance (Xiao et al., 2016). However, it is difficult to confirm that a given PTM is 'non-functional' – even though many CFIm PTMs showed no effect on APA regulation in the complementation assay, it may be that they have important transcript- or condition-specific effects, or roles in non-APA-related processes, which are not picked up by the northern blotting approach. Analysis of PTM site mutants using the transcriptome-wide 3' end sequencing approach would be one way to explore the possibility that specific CFIm PTMs regulate APA at specific groups of transcripts. Measuring the levels of PTMs and the effect of their mutagenesis under different cellular conditions and in different

cell types may also be informative. While a preliminary comparison of quiescent and proliferating cells did not reveal any clear differences in PTM profiles (see Appendix 1.4), this did not include the CFIm68 RS domain phosphorylation sites, and PTMs may make an important contribution to APA in other physiological conditions such as stress. Studying the role of CFIm PTMs in neuronal cells would also be of particular interest, due to their particularly long 3'UTRs and the demonstrated importance of CFIm-regulated APA in a glioblastoma model (Masamha et al., 2014; Miura et al., 2013). A further point to note is the possibility that PTMs co-ordinately regulate CFIm function, as such multi-site 'PTM codes' are known to be prevalent (Lothrop et al., 2013). This was partially explored, for example through the analysis of a multiple lysine-to-arginine acetylation site mutant in CFIm25, which showed no changes in function, but it cannot be ruled out that other PTM site mutations in combination may lead to altered PTM site use.

Aside from possible PTM-mediated regulation, it is worth considering other ways in which CFIm activity may be controlled in different physiological settings. The demonstration that CFIm25 can bind (though not hydrolyse) ATP and Ap_4A, and that binding of Ap_4A and RNA by CFIm25 are mutually exclusive (Coseno et al., 2008), raises the intriguing possibility that CFIm activity could be regulated by these nucleotides. Ap_4A is considered a 'stress-induced' signalling molecule, with a role in promoting apoptosis and possibly regulating gene expression, though its functions have not been fully defined (Marriott et al., 2016). It would therefore be particularly interesting to explore potential links between Ap_4A and CFIm-regulated APA in the context of different types of cell stress. It is also possible that CFIm activity is regulated indirectly, for example through interactions with a protein binding partner that is cell cycle-regulated. Such regulation may be harder to dissect, but a more comprehensive study of CFIm binding partners would be a useful starting point.

6.5 Concluding remarks

This thesis provides a number of insights into the function and regulation of CFIm, but arguably the finding of most significance is the novel demonstration that the RS domain of CFIm68 makes an

important contribution to the regulation of APA, potentially in a phosphorylation-dependent manner. As discussed above, further investigation of the mechanistic involvement of this domain in poly(A) site selection would be a clear priority for future study. Despite attempts to address them here, some major questions surrounding the mechanism of action of CFIm and its importance as a physiological APA regulator remain. While APA is clearly a complex phenomenon, likely subject to multi-factorial regulation, further dissection of the function and regulation of CFIm will no doubt be a crucial step towards improving our understanding of this process.

References

Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**, 720-731.

Ahn, S.H., Kim, M., and Buratowski, S. (2004). Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3' end processing. *Mol Cell* **13**, 67-76.

Ainger, K., Avossa, D., Diana, A.S., Barry, C., Barbarese, E., and Carson, J.H. (1997). Transport and localization elements in myelin basic protein mRNA. *J Cell Biol* **138**, 1077-1087.

Alt, F.W., Bothwell, A.L.M., Knapp, M., Siden, E., Mather, E., Koshland, M., and Baltimore, D. (1980). Synthesis of Secreted and Membrane-Bound Immunoglobulin-Mu Heavy-Chains Is Directed by Messenger-Rnas That Differ at Their 3' Ends. *Cell* **20**, 293-301.

Amrein, H., Gorman, M., and Nothiger, R. (1988). The sex-determining gene tra-2 of Drosophila encodes a putative RNA binding protein. *Cell* **55**, 1025-1035.

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

Araki, S., Dairiki, R., Nakayama, Y., Murai, A., Miyashita, R., Iwatani, M., Nomura, T., and Nakanishi, O. (2015). Inhibitors of CLK protein kinases suppress cell growth and induce apoptosis by modulating pre-mRNA splicing. *PLoS One* **10**, e0116929.

Bai, Y., Auperin, T.C., Chou, C.Y., Chang, G.G., Manley, J.L., and Tong, L. (2007). Crystal structure of murine CstF-77: dimeric association and implications for polyadenylation of mRNA precursors. *Mol Cell* **25**, 863-875.

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

Batra, R., Charizanis, K., Manchanda, M., Mohan, A., Li, M., Finn, D.J., Goodwin, M., Zhang, C., Sobczak, K., Thornton, C.A.*, et al.* (2014). Loss of MBNL leads to disruption of developmentally regulated alternative polyadenylation in RNA-mediated disease. *Mol Cell* **56**, 311-322.

Bava, F.A., Eliscovich, C., Ferreira, P.G., Minana, B., Ben-Dov, C., Guigo, R., Valcarcel, J., and Mendez, R. (2013). CPEB1 coordinates alternative 3 ' -UTR formation with translational regulation. *Nature* **495**, 121-125.

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

Beck, A.H., Weng, Z.M., Witten, D.M., Zhu, S., Foley, J.W., Lacroute, P., Smith, C.L., Tibshirani, R., van de Rijn, M., Sidow, A.*, et al.* (2010). 3 '-End Sequencing for Expression Quantification (3SEQ) from Archival Tumor Samples. *PLoS One* **5**, e8768.

Bensaude, O. (2011). Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? *Transcription* **2**, 103-108.

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

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

Bethune, J., Artus-Revel, C.G., and Filipowicz, W. (2012). Kinetic analysis reveals successive steps leading to miRNA-mediated silencing in mammalian cells. *EMBO Rep* **13**, 716-723.

Bienroth, S., Keller, W., and Wahle, E. (1993). Assembly of a processive messenger RNA polyadenylation complex. *EMBO J* **12**, 585-594.

Brown, K.M., and Gilmartin, G.M. (2003). A mechanism for the regulation of pre-mRNA 3' processing by human cleavage factor Im. *Mol Cell* **12**, 1467-1476.

Caceres, J.F., Misteli, T., Screaton, G.R., Spector, D.L., and Krainer, A.R. (1997). Role of the modular domains of SR proteins in subnuclear localization and alternative splicing specificity. *J Cell Biol* **138**, 225-238.

Caceres, J.F., Screaton, G.R., and Krainer, A.R. (1998). A specific subset of SR proteins shuttles continuously between the nucleus and the cytoplasm. *Genes Dev* **12**, 55-66.

Callebaut, I., Moshous, D., Mornon, J.P., and de Villartay, J.P. (2002). Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family. *Nucleic Acids Res* **30**, 3592-3601.

Caput, D., Beutler, B., Hartog, K., Thayer, R., Brown-Shimer, S., and Cerami, A. (1986). Identification of a common nucleotide sequence in the 3'-untranslated region of mRNA molecules specifying inflammatory mediators. *Proc Natl Acad Sci U S A* **83**, 1670-1674.

Cardinale, S., Cisterna, B., Bonetti, P., Aringhieri, C., Biggiogera, M., and Barabino, S.M.L. (2007). Subnuclear localization and dynamics of the pre-mRNA 3 ' end processing factor mammalian cleavage factor I 68-kDa subunit. *Mol Biol Cell* **18**, 1282-1292.

Cazalla, D., Zhu, J., Manche, L., Huber, E., Krainer, A.R., and Caceres, J.F. (2002). Nuclear export and retention signals in the RS domain of SR proteins. *Mol Cell Biol* **22**, 6871-6882.

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

Chang, H., Lim, J., Ha, M., and Kim, V.N. (2014). TAIL-seq: Genome-wide Determination of Poly(A) Tail Length and 3' End Modifications. *Mol Cell* **53**, 1044-1052.

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

Chen, C.Y., Xu, N., and Shyu, A.B. (2002). Highly selective actions of HuR in antagonizing AU-rich element-mediated mRNA destabilization. *Mol Cell Biol* **22**, 7268-7278.

Cho, S., Hoang, A., Sinha, R., Zhong, X.Y., Fu, X.D., Krainer, A.R., and Ghosh, G. (2011). Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. *Proc Natl Acad Sci U S A* **108**, 8233-8238.

Christofori, G., and Keller, W. (1989). Poly(A) polymerase purified from HeLa cell nuclear extract is required for both cleavage and polyadenylation of pre-mRNA in vitro. *Mol Cell Biol* **9**, 193-203.

Cmarko, D., Verschure, P.J., Martin, T.E., Dahmus, M.E., Krause, S., Fu, X.D., van Driel, R., and Fakan, S. (1999). Ultrastructural analysis of transcription and splicing in the cell nucleus after bromo-UTP microinjection. *Mol Biol Cell* **10**, 211-223.

Colgan, D.F., and Manley, J.L. (1997). Mechanism and regulation of mRNA polyadenylation. *Genes Dev* **11**, 2755-2766.

Colwill, K., Feng, L.L., Yeakley, J.M., Gish, G.D., Caceres, J.F., Pawson, T., and Fu, X.D. (1996a). SRPK1 and Clk/Sty protein kinases show distinct substrate specificities for serine/arginine-rich splicing factors. *J Biol Chem* **271**, 24569-24575.

Colwill, K., Pawson, T., Andrews, B., Prasad, J., Manley, J.L., Bell, J.C., and Duncan, P.I. (1996b). The Clk/Sty protein kinase phosphorylates SR splicing factors and regulates their intranuclear distribution. *EMBO J* **15**, 265-275.

Connelly, S., and Manley, J.L. (1988). A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev* **2**, 440-452.

Coseno, M., Martin, G., Berger, C., Gilmartin, G., Keller, W., and Doublie, S. (2008). Crystal structure of the 25 kDa subunit of human cleavage factor Im. *Nucleic Acids Res* **36**, 3474-3483.

Crenshaw, E.B., 3rd, Russo, A.F., Swanson, L.W., and Rosenfeld, M.G. (1987). Neuron-specific alternative RNA processing in transgenic mice expressing a metallothionein-calcitonin fusion gene. *Cell* **49**, 389-398.

Csizmok, V., Follis, A.V., Kriwacki, R.W., and Forman-Kay, J.D. (2016). Dynamic Protein Interaction Networks and New Structural Paradigms in Signaling. *Chem Rev* **116**, 6424-6462.

Cui, Y., and Denis, C.L. (2003). In vivo evidence that defects in the transcriptional elongation factors RPB2, TFIIS, and SPT5 enhance upstream poly(A) site utilization. *Mol Cell Biol* **23**, 7887-7901.

De Iaco, A., Santoni, F., Vannier, A., Guipponi, M., Antonarakis, S., and Luban, J. (2013). TNPO3 protects HIV-1 replication from CPSF6-mediated capsid stabilization in the host cell cytoplasm. *Retrovirology* **10**, 20.

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

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

DeMaria, C.T., and Brewer, G. (1996). AUF1 binding affinity to A+U-rich elements correlates with rapid mRNA degradation. *J Biol Chem* **271**, 12179-12184.

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

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

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

Duncan, P.I., Stojdl, D.F., Marius, R.M., Scheit, K.H., and Bell, J.C. (1998). The Clk2 and Clk3 dual-specificity protein kinases regulate the intranuclear distribution of SR proteins and influence pre-mRNA splicing. *Exp Cell Res* **241**, 300-308.

Dye, M.J., and Proudfoot, N.J. (1999). Terminal exon definition occurs cotranscriptionally and promotes termination of RNA polymerase II. *Mol Cell* **3**, 371-378.

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R., and Hood, L. (1980). 2 Messenger-Rnas Can Be Produced from a Single Immunoglobulin-Mu Gene by Alternative Rna Processing Pathways. *Cell* **20**, 313-319.

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

Edwards, R.A., Lee, M.S., Tsutakawa, S.E., Williams, R.S., Nazeer, I., Kleiman, F.E., Tainer, J.A., and Glover, J.N. (2008). The BARD1 C-terminal domain structure and interactions with polyadenylation factor CstF-50. *Biochemistry* **47**, 11446-11456.

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

Erkelenz, S., Mueller, W.F., Evans, M.S., Busch, A., Schoneweis, K., Hertel, K.J., and Schaal, H. (2013). Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA* **19**, 96-102.

Flavell, S.W., Kim, T.K., Gray, J.M., Harmin, D.A., Hemberg, M., Hong, E.J., Markenscoff-Papadimitriou, E., Bear, D.M., and Greenberg, M.E. (2008). Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. *Neuron* **60**, 1022-1038.

Fong, N., and Bentley, D.L. (2001). Capping, splicing, and 3' processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD. *Genes Dev* **15**, 1783-1795.

Fong, N., Brannan, K., Erickson, B., Kim, H., Cortazar, M.A., Sheridan, R.M., Nguyen, T., Karp, S., and Bentley, D.L. (2015). Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Mol Cell* **60**, 256-267.

Friedlander, M.R., Lizano, E., Houben, A.J., Bezdan, D., Banez-Coronel, M., Kudla, G., Mateu-Huertas, E., Kagerbauer, B., Gonzalez, J., Chen, K.C*., et al.* (2014). Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol* **15**, R57.

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

Fu, X.D., and Maniatis, T. (1990). Factor required for mammalian spliceosome assembly is localized to discrete regions in the nucleus. *Nature* **343**, 437-441.

Fu, X.D., and Maniatis, T. (1992). The 35-kDa mammalian splicing factor SC35 mediates specific interactions between U1 and U2 small nuclear ribonucleoprotein particles at the 3' splice site. *Proc Natl Acad Sci U S A* **89**, 1725-1729.

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

Fuke, H., and Ohno, M. (2008). Role of poly (A) tail as an identity element for mRNA nuclear export. *Nucleic Acids Res* **36**, 1037-1049.

Fukuhara, T., Hosoya, T., Shimizu, S., Sumi, K., Oshiro, T., Yoshinaka, Y., Suzuki, M., Yamamoto, N., Herzenberg, L.A., Herzenberg, L.A.*, et al.* (2006). Utilization of host SR protein kinases and RNA-splicing machinery during viral replication. *Proc Natl Acad Sci U S A* **103**, 11329-11333.

Fukumitsu, H., Soumiya, H., and Furukawa, S. (2012). Knockdown of pre-mRNA cleavage factor Im 25 kDa promotes neurite outgrowth. *Biochem Biophys Res Commun* **425**, 848-853.

Garneau, N.L., Wilusz, J., and Wilusz, C.J. (2007). The highways and byways of mRNA decay. *Nat Rev Mol Cell Biol* **8**, 113-126.

Gennarino, V.A., Alcott, C.E., Chen, C.A., Chaudhury, A., Gillentine, M.A., Rosenfeld, J.A., Parikh, S., Wheless, J.W., Roeder, E.R., Horovitz, D.D.*, et al.* (2015). NUDT21-spanning CNVs lead to neuropsychiatric disease and altered MeCP2 abundance via alternative polyadenylation. *Elife* **4**, e10782.

Ghosh, G., and Adams, J.A. (2011). Phosphorylation mechanism and structure of serine-arginine protein kinases. *FEBS J* **278**, 587-597.

Giannakouros, T., Nikolakaki, E., Mylonis, I., and Georgatsou, E. (2011). Serine-arginine protein kinases: a small protein kinase family with a large cellular presence. *FEBS J* **278**, 570-586.

Gil, A., and Proudfoot, N.J. (1987). Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation. *Cell* **49**, 399-406.

Glover-Cutter, K., Kim, S., Espinosa, J., and Bentley, D.L. (2008). RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nat Struct Mol Biol* **15**, 71-78.

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

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

Gruber, A.R., Martin, G., Muller, P., Schmidt, A., Gruber, A.J., Gumienny, R., Mittal, N., Jayachandran, R., Pieters, J., Keller, W.*, et al.* (2014). Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nat Commun* **5**, 5465.

Gui, J.F., Lane, W.S., and Fu, X.D. (1994a). A serine kinase regulates intracellular localization of splicing factors in the cell cycle. *Nature* **369**, 678-682.

Gui, J.F., Tronchere, H., Chandler, S.D., and Fu, X.D. (1994b). Purification and characterization of a kinase specific for the serine- and arginine-rich pre-mRNA splicing factors. *Proc Natl Acad Sci U S A* **91**, 10824-10828.

Guo, A., Gu, H., Zhou, J., Mulhern, D., Wang, Y., Lee, K.A., Yang, V., Aguiar, M., Kornhauser, J., Jia, X.*, et al.* (2014). Immunoaffinity enrichment and mass spectrometry analysis of protein methylation. *Mol Cell Proteomics* **13**, 372-387.

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835-840.

Hanada, T., Weitzer, S., Mair, B., Bernreuther, C., Wainger, B.J., Ichida, J., Hanada, R., Orthofer, M., Cronin, S.J., Komnenovic, V.*, et al.* (2013). CLP1 links tRNA metabolism to progressive motor-neuron loss. *Nature* **495**, 474-480.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S.*, et al.* (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515-1526.

Haynes, C., and Iakoucheva, L.M. (2006). Serine/arginine-rich splicing factors belong to a class of intrinsically disordered proteins. *Nucleic Acids Res* **34**, 305-312.

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

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

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

Holcik, M., and Sonenberg, N. (2005). Translational control in stress and apoptosis. *Nat Rev Mol Cell Biol* **6**, 318-327.

Hon, L.S., and Zhang, Z. (2007). The roles of binding site arrangement and combinatorial targeting in microRNA repression of gene expression. *Genome Biol* **8**, R166.

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

Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* **43**, D512-520.

Huang, Y., Gattoni, R., Stevenin, J., and Steitz, J.A. (2003). SR splicing factors serve as adapter proteins for TAP-dependent mRNA export. *Mol Cell* **11**, 837-843.

Humphreys, D.T., Westman, B.J., Martin, D.I., and Preiss, T. (2005). MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function. *Proc Natl Acad Sci U S A* **102**, 16961-16966.

Iliuk, A., Liu, X.S., Xue, L., Liu, X., and Tao, W.A. (2012). Chemical visualization of phosphoproteomes on membrane. *Mol Cell Proteomics* **11**, 629-639.

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

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

Ji, Z., Lee, J.Y., Pan, Z., Jiang, B., and Tian, B. (2009). Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc Natl Acad Sci U S A* **106**, 7028-7033.

Ji, Z., Luo, W., Li, W., Hoque, M., Pan, Z., Zhao, Y., and Tian, B. (2011). Transcriptional activity regulates alternative cleavage and polyadenylation. *Mol Syst Biol* **7**, 534.

Ji, Z., and Tian, B. (2009). Reprogramming of 3 ' Untranslated Regions of mRNAs by Alternative Polyadenylation in Generation of Pluripotent Stem Cells from Different Cell Types. *PLoS One* **4**, e8419.

Jiang, X.R., Jimenez, G., Chang, E., Frolkis, M., Kusler, B., Sage, M., Beeche, M., Bodnar, A.G., Wahl, G.M., Tlsty, T.D.*, et al.* (1999). Telomerase expression in human somatic cells does not induce changes associated with a transformed phenotype. *Nat Genet* **21**, 111-114.

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

Katahira, J., Okuzaki, D., Inoue, H., Yoneda, Y., Maehara, K., and Ohkawa, Y. (2013). Human TREX component Thoc5 affects alternative polyadenylation site choice by recruiting mammalian cleavage factor I. *Nucleic Acids Res* **41**, 7060-7072.

Kataoka, N., Bachorik, J.L., and Dreyfuss, G. (1999). Transportin-SR, a nuclear import receptor for SR proteins. *J Cell Biol* **145**, 1145-1152.

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

Keller, W., Bienroth, S., Lang, K.M., and Christofori, G. (1991). Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal AAUAAA. *EMBO J* **10**, 4241-4249.

Kim, H., and Lee, Y. (2001). Interaction of poly(A) polymerase with the 25-kDa subunit of cleavage factor I. *Biochem Biophys Res Commun* **289**, 513-518.

Kim, M., Krogan, N.J., Vasiljeva, L., Rando, O.J., Nedea, E., Greenblatt, J.F., and Buratowski, S. (2004). The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature* **432**, 517-522.

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

Koizumi, J., Okamoto, Y., Onogi, H., Mayeda, A., Krainer, A.R., and Hagiwara, M. (1999). The subcellular localization of SF2/ASF is regulated by direct interaction with SR protein kinases (SRPKs). *J Biol Chem* **274**, 11125-11131.

Kolev, N.G., Yario, T.A., Benson, E., and Steitz, J.A. (2008). Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3'-end maturation. *EMBO Rep* **9**, 1013-1018.

Krainer, A.R., Conway, G.C., and Kozak, D. (1990). Purification and characterization of pre-mRNA splicing factor SF2 from HeLa cells. *Genes Dev* **4**, 1158-1171.

Kubo, T., Wada, T., Yamaguchi, Y., Shimizu, A., and 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.

Kuhn, U., Gundel, M., Knoth, A., Kerwitz, Y., Rudel, S., and Wahle, E. (2009). Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. *J Biol Chem* **284**, 22803-22814.

Kyburz, A., Friedlein, A., Langen, H., and Keller, W. (2006). Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3' end processing and splicing. *Mol Cell* **23**, 195-205.

Kyriakopoulou, C.B., Nordvarg, H., and Virtanen, A. (2001). A novel nuclear human poly(A) polymerase (PAP), PAP gamma. *J Biol Chem* **276**, 33504-33511.

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

Lai, M.C., Lin, R.I., Huang, S.Y., Tsai, C.W., and Tarn, W.Y. (2000). A human importin-beta family protein, transportin-SR2, interacts with the phosphorylated RS domain of SR proteins. *J Biol Chem* **275**, 7950-7957.

Lai, M.C., and Tarn, W.Y. (2004). Hypophosphorylated ASF/SF2 binds TAP and is present in messenger ribonucleoproteins. *J Biol Chem* **279**, 31745-31749.

Lai, W.S., Parker, J.S., Grissom, S.F., Stumpo, D.J., and Blackshear, P.J. (2006). Novel mRNA targets for tristetraprolin (TTP) identified by global analysis of stabilized transcripts in TTP-deficient fibroblasts. *Mol Cell Biol* **26**, 9196-9208.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359.

Lawrence, J.B., and Singer, R.H. (1986). Intracellular localization of messenger RNAs for cytoskeletal proteins. *Cell* **45**, 407-415.

Lee, K., Ambrose, Z., Martin, T.D., Oztop, I., Mulky, A., Julias, J.G., Vandegraaff, N., Baumann, J.G., Wang, R., Yuen, W.*, et al.* (2010). Flexible use of nuclear import pathways by HIV-1. *Cell Host Microbe* **7**, 221-233.

Legrand, P., Pinaud, N., Minvielle-Sebastia, L., and Fribourg, S. (2007). The structure of the CstF-77 homodimer provides insights into CstF assembly. *Nucleic Acids Res* **35**, 4515-4522.

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. (2003). Prediction of mammalian microRNA targets. *Cell* **115**, 787-798.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

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

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

Li, W., You, B., Hoque, M., Zheng, D., Luo, W., Ji, Z., Park, J.Y., Gunderson, S.I., Kalsotra, A., Manley, J.L.*, et al.* (2015). Systematic profiling of poly(a)+ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. *PLoS Genet* **11**, e1005166.

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

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.*, et al.* (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464-469.

Lienhard, G.E. (2008). Non-functional phosphorylations? *Trends Biochem Sci* **33**, 351-352.

Lin, Y.C., Boone, M., Meuris, L., Lemmens, I., Van Roy, N., Soete, A., Reumers, J., Moisse, M., Plaisance, S., Drmanac, R.*, et al.* (2014). Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat Commun* **5**, 4767.

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

Long, J.C., and Caceres, J.F. (2009). The SR protein family of splicing factors: master regulators of gene expression. *Biochem J* **417**, 15-27.

Lothrop, A.P., Torres, M.P., and Fuchs, S.M. (2013). Deciphering post-translational modification codes. *FEBS Lett* **587**, 1247-1257.

Lou, H., Gagel, R.F., and Berget, S.M. (1996). An intron enhancer recognized by splicing factors activates polyadenylation. *Genes Dev* **10**, 208-219.

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

Lovci, M.T., Bengtson, M.H., and Massirer, K.B. (2016). Post-Translational Modifications and RNA-Binding Proteins. *Adv Exp Med Biol* **907**, 297-317.

Ludwig, N., Leidinger, P., Becker, K., Backes, C., Fehlmann, T., Pallasch, C., Rheinheimer, S., Meder, B., Stahler, C., Meese, E.*, et al.* (2016). Distribution of miRNA expression across human tissues. *Nucleic Acids Res* **44**, 3865-3877.

Lukacs, G.L., and Verkman, A.S. (2012). CFTR: folding, misfolding and correcting the DeltaF508 conformational defect. *Trends Mol Med* **18**, 81-91.

MacDonald, C.C., Wilusz, J., and Shenk, T. (1994). The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. *Mol Cell Biol* **14**, 6647-6654.

Maertens, G.N., Cook, N.J., Wang, W., Hare, S., Gupta, S.S., Oztop, I., Lee, K., Pye, V.E., Cosnefroy, O., Snijders, A.P.*, et al.* (2014). Structural basis for nuclear import of splicing factors by human Transportin 3. *Proc Natl Acad Sci U S A* **111**, 2728-2733.

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

Manley, J.L., and Krainer, A.R. (2010). A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). *Genes Dev* **24**, 1073-1074.

Marriott, A.S., Vasieva, O., Fang, Y., Copeland, N.A., McLennan, A.G., and Jones, N.J. (2016). NUDT2 Disruption Elevates Diadenosine Tetraphosphate (Ap4A) and Down-Regulates Immune Response and Cancer Promotion Genes. *PLoS One* **11**, e0154674.

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

Martin, G., Moglich, A., Keller, W., and Doublie, S. (2004). Biochemical and structural insights into substrate binding and catalytic mechanism of mammalian poly(A) polymerase. *J Mol Biol* **341**, 911-925.

Martin, G., Ostareck-Lederer, A., Chari, A., Neuenkirchen, N., Dettwiler, S., Blank, D., Ruegsegger, U., Fischer, U., and Keller, W. (2010). Arginine methylation in subunits of mammalian pre-mRNA cleavage factor I. *RNA* **16**, 1646-1659.

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

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17**, 10-12.

Martincic, K., Alkan, S.A., Cheatle, A., Borghesi, L., and Milcarek, C. (2009). Transcription elongation factor ELL2 directs immunoglobulin secretion in plasma cells by stimulating altered RNA processing. *Nat Immunol* **10**, 1102-1109.

Martincic, K., Campbell, R., Edwalds-Gilbert, G., Souan, L., Lotze, M.T., and Milcarek, C. (1998). Increase in the 64-kDa subunit of the polyadenylation/cleavage stimulatory factor during the G0 to S phase transition. *Proc Natl Acad Sci U S A* **95**, 11095-11100.

Masamha, C.P., Xia, Z., Peart, N., Collum, S., Li, W., Wagner, E.J., and Shyu, A.B. (2016). CFIm25 regulates glutaminase alternative terminal exon definition to modulate miR-23 function. *RNA* **22**, 830-838.

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

Mata, J. (2013). Genome-wide mapping of polyadenylation sites in fission yeast reveals widespread alternative polyadenylation. *RNA Biol* **10**, 1407-1414.

Mathew, R., Hartmuth, K., Mohlmann, S., Urlaub, H., Ficner, R., and Luhrmann, R. (2008). Phosphorylation of human PRP28 by SRPK2 is required for integration of the U4/U6-U5 tri-snRNP into the spliceosome. *Nat Struct Mol Biol* **15**, 435-443.

Matoulkova, E., Michalova, E., Vojtesek, B., and Hrstka, R. (2012). The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA Biol* **9**, 563-576.

Matsuoka, S., Ballif, B.A., Smogorzewska, A., McDonald, E.R., 3rd, Hurov, K.E., Luo, J., Bakalarski, C.E., Zhao, Z., Solimini, N., Lerenthal, Y.*, et al.* (2007). ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science* **316**, 1160-1166.

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

McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S.D., Wickens, M., and Bentley, D.L. (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* **385**, 357-361.

McLennan, A.G. (2006). The Nudix hydrolase superfamily. *Cell Mol Life Sci* **63**, 123-143.

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

Melcak, I., Cermanova, S., Jirsova, K., Koberna, K., Malinsky, J., and Raska, I. (2000). Nuclear pre-mRNA compartmentalization: trafficking of released transcripts to splicing factor reservoirs. *Mol Biol Cell* **11**, 497-510.

Mermoud, J.E., Cohen, P., and Lamond, A.I. (1992). Ser/Thr-specific protein phosphatases are required for both catalytic steps of pre-mRNA splicing. *Nucleic Acids Res* **20**, 5263-5269.

Mermoud, J.E., Cohen, P.T., and Lamond, A.I. (1994). Regulation of mammalian spliceosome assembly by a protein phosphorylation mechanism. *EMBO J* **13**, 5679-5688.

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

Millevoi, S., Loulergue, C., Dettwiler, S., Karaa, S.Z., Keller, W., Antoniou, M., and 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.

Misteli, T., Caceres, J.F., Clement, J.Q., Krainer, A.R., Wilkinson, M.F., and Spector, D.L. (1998). Serine phosphorylation of SR proteins is required for their recruitment to sites of transcription in vivo. *J Cell Biol* **143**, 297-307.

Misteli, T., Caceres, J.F., and Spector, D.L. (1997). The dynamics of a pre-mRNA splicing factor in living cells. *Nature* **387**, 523-527.

Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J.O., and Lai, E.C. (2013). Widespread and extensive lengthening of 3 ' UTRs in the mammalian brain. *Genome Res* **23**, 812-825.

Moll, P., Ante, M., Seitz, A., and Reda, T. (2014). QuantSeq 3′ mRNA sequencing for RNA quantification. *Nat Methods* **11**.

Moreno-Morcillo, M., Minvielle-Sebastia, L., Mackereth, C., and Fribourg, S. (2011). Hexameric architecture of CstF supported by CstF-50 homodimerization domain structure. *RNA* **17**, 412-418.

Mueller, A.A., Cheung, T.H., and Rando, T.A. (2013). All's well that ends well: alternative polyadenylation and its implications for stem cell biology. *Curr Opin Cell Biol* **25**, 222-232.

Muraki, M., Ohkawara, B., Hosoya, T., Onogi, H., Koizumi, J., Koizumi, T., Sumi, K., Yomoda, J., Murray, M.V., Kimura, H.*, et al.* (2004). Manipulation of alternative splicing by a newly developed inhibitor of Clks. *J Biol Chem* **279**, 24246-24254.

Murthy, K.G., and 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.

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

Nakagawa, O., Arnold, M., Nakagawa, M., Hamada, H., Shelton, J.M., Kusano, H., Harris, T.M., Childs, G., Campbell, K.P., Richardson, J.A.*, et al.* (2005). Centronuclear myopathy in mice lacking a novel muscle-specific protein kinase transcriptionally regulated by MEF2. *Genes Dev* **19**, 2066-2077.

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

Neugebauer, K.M., and Roth, M.B. (1997). Distribution of pre-mRNA splicing factors at sites of RNA polymerase II transcription. *Genes Dev* **11**, 1148-1159.

Neugebauer, K.M., Stolk, J.A., and Roth, M.B. (1995). A conserved epitope on a subset of SR proteins defines a larger family of Pre-mRNA splicing factors. *J Cell Biol* **129**, 899-908.

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

Ngo, J.C., Chakrabarti, S., Ding, J.H., Velazquez-Dones, A., Nolen, B., Aubol, B.E., Adams, J.A., Fu, X.D., and Ghosh, G. (2005). Interplay between SRPK and Clk/Sty kinases in phosphorylation of the splicing factor ASF/SF2 is regulated by a docking motif in ASF/SF2. *Mol Cell* **20**, 77-89.

Ngo, J.C., Giang, K., Chakrabarti, S., Ma, C.T., Huynh, N., Hagopian, J.C., Dorrestein, P.C., Fu, X.D., Adams, J.A., and Ghosh, G. (2008). A sliding docking interaction is essential for sequential and processive phosphorylation of an SR protein by SRPK1. *Mol Cell* **29**, 563-576.

Ni, Z., Schwartz, B.E., Werner, J., Suarez, J.R., and Lis, J.T. (2004). Coordination of transcription, RNA processing, and surveillance by P-TEFb kinase on heat shock genes. *Mol Cell* **13**, 55-65.

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

Nojima, T., Gomes, T., Grosso, A.R., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161**, 526-540.

Nunes, N.M., Li, W., Tian, B., and 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.

Oktaba, K., Zhang, W., Lotz, T.S., Jun, D.J., Lemke, S.B., Ng, S.P., Esposito, E., Levine, M., and Hilgers, V. (2015). ELAV Links Paused Pol II to Alternative Polyadenylation in the Drosophila Nervous System. *Mol Cell* **57**, 341-348.

Olsen, J.V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M.L., Jensen, L.J., Gnad, F., Cox, J., Jensen, T.S., Nigg, E.A.*, et al.* (2010). Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci Signal* **3**, ra3.

Pinto, P.A., 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.*, et al.* (2011). RNA polymerase II kinetics in polo polyadenylation signal selection. *EMBO J* **30**, 2431-2444.

Proudfoot, N.J. (1989). How RNA polymerase II terminates transcription in higher eukaryotes. *Trends Biochem Sci* **14**, 105-110.

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

Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281-2308.

Rasheedi, S., Shun, M.C., Serrao, E., Sowd, G.A., Qian, J., Hao, C., Dasgupta, T., Engelman, A.N., and Skowronski, J. (2016). The Cleavage and Polyadenylation Specificity Factor 6 (CPSF6) Subunit of the Capsid-recruited Pre-messenger RNA Cleavage Factor I (CFIm) Complex Mediates HIV-1 Integration into Genes. *J Biol Chem* **291**, 11809-11819.

Riaz, M., Raz, Y., van Putten, M., Paniagua-Soriano, G., Krom, Y.D., Florea, B.I., and Raz, V. (2016). PABPN1-Dependent mRNA Processing Induces Muscle Wasting. *PLoS Genet* **12**, e1006031.

Rosenwald, A., Wright, G., Wiestner, A., Chan, W.C., Connors, J.M., Campo, E., Gascoyne, R.D., Grogan, T.M., Muller-Hermelink, H.K., Smeland, E.B.*, et al.* (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell* **3**, 185-197.

Rosonina, E., Bakowski, M.A., McCracken, S., and Blencowe, B.J. (2003). Transcriptional activators control splicing and 3'-end cleavage levels. *J Biol Chem* **278**, 43034-43040.

Ruegsegger, U., Beyer, K., and Keller, W. (1996). Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors. *J Biol Chem* **271**, 6107-6113.

Ruegsegger, U., Blank, D., and Keller, W. (1998). Human pre-mRNA cleavage factor Im is related to spliceosomal SR proteins and can be reconstituted in vitro from recombinant subunits. *Mol Cell* **1**, 243-253.

Ruepp, M.D., Aringhieri, C., Vivarelli, S., Cardinale, S., Paro, S., Schumperli, D., and Barabino, S.M. (2009). Mammalian pre-mRNA 3' end processing factor CF I m 68 functions in mRNA export. *Mol Biol Cell* **20**, 5211-5223.

Ruepp, M.D., Vivarelli, S., Pillai, R.S., Kleinschmidt, N., Azzouz, T.N., Barabino, S.M., and Schumperli, D. (2010). The 68 kDa subunit of mammalian cleavage factor I interacts with the

U7 small nuclear ribonucleoprotein and participates in 3'-end processing of animal histone mRNAs. *Nucleic Acids Res* **38**, 7637-7650.

Rush, J., Moritz, A., Lee, K.A., Guo, A., Goss, V.L., Spek, E.J., Zhang, H., Zha, X.M., Polakiewicz, R.D., and Comb, M.J. (2005). Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat Biotechnol* **23**, 94-101.

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

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

Sasado, T., Kondoh, H., Furutani-Seiki, M., and Naruse, K. (2017). Mutation in cpsf6/CFIm68 (Cleavage and Polyadenylation Specificity Factor Subunit 6) causes short 3'UTRs and disturbs gene expression in developing embryos, as revealed by an analysis of primordial germ cell migration using the medaka mutant naruto. *PLoS One* **12**, e0172467.

Schonemann, L., Kuhn, U., Martin, G., Schafer, P., Gruber, A.R., Keller, W., Zavolan, M., and Wahle, E. (2014). Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. *Genes Dev* **28**, 2381-2393.

Setzer, D.R., McGrogan, M., Nunberg, J.H., and Schimke, R.T. (1980). Size heterogeneity in the 3' end of dihydrofolate reductase messenger RNAs in mouse cells. *Cell* **22**, 361-370.

Shaw, G., and Kamen, R. (1986). A conserved AU sequence from the 3' untranslated region of GM-CSF mRNA mediates selective mRNA degradation. *Cell* **46**, 659-667.

Sheets, M.D., Ogg, S.C., and 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**, 5799-5805.

Shell, S.A., Hesse, C., Morris, S.M., Jr., and 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**, 39950-39961.

Shen, H., and Green, M.R. (2004). A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly. *Mol Cell* **16**, 363-373.

Shen, H., Kan, J.L., and Green, M.R. (2004). Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. *Mol Cell* **13**, 367-376.

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

Shimazu, T., Horinouchi, S., and Yoshida, M. (2007). Multiple histone deacetylases and the CREB-binding protein regulate pre-mRNA 3'-end processing. *J Biol Chem* **282**, 4470-4478.

Shimberg, G.D., Michalek, J.L., Oluyadi, A.A., Rodrigues, A.V., Zucconi, B.E., Neu, H.M., Ghosh, S., Sureschandra, K., Wilson, G.M., Stemmler, T.L.*, et al.* (2016). Cleavage and polyadenylation specificity factor 30: An RNA-binding zinc-finger protein with an unexpected 2Fe-2S cluster. *Proc Natl Acad Sci U S A* **113**, 4700-4705.

Skourti-Stathaki, K., Kamieniarz-Gdula, K., and Proudfoot, N.J. (2014). R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature* **516**, 436-439.

Skourti-Stathaki, K., Proudfoot, N.J., and Gromak, N. (2011). Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol Cell* **42**, 794-805.

Sommer, J., Garbers, C., Wolf, J., Trad, A., Moll, J.M., Sack, M., Fischer, R., Grotzinger, J., Waetzig, G.H., Floss, D.M*., et al.* (2014). Alternative intronic polyadenylation generates the interleukin-6 trans-signaling inhibitor sgp130-E10. *J Biol Chem* **289**, 22140-22150.

Sonenberg, N., and Hinnebusch, A.G. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136**, 731-745.

Sowd, G.A., Serrao, E., Wang, H., Wang, W., Fadel, H.J., Poeschla, E.M., and Engelman, A.N. (2016). A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proc Natl Acad Sci U S A* **113**, E1054-1063.

Spector, D.L., Fu, X.D., and Maniatis, T. (1991). Associations between distinct pre-mRNA splicing components and the cell nucleus. *EMBO J* **10**, 3467-3481.

Spector, D.L., and Lamond, A.I. (2011). Nuclear speckles. *Cold Spring Harb Perspect Biol* **3**, a000646.

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

Spies, N., Nielsen, C.B., Padgett, R.A., and Burge, C.B. (2009). Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell* **36**, 245-254.

Sullivan, K.D., Steiniger, M., and Marzluff, W.F. (2009). A core complex of CPSF73, CPSF100, and Symplekin may form two different cleavage factors for processing of poly(A) and histone mRNAs. *Mol Cell* **34**, 322-332.

Sun, Y., Zhang, Y., Hamilton, K., Manley, J.L., Shi, Y., Walz, T., and Tong, L. (2017). Molecular basis for the recognition of the human AAUAAA polyadenylation signal. *Proc Natl Acad Sci U S A*.

Tabaska, J.E., and Zhang, M.Q. (1999). Detection of polyadenylation signals in human DNA sequences. *Gene* **231**, 77-86.

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

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

Takagaki, Y., Ryner, L.C., and Manley, J.L. (1988). Separation and characterization of a poly(A) polymerase and a cleavage/specificity factor required for pre-mRNA polyadenylation. *Cell* **52**, 731-742.

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

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

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

Tintaru, A.M., Hautbergue, G.M., Hounslow, A.M., Hung, M.L., Lian, L.Y., Craven, C.J., and Wilson, S.A. (2007). Structural and functional analysis of RNA and TAP binding to SF2/ASF. *EMBO Rep* **8**, 756-762.

Ulitsky, I., Shkumatava, A., Jan, C.H., Subtelny, A.O., Koppstein, D., Bell, G.W., Sive, H., and Bartel, D.P. (2012). Extensive alternative polyadenylation during zebrafish development. *Genome Res* **22**, 2054-2066.

Velazquez-Dones, A., Hagopian, J.C., Ma, C.T., Zhong, X.Y., Zhou, H., Ghosh, G., Fu, X.D., and Adams, J.A. (2005). Mass spectrometric and kinetic analysis of ASF/SF2 phosphorylation by SRPK1 and Clk/Sty. *J Biol Chem* **280**, 41761-41768.

Venkataraman, K., Brown, K.M., and Gilmartin, G.M. (2005). Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. *Genes Dev* **19**, 1315-1327.

Wahle, E. (1991). Purification and characterization of a mammalian polyadenylate polymerase involved in the 3' end processing of messenger RNA precursors. *J Biol Chem* **266**, 3131-3139.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476.

Wang, H.Y., Lin, W., Dyck, J.A., Yeakley, J.M., Songyang, Z., Cantley, L.C., and Fu, X.D. (1998). SRPK2: a differentially expressed SR protein-specific kinase involved in mediating the interaction and localization of pre-mRNA splicing factors in mammalian cells. *J Cell Biol* **140**, 737-750.

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-1101.

Weber, C., Schreiber, T.B., and Daub, H. (2012). Dual phosphoproteomics and chemical proteomics analysis of erlotinib and gefitinib interference in acute myeloid leukemia cells. *J Proteomics* **75**, 1343-1356.

West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* **432**, 522-525.

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

White, E., Kamieniarz-Gdula, K., Dye, M.J., and Proudfoot, N.J. (2013). AT-rich sequence elements promote nascent transcript cleavage leading to RNA polymerase II termination. *Nucleic Acids Res* **41**, 1797-1806.

Wickens, M., and Stephenson, P. (1984). Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3' end formation. *Science* **226**, 1045-1051.

Wiestner, A., Tehrani, M., Chiorazzi, M., Wright, G., Gibellini, F., Nakayama, K., Liu, H., Rosenwald, A., Muller-Hermelink, H.K., Ott, G.*, et al.* (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**, 4599-4606.

Winters, M.A., and Edmonds, M. (1973a). A poly(A) polymerase from calf thymus. Characterization of the reaction product and the primer requirement. *J Biol Chem* **248**, 4763-4768.

Winters, M.A., and Edmonds, M. (1973b). A poly(A) polymerase from calf thymus. Purification and properities of the enzyme. *J Biol Chem* **248**, 4756-4762.

Wisniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. *Nat Methods* **6**, 359-362.

Wu, J.Y., and Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* **75**, 1061-1070.

Wu, L., Fan, J., and Belasco, J.G. (2006). MicroRNAs direct rapid deadenylation of mRNA. *Proc Natl Acad Sci U S A* **103**, 4034-4039.

Wu, X., Zhang, Y., and Li, Q.Q. (2016). PlantAPA: A Portal for Visualization and Analysis of Alternative Polyadenylation in Plants. *Front Plant Sci* **7**, 889.

Xia, Z., Donehower, L.A., Cooper, T.A., Neilson, J.R., Wheeler, D.A., Wagner, E.J., and Li, W. (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun* **5**, 5274.

Xiao, Q., Miao, B., Bi, J., Wang, Z., and Li, Y. (2016). Prioritizing functional phosphorylation sites based on multiple feature integration. *Sci Rep* **6**, 24735.

Yan, J., and Marr, T.G. (2005). Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Res* **15**, 369-375.

Yang, Q., Coseno, M., Gilmartin, G.M., and Doublie, S. (2011a). Crystal Structure of a Human Cleavage Factor CFI(m)25/CFI(m)68/RNA Complex Provides an Insight into Poly(A) Site Recognition and RNA Looping. *Structure* **19**, 368-377.

Yang, Q., and Doublie, S. (2011). Structural biology of poly(A) site definition. *Wiley Interdiscip Rev RNA* **2**, 732-747.

Yang, Q., Gilmartin, G.M., and Doublie, S. (2010). Structural basis of UGUA recognition by the Nudix protein CFI(m)25 and implications for a regulatory role in mRNA 3' processing. *Proc Natl Acad Sci U S A* **107**, 10062-10067.

Yang, Q., Gilmartin, G.M., and Doublie, S. (2011b). The structure of human cleavage factor I(m) hints at functions beyond UGUA-specific RNA binding: a role in alternative polyadenylation and a potential link to 5' capping and splicing. *RNA Biol* **8**, 748-753.

Yang, Q., Nausch, L.W., Martin, G., Keller, W., and Doublie, S. (2014). Crystal structure of human poly(A) polymerase gamma reveals a conserved catalytic core for canonical poly(A) polymerases. *J Mol Biol* **426**, 43-50.

Yang, Y., Li, W., Hoque, M., Hou, L., Shen, S., Tian, B., and Dynlacht, B.D. (2016). PAF Complex Plays Novel Subunit-Specific Roles in Alternative Cleavage and Polyadenylation. *PLoS Genet* **12**, e1005794.

Yao, C., Choi, E.A., Weng, L., Xie, X., Wan, J., Xing, Y., Moresco, J.J., Tu, P.G., Yates, J.R., 3rd, and Shi, Y. (2013). Overlapping and distinct functions of CstF64 and CstF64tau in mammalian mRNA 3' processing. *RNA* **19**, 1781-1790.

Yao, C.G., Biesinger, J., Wan, J., Weng, L.J., Xing, Y., Xie, X.H., and Shi, Y.S. (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.

Yeakley, J.M., Tronchere, H., Olesen, J., Dyck, J.A., Wang, H.Y., and Fu, X.D. (1999). Phosphorylation regulates in vivo interaction and molecular targeting of serine/arginine-rich pre-mRNA splicing factors. *J Cell Biol* **145**, 447-455.

Yun, C.Y., Velazquez-Dones, A.L., Lyman, S.K., and Fu, X.D. (2003). Phosphorylation-dependent and -independent nuclear import of RS domain-containing splicing factors and regulators. *J Biol Chem* **278**, 18050-18055.

Zahler, A.M., Lane, W.S., Stolk, J.A., and Roth, M.B. (1992). SR proteins: a conserved family of pre-mRNA splicing factors. *Genes Dev* **6**, 837-847.

Zhang, H., Lee, J.Y., and Tian, B. (2005). Biased alternative polyadenylation in human tissues. *Genome Biol* **6**, R100.

Zhang, H., Rigo, F., and Martinson, H.G. (2015). Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site. *Mol Cell* **59**, 437-448.

Zhang, Z., Wu, C.H., and Gilmour, D.S. (2004). Analysis of polymerase II elongation complexes by native gel electrophoresis. Evidence for a novel carboxyl-terminal domain-mediated termination mechanism. *J Biol Chem* **279**, 23223-23228.

Zhong, X.Y., Wang, P., Han, J., Rosenfeld, M.G., and Fu, X.D. (2009). SR proteins in vertical integration of gene expression from transcription to RNA processing to translation. *Mol Cell* **35**, 1-10.

Zhou, Z., and Fu, X.D. (2013). Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma* **122**, 191-207.

Zhou, Z., Licklider, L.J., Gygi, S.P., and Reed, R. (2002). Comprehensive proteomic analysis of the human spliceosome. *Nature* **419**, 182-185.

Zuo, P., and Maniatis, T. (1996). The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. *Genes Dev* **10**, 1356-1368.

Appendix 1

Studying proliferation-associated regulation of CFIm and APA

A1.1 Introduction

A hypothesis of particular interest at the start of the project was that regulated changes in CFIm activity may drive global 3'UTR shortening upon the entry of cells intro proliferation. This was based on the identification of CFIm as a key global APA regulator, as well as the overlap between CFIm-regulated and differentiation-associated APA events (see Section 1.4), suggesting that CFIm may be an important physiological APA regulator. The BJ-5ta fibroblast cell line was therefore employed as a model system in which synchronised cell cycle progression could be achieved, and CFIm expression and PTMs, as well as APA patterns, were studied in this system. Surprisingly, the previously reported proliferation-associated APA shift could not be fully replicated in this study, and preliminary data suggested no obvious changes in CFIm dynamics, so this avenue of the project was not explored further. Nonetheless, the results are included here as an appendix for general interest.

A1.2 Serum starvation and re-feeding to induce synchronous cell cycle progression in BJ-5ta fibroblasts

To study CFIm in the context of the transition into proliferation, human hTERT-immortalised BJ-5ta fibroblasts were used. These cells show serum-dependent proliferation, such that serum deprivation and re-feeding can induce quiescence followed by relatively synchronous progression of the population through the cell cycle. Another advantage of these cells is that they are diploid, and thus more karyotypically normal than many commonly studied human cell lines (Jiang et al., 1999). Furthermore, proliferation-associated global 3'UTR shortening had been reported in this cell line previously (Elkon

et al., 2012). These cells were therefore considered a useful, physiologically relevant model for studying proliferation-linked APA.

A first step in using these cells was to characterise the serum-dependent cell cycle time course and select appropriate time points for study. Following 48 h of serum deprivation, serum was added to the cells and samples were taken every 3 h from 14 h to 27 h post-serum addition. DNA content was measured as an indication of cell cycle stage, using propidium iodide staining and flow cytometry (Figure A.1). The analysis showed that cell cycle synchronisation was successful, with a single peak for starved cells representing quiescence, with a 2n DNA content, followed by progressively increasing average DNA content at subsequent timepoints (Figure A.1). The 20 h sample appeared to represent an S-phase population, with DNA contents ranging from 2n to 4n, while the 23 h sample had a clear G2/M-phase 4n peak. The 0 h, 20 h and 23 h time points were therefore selected as appropriate time points covering the transition of cells from quiescence into the proliferative state. This time course was highly reproducible upon repeated serum starvation experiments (data not shown).

A1.3 Analysis of APA patterns across a cell cycle time course

Having established the cell cycle time course, the next step was to verify the expected APA-mediated 3'UTR shortening upon entry of the BJ-5ta cells into proliferation from quiescence. A previous global 3' end sequencing study had identified widespread 3'UTR shortening in the same cell type when comparing asynchronously proliferating cells to confluent, contact-inhibited quiescent cells (Elkon et al., 2012). This system was arguably a less rigorously controlled proliferation model than the serum starvation method, and it was therefore expected that the APA shift might be even more dramatic between serum-starved and serum re-fed cells. As a starting point, the northern blotting approach with *TIMP2* and *SDC2* was used to assess APA in contact-inhibited vs. asynchronously proliferating cells and in the 0 h, 20 h and 23 h serum re-fed samples. Analysis of these transcripts was deemed appropriate considering that *TIMP2* showed one of the most dramatic proliferation-linked 3'UTR shortening

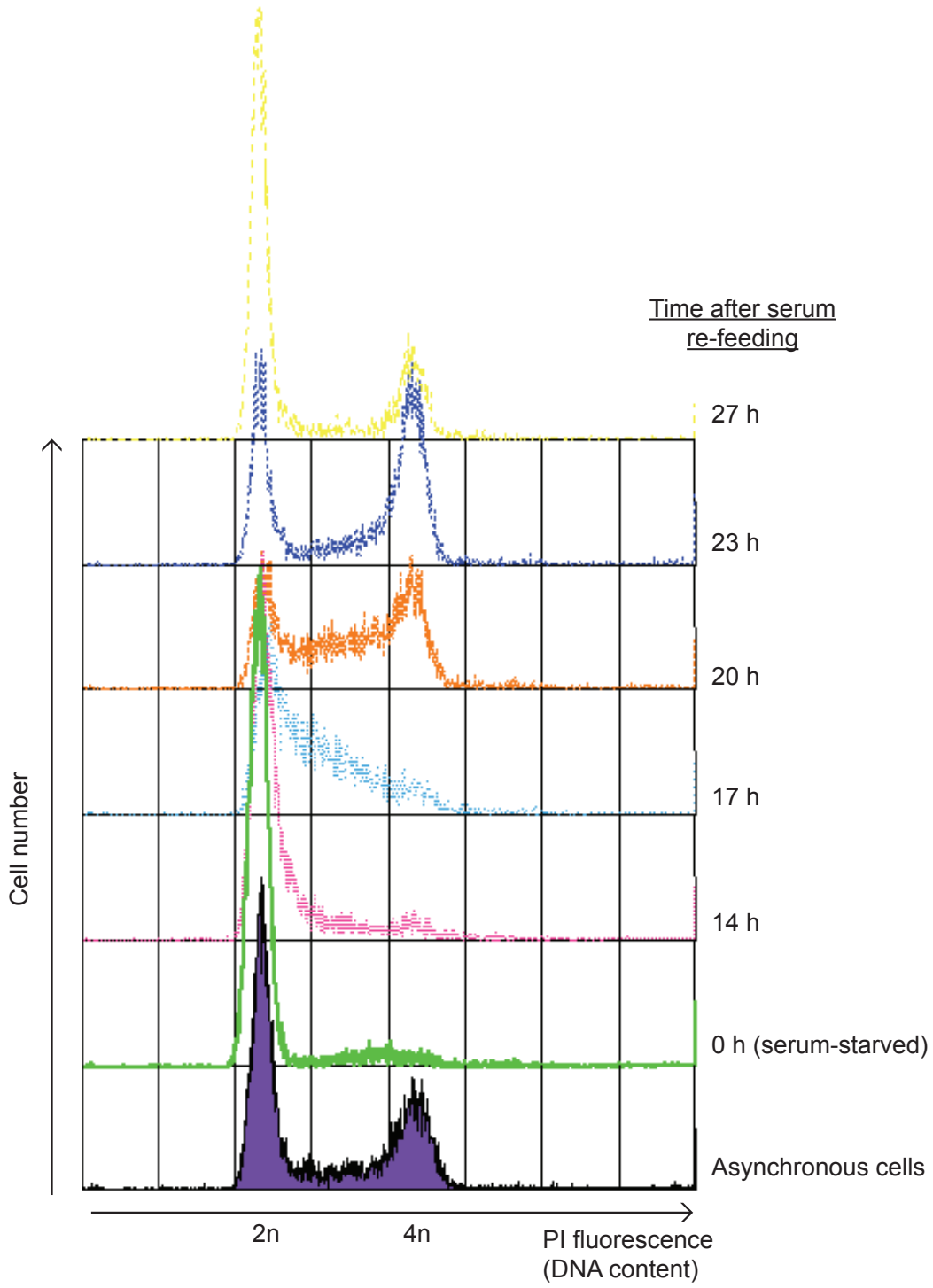


Figure A.1: Cell cycle analysis in BJ-5ta fibroblasts. BJ-5ta fibroblasts were serum-starved for 48 h, then serum was re-added and samples were harvested at the given times. Ethanol-fixed cells were stained with propidium iodide (PI) and fluorescence of single cells was measured using a flow cytometer. Histograms are overlaid to compare cell cycle stages - 2n represents inferred normal diploid DNA content of cells in G1; 4n represents cells in G2 or M; cells with intermediate DNA content are in S phase.

responses in the previous study (Elkon et al., 2012). Flow cytometry confirmed the cell cycle profiles were as expected (data not shown).

Surprisingly, no change in APA pattern for *TIMP2* or *SDC2* was observed between quiescent and proliferating cells using either comparison method (Figure A.2A). This was in stark contrast to the previous report, and suggested a limited effect of proliferation on poly(A) site selection for these transcripts. To explore this result further, the 3' end sequencing approach (see Section 3.5) was employed to assess transcriptome-wide APA in the quiescent and proliferating cells. Comparisons of transcriptome-wide P/D ratios between appropriate pairs of samples are plotted in Figure A.2B. In all three comparisons, more transcripts showed 3'UTR shortening than lengthening in the proliferating cells, providing general support for the previously reported directional bias. However, the numbers of transcripts affected were relatively low, and amongst the APA changes, the proportion of 3'UTR lengthening events was relatively high. In the asynchronous vs. confluent comparison, only 5.2% of APA-regulated transcripts showed a significant change in APA site use between the two conditions, 67.2% of which were 3'UTR shortening in the proliferative condition. These effects are much smaller than those observed in a previous study using exactly the same system – in that case, 34.6% of transcripts underwent significant APA changes, 83% of which were 3'UTR shortening (Elkon et al., 2012).

Overall, these observations raise doubts as to the reproducibility and extent of proliferation-associated 3'UTR shortening, and call into question the usefulness of BJ-5ta cells for studying APA regulators.

A1.4 Analysis of CFIm expression and PTM profiles across a cell cycle time course

Though the observed APA modulation in proliferating BJ-5ta cells was relatively limited, there was a general bias towards 3'UTR shortening, and it was still possible that CFIm could be regulating some of the APA changes observed. An overlap analysis between transcripts showing 3'UTR shortening in 25KD

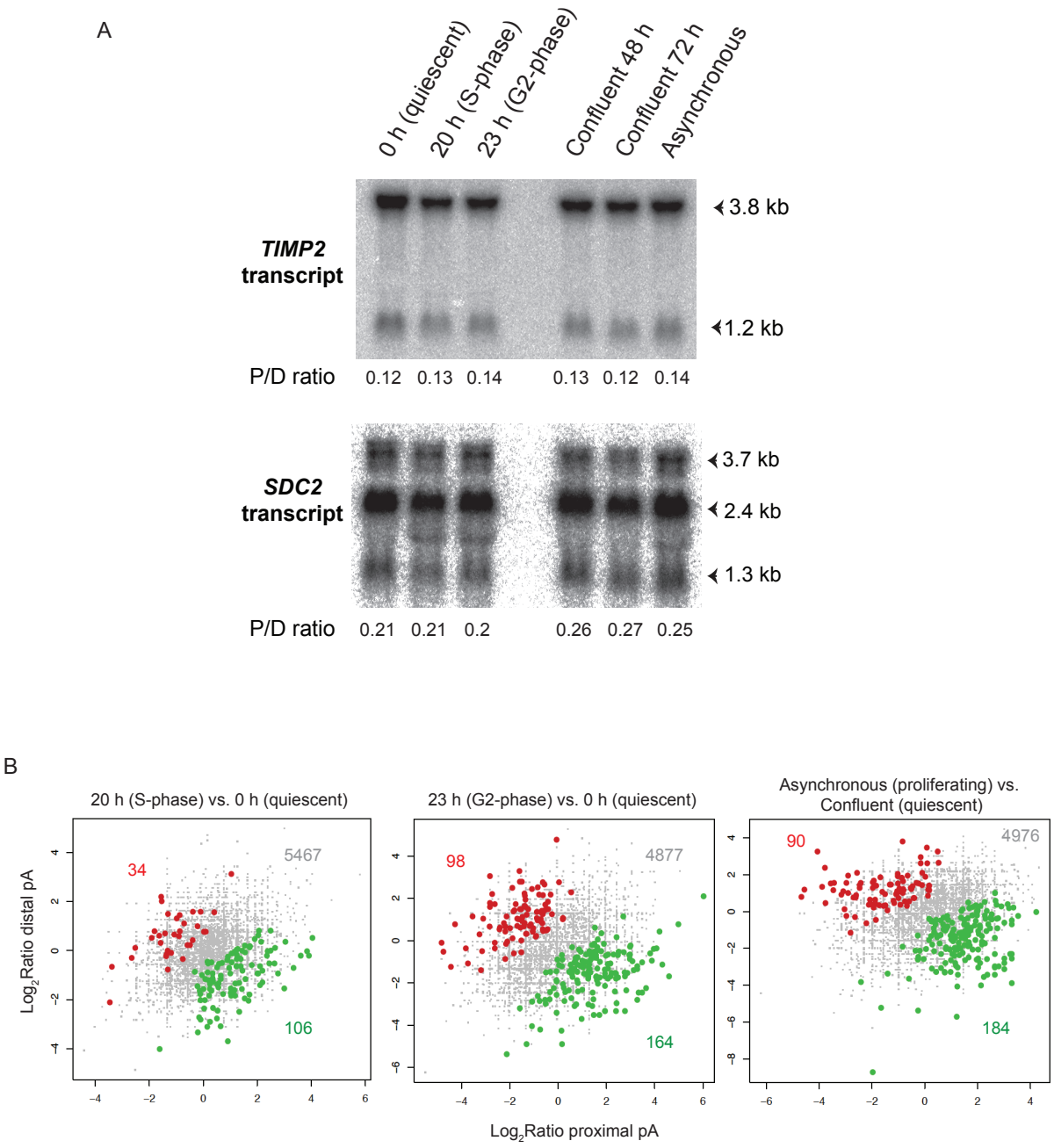


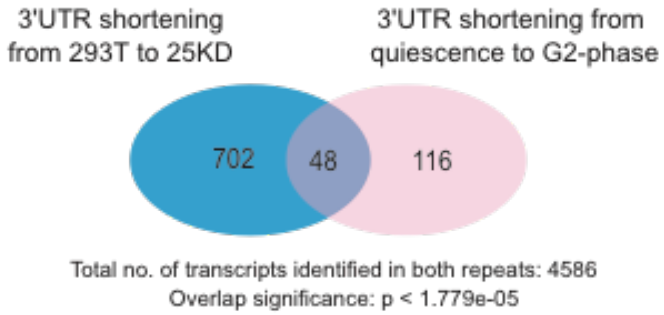
Figure A.2: Comparison of APA patterns between quiescent and proliferating cells. RNA was isolated from the given BJ-5ta cell samples. 0 h, 20 h and 23 h refer to times after serum re-feeding as represented in Figure A.1. 'Confluent' cells were grown until they completely covered the dish and left in that state for the given time. 'Asynchronous' cells were non-confluent, normally proliferating cells harvested two days after seeding. (A) Northern blotting and P/D ratio calculations were performed on the given samples as described in Figure 3.3. (B) 3' end sequencing analysis was carried out on the given samples, with preparation of libraries, data analysis and plotting of transcript dots performed as described in Figure 3.5.

cells (see Figure 5.9A, red dots) and in G2-phase cells compared to quiescent cells (see Figure A.2B, middle panel, green dots) revealed a significant overlap, with 48 of the 164 G2-shortened transcripts also showing 3'UTR shortening in 25KD cells (Figure A.3A). This suggested a potential involvement of CFIm regulation in the system, and as a next step expression levels of CFIm across the time course were analysed to explore the possibility that down-regulation of CFIm25/CFIm68 expression might underlie the observed 3'UTR shortening events. However, western blot analysis revealed no obvious change in expression levels of CFIm subunits across the time course (Figure A.3B).

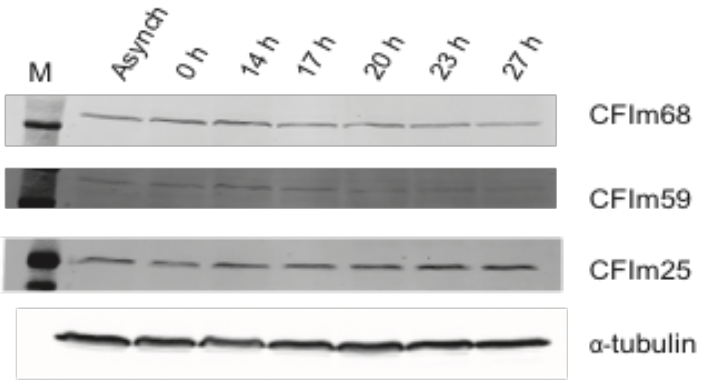
Another hypothesis of interest was that in the absence of changes in expression level, CFIm activity may be regulated by PTMs. The CFIm IP/MS approach (see Section 5.3) was applied to BJ-5ta cells from the 0 h and 20 h (quiescent vs. S-phase) time points, and a label-free quantitative analysis of the data was performed. Only four modifications were detected with high enough confidence for the analysis to be performed, and none of these showed more than a two-fold difference in abundance between the two conditions (Figure A.3C). This was deemed well within the potential error of measurement, such that no meaningful conclusion about increased/decreased levels of these PTMs could be drawn. This suggests that a change in levels of acetylation of K23/K173 on CFIm25, K139 on CFIm68 and/or phosphorylation of S46/47/48 on CFIm59 are unlikely to be involved in proliferation-associated changes in APA regulation. As this approach identified very few PTMs, it remains possible that other modifications undergo important changes that were not detected. Clearly, further method optimisation, and possibly use of the pIMAGO approach, would be needed to explore this further.

In light of the data discussed here, the decision was taken not to explore potential roles for CFIm in proliferation-associated APA any further. While there was some degree of general 3'UTR shortening in the BJ-5ta proliferation model, the extent of this was limited, and there were no preliminary indications that CFIm undergoes large proliferation-associated changes in expression or modification. Future investigation of other model systems or aspects of CFIm may yet reveal important regulatory roles for this complex, though it will first be necessary to resolve the apparent discrepancies between this study and others in identification of so-called ‘global’ APA modulation.

A



B



C

Modification	Relative quantification (S-phase/quiescent)
CFIm25 K23 acetylation	0.56
CFIm25 K173 acetylation	0.78
CFIm68 K139 acetylation	0.53
CFIm59 S46/S47/S48 phosphorylation (ambiguous site)	0.81

Figure A.3: Analysis of CFIm in quiescent and proliferating cells. (A) Venn diagram of overlap between transcript identities showing 3'UTR shortening in 25KD cells (compared to 293T parent) and in G2-phase BJ-5ta fibroblasts (compared to serum-starved quiescent population). Overlap significance was calculated as described in Figure 3.5. (B) Whole cell lysates from BJ-5ta cells at the given timepoints (as defined in Figure A.1) were subjected to western blotting with the given antibodies to measure CFIm subunit expression across a cell cycle timecourse. M = marker lane. (C) MS analysis of CFIm was performed to detect PTMs in quiescent (starved) and S-phase (20 h timepoint) BJ-5ta cells. Label-free quantitation of PTMs was performed using MaxQuant software, with normalisation of modified peptides to total protein level. S-phase values were divided by quiescent phase values to give the relative quantification of each modification in proliferating compared to quiescent cells.

Appendix 2

Table of oligonucleotides

All oligonucleotides were purchased from Life Technologies, and are listed in approximately the order they first appear within the thesis.

Code	Oligo name	Purpose	Sequence
JG245	CFIm25 CRISPR site Fwd	Amplification of *NUDT21* exon 1 CRISPR-targeted site for sequencing	CGCACTTACAGGTTGATGGT
JG246	CFIm25 CRISPR site Rvs	Amplification of *NUDT21* CRISPR-targeted site for sequencing	GCTGAGGGGATTGCAGATAA
JG97	M13 Fwd	Sequencing/colony PCR for pCR2.1	GTAAAACGACGGCCAGT
JG98	M13 Rvs	Sequencing/colony PCR for pCR2.1	CAGGAAACAGCTATGAC
JG247	CFIm68 CRISPR site Fwd	Amplification of *CPSF6* CRISPR-targeted site for sequencing	TGAGGGGGAAAATATCTTGCAGT
JG248	CFIm68 CRISPR site Rvs	Amplification of *CPSF6* CRISPR-targeted site for sequencing	CCTCCTTCCAATGTAAACAATCATG
JG67	TIMP2 Fwd	Amplification of *TIMP2* CDS for northern probe	CGCAACAGGCGTTTTGCAAT
JG68	TIMP2 Rvs	Amplification of *TIMP2* CDS for northern probe	TGGTGCCCGTTGATGTTCTT
JG69	SDC2 Fwd	Amplification of *SDC2* CDS for northern probe	TGTACCTTGACAACAGCTCC
JG70	SDC2 Rvs	Amplification of *SDC2* CDS for northern probe	GCCAATAACTCCACCAGCAA
JG75	ACTB Fwd	Amplification of *ACTB* CDS for northern probe	GGATTCCTATGTGGGCGACG
JG76	ACTB Rvs	Amplification of *ACTB* CDS for northern probe	GTAGTCAGTCAGGTCCCGGC

JG249	*Hin*DIII-CFIm25 Fwd	Original amplification of *NUDT21* CDS from cDNA	GAGAAAGCTTAGCATGTCTGTGGTACCGCCC
JG250	CFIm25-*Bam*HI Rvs	Original amplification of *NUDT21* CDS from cDNA	CGGGATCCTCAGTTGTAAATAAAATTGAACCTGC
JG58	*Not*I-CFIm68 Fwd	Original amplification of *CPSF6* CDS from cDNA	GGGGAGCGGCCGCATGGCGGACGGCGTG
JG61	CFIm68-*Xho*I Rvs	Original amplification of *CPSF6* CDS from cDNA	GGGGCTCGAGCTAACGATGACGATATTCGCGCTC
JG133	CMV Fwd	Sequencing plasmid inserts	CGCAAATGGGCGGTAGGCGTG
JG132	BGH Rvs	Sequencing plasmid inserts	TAGAAGGCACAGTCGAGG
JG52	CFIm68 seq Fwd	Sequencing of *CPSF6* (middle section of gene) following mutagenesis	ACGCCATTGAGTGAAGCTGA
JG124	CFIm25 Y158A Y160A Fwd	Mutagenesis	TGAACCTCCTCAGGCTCCAGCTATTCCTGCAC
JG125	CFIm25 Y158A Y160A Rvs	Mutagenesis	GTGCAGGAATAGCTGGAGCCTGAGGAGGTTCA
JG163	CFIm68 Δ116-122 Fwd	Mutagenesis	TGGAGATAAAATTTTTTTCAAAGGGGTTTGCCCTTG
JG164	CFIm68 Δ116-122 Rvs	Mutagenesis	CAAGGGCAAACCCCTTTGAAAAAAATTTTATCTCCA
JG159	CFIm68 ΔN117 Fwd	Mutagenesis	GGAGATAAAATTTTTTGAACGAGCAAATGGCCAGTCA
JG160	CFIm68 ΔN117 Rvs	Mutagenesis	TGACTGGCCATTTGCTCGTTCAAAAAATTTTATCTCC
JG161	CFIm68 ΔN120 Fwd	Mutagenesis	AATTTTTTGAAAATCGAGCAGGCCAGTCAAAGGGGT
JG162	CFIm68 ΔN120 Rvs	Mutagenesis	ACCCCTTTGACTGGCCTGCTCGATTTTCAAAAAATT
JG126	CFIm25 E55A Fwd	Mutagenesis	AGAGCCCCTCTACGCGAAGGACAGCTCTGT
JG127	CFIm25 E55A Rvs	Mutagenesis	ACAGAGCTGTCCTTCGCGTAGAGGGGCTCT

JG128	CFIm25 R63A Fwd	Mutagenesis	GCTCTGTTGCAGCCGCATTTCAGCGCATGAG
JG129	CFIm25 R63A Rvs	Mutagenesis	CTCATGCGCTGAAATGCGGCTGCAACAGAGC
JG130	CFIm25 F104A Fwd	Mutagenesis	GCTGGGAACAACTTTCGCCAAACTACCTGGTGGTG
JG131	CFIm25 F104A Rvs	Mutagenesis	CACCACCAGGTAGTTTGGCGAAAGTTGTTCCCAGC
JG134	CFIm25 F103A Fwd	Mutagenesis	GCTGGGAACAACTGCCTTCAAACTACCTGGTGGTG
JG135	CFIm25 F103A Rvs	Mutagenesis	CACCACCAGGTAGTTTGAAGGCAGTTGTTCCCAGC
JG181	CFIm68 ΔRS1 Fwd	Mutagenesis	GATTGCCTTCATGGAATTCGACGTCATAAATCCCGT
JG182	CFIm68 ΔRS1 Rvs	Mutagenesis	ACGGGATTTATGACGTCGAATTCCATGAAGGCAATC
JG165	*Kpn*I-CFIm68 ΔRS2 Fwd	Mutagenesis (PCR amplification followed by digest/ligation into pcDNA5/FRT)	CGCGGTACCAGCGGCCGCATGGC
JG185	CFIm68 ΔRS2-*Xho*I Rvs	Mutagenesis (PCR amplification followed by digest/ligation into pcDNA5/FRT)	CCGCCTCGAGCTATTTATGACGTCGACTCTTTTCTCG
JG241	CFIm68 ΔpNLS Fwd	Mutagenesis	CCATAGTAGATCACGAGAAAAGAGTTCCCGTAGTAGAG
JG242	CFIm68 ΔpNLS Rvs	Mutagenesis	CTCTACTACGGGAACTCTTTTCTCGTGATCTACTATGG
JG189	SDC2 TSS Fwd	qPCR analysis of pol II ChIP-seq samples	GAGGGAGCCAGAGGAAAAG
JG190	SDC2 TSS Rvs	qPCR analysis of pol II ChIP-seq samples	GTATCTGGAGAGCAGCGC
JG191	SDC2 DS2 Fwd	qPCR analysis of pol II ChIP-seq samples (1 kb downstream of SDC2)	TGTATATGGTCTTTAGCAGGGC
JG192	SDC2 DS2 Rvs	qPCR analysis of pol II ChIP-seq samples (1 kb downstream of SDC2)	CTATTTGCGGGTTGGATTGAC

JG195	SDC2 DS early Fwd	qPCR analysis of pol II ChIP-seq samples (300 bp downstream of SDC2)	AGTGAATATGGCTTTGTGTTTGAG
JG196	SDC2 DS early Rvs	qPCR analysis of pol II ChIP-seq samples (300 bp downstream of SDC2)	AACCCAAGTGCTTCCTATTCC
JG19	CFIm25 R15A Fwd	Mutagenesis	GACCGGCTGGCCCGCGGGGGTCACTCAGT
JG20	CFIm25 R15A Rvs	Mutagenesis	ACTGAGTGACCCCCGCGGGCCAGCCGGTC
JG138	CFIm25 K23R Fwd	Mutagenesis	ACTCAGTTCGGCAACAGGTACATCCAGCAGACGAA
JG139	CFIm25 K23R Rvs	Mutagenesis	TTCGTCTGCTGGATGTACCTGTTGCCGAACTGAGT
JG27	CFIm25 Y24F Fwd	Mutagenesis	TCAGTTCGGCAACAAGTTCATCCAGCAGACGAAG
JG28	CFIm25 Y24F Rvs	Mutagenesis	CTTCGTCTGCTGGATGAACTTGTTGCCGAACTGA
JG118	CFIm25 K29R Fwd	Mutagenesis	GTACATCCAGCAGACGAGGCCCCTCACCCTGGA
JG119	CFIm25 K29R Rvs	Mutagenesis	TCCAGGGTGAGGGGCCTCGTCTGCTGGATGTAC
JG251	CFIm25 Y40F Fwd	Mutagenesis	CGCACCATCAACCTGTTCCCTCTTACCAATTATAC
JG252	CFIm25 Y40F Rvs	Mutagenesis	GTATAATTGGTAAGAGGGAACAGGTTGATGGTGCG
JG144	CFIm25 K56R Fwd	Mutagenesis	CCTCTACGAGAGGGACAGCTCTGTTGC
JG145	CFIm25 K56R Rvs	Mutagenesis	GCAACAGAGCTGTCCCTCTCGTAGAGG
JG146	CFIm25 K73R Fwd	Mutagenesis	AGGGAAGAATTTGATAGAATTGGAATGAGGAG
JG147	CFIm25 K73R Rvs	Mutagenesis	CTCCTCATTCCAATTCTATCAAATTCTTCCCT
JG148	CFIm25 T101A T102A Fwd	Mutagenesis	GCAGCTGGGAGCAGCTTTCTTCAAACTAC
JG149	CFIm25 T101A T102A Rvs	Mutagenesis	GTAGTTTGAAGAAAGCTGCTCCCAGCTGC

JG140	CFIm25 K122R Fwd	Mutagenesis	AAGTTGAAGGACTAAGACGCTTAATGACAG
JG141	CFIm25 K122R Rvs	Mutagenesis	CTGTCATTAAGCGTCTTAGTCCTTCAACTT
JG116	CFIm25 K173R Fwd	Mutagenesis	GCCTAAGGAACATAAGAGGTTGTTTCTGGTTCAGC
JG117	CFIm25 K173R Rvs	Mutagenesis	GCTGAACCAGAAACAACCTCTTATGTTCCTTAGGC
JG152	CFIm25 K173E Fwd	Mutagenesis	GCCTAAGGAACATAAGGAGTTGTTTCTGGTTCAGC
JG153	CFIm25 K173E Rvs	Mutagenesis	GCTGAACCAGAAACAACTCCTTATGTTCCTTAGGC
JG142	CFIm25 K189R Fwd	Mutagenesis	TGCAGTCCCTAGAAATTACAAGCTGGTAGC
JG143	CFIm25 K189R Rvs	Mutagenesis	GCTACCAGCTTGTAATTTCTAGGGACTGCA
JG9	CFIm68 Y74F Y76F Fwd	Mutagenesis	GCACCAAATGTTGTCTTTACATTTACTGGAAAGAGAATTG
JG10	CFIm68 Y74F Y76F Rvs	Mutagenesis	CAATTCTCTTTCCAGTAAATGTAAAGACAACATTTGGTGC
JG120	CFIm68 K139R Fwd	Mutagenesis	GCATCTTCAAAAAGGTTAATGGATCTGTTACCT
JG121	CFIm68 K139R Rvs	Mutagenesis	AGGTAACAGATCCATTAACCTTTTTGAAGATGC
JG122	CFIm68 K161R Fwd	Mutagenesis	GTAACTCCATGCAATAGACAGTTCCTGAGTCA
JG123	CFIm68 K161R Rvs	Mutagenesis	TGACTCAGGAACTGTCTATTGCATGGAGTTAC
JG93	CFIm68 S165A Fwd	Mutagenesis	ACAGTTCCTGGCTCAATTTGAAATGCAGTCCAGGA
JG94	CFIm68 S165A Rvs	Mutagenesis	TCCTGGACTGCATTTCAAATTGAGCCAGGAACTGT
JG175	CFIm68 S165D Fwd	Mutagenesis	ACAGTTCCTGGATCAATTTGAAATGCAGTCCAGG

JG176	CFIm68 S165D Rvs	Mutagenesis	CCTGGACTGCATTTCAAATTGATCCAGGAACTGT
JG169	CFIm68 R202A Fwd	Mutagenesis	GCATTTCCACAAGGTGGTGCAGGACGGGGCCG
JG170	CFIm68 R202A Rvs	Mutagenesis	CGGCCCCGTCCTGCACCACCTTGTGGAAATGC
JG171	CFIm68 R204A Fwd	Mutagenesis (used after R202A)	TCCACAAGGTGGTGCAGGAGCGGGCCGTTTTC
JG172	CFIm68 R204A Rvs	Mutagenesis (used after R202A)	GAAAACGGCCCGCTCCTGCACCACCTTGTGGA
JG173	CFIm68 R206A Fwd	Mutagenesis (used after R202A and R204A)	TGGTGCAGGAGCGGGCGCTTTTCCAGGGGC
JG174	CFIm68 R206A Rvs	Mutagenesis (used after R202A and R204A)	GCCCCTGGAAAAGCGCCCGCTCCTGCACCA
JG157	CFIm68 T404A T407A Fwd	Mutagenesis	GAAATGGATGCTGCAAGAGCGCCATTGAGTG
JG158	CFIm68 T404A T407A Rvs	Mutagenesis	CACTCAATGGCGCTCTTGCAGCATCCATTTC
JG177	CFIm68 T404D Fwd	Mutagenesis	GAAGGGAAATGGATGATGCAAGAACGCCATTGAGT
JG178	CFIm68 T404D Rvs	Mutagenesis	ACTCAATGGCGTTCTTGCATCATCCATTTCCCTTC
JG179	CFIm68 T407D Fwd	Mutagenesis (used after T404D)	GGAAATGGATGATGCAAGAGACCCATTGAGTGAAGCTGAAT
JG180	CFIm68 T407D Rvs	Mutagenesis (used after T404D)	ATTCAGCTTCACTCAATGGGTCTCTTGCATCATCCATTTCC
JG229	CFIm68 Y485F Fwd	Mutagenesis	GGAATTGAGTCCAAGTCTTTTGGTTCTGGATCAAGACG
JG230	CFIm68 Y485F Rvs	Mutagenesis	CGTCTTGATCCAGAACCAAAAGACTTGGACTCAATTCC
JG201	CFIm68 S511A Fwd	Mutagenesis	TCGACGTCATAAAGCCCGTAGTAGAGACCGTC
JG202	CFIm68 S511A Rvs	Mutagenesis	GACGGTCTCTACTACGGGCTTTATGACGTCGA

JG211	CFIm68 S511D Fwd	Mutagenesis	CGACGTCATAAAGACCGTAGTAGAGACCGTCATGA
JG212	CFIm68 S511D Rvs	Mutagenesis	TCATGACGGTCTCTACTACGGTCTTTATGACGTCG
JG215	CFIm68 S513A Fwd	Mutagenesis	CGTCATAAATCCCGTGCTAGAGACCGTCATGAC
JG216	CFIm68 S513A Rvs	Mutagenesis	GTCATGACGGTCTCTAGCACGGGATTTATGACG
JG217	CFIm68 S513D Fwd	Mutagenesis (used on wild type)	CGTCATAAATCCCGTGATAGAGACCGTCATGAC
JG218	CFIm68 S513D Rvs	Mutagenesis (used on wild type)	GTCATGACGGTCTCTATCACGGGATTTATGACG
JG213	CFIm68 S513D 2 Fwd	Mutagenesis (used after S511D)	CGTCATAAAGACCGTGATAGAGACCGTCATGAC
JG214	CFIm68 S513D 2 Rvs	Mutagenesis (used after S511D)	GTCATGACGGTCTCTATCACGGTCTTTATGACG
JG203	CFIm68 S511A S513A Fwd	Mutagenesis	CGACGTCATAAAGCCCGTGCTAGAGACCGTCATGAC
JG204	CFIm68 S511A S513A Rvs	Mutagenesis	GTCATGACGGTCTCTAGCACGGGCTTTATGACGTCG
JG15	CFIm68 Y520F Y521F Fwd	Mutagenesis	ACCGTCATGACGATTTTTTCAGAGAGAGAAGCA
JG16	CFIm68 Y520F Y521F Rvs	Mutagenesis	TGCTTCTCTCTCTGAAAAAATCGTCATGACGGT
JG207	CFIm68 S525A Fwd	Mutagenesis	ATTACAGAGAGAGAGCCAGAGAACGAGAGAGG
JG208	CFIm68 S525A Rvs	Mutagenesis	CCTCTCTCGTTCTCTGGCTCTCTCTCTGTAAT
JG209	CFIm68 S525D Fwd	Mutagenesis	ATTACAGAGAGAGAGACAGAGAACGAGAGAGG
JG210	CFIm68 S525D Rvs	Mutagenesis	CCTCTCTCGTTCTCTGTCTCTCTCTCTGTAAT
JG243	CFIm68 Y548F Fwd	Mutagenesis	GACCGAGAGCGCGAATTTCGTCATCGTTAGCT

JG244	CFIm68 Y548F Rvs	Mutagenesis	AGCTAACGATGACGAAATTCGCGCTCTCGGTC
JG231	CFIm68 RS1 S-to-A 1 Fwd	Fragment for annealing to make RS1 S-to-A mutant	TAAACAATCCAAAGTATCTGCTGATGATCGTTGCAAAGTTCT TATTAGTTCTTT
JG236	CFIm68 RS1 S-to-A 1 Rvs	Fragment for annealing to make RS1 S-to-A mutant	TGAAGGCAATCTTGCAAAGAACTAATAAGAACTTTGCAACG ATCATCAGCAGATACTTTGGATTGTTTAAT
JG232	CFIm68 RS1 S-to-A 2 Fwd	Fragment for annealing to make RS1 S-to-A mutant	GCAAGATTGCCTTCATGGAATTGAGGCCAAGGCTTATGGTGC TGGAGCAAGACGTGAACGAG
JG237	CFIm68 RS1 S-to-A 2 Rvs	Fragment for annealing to make RS1 S-to-A mutant	GGTCCCTCTCTCTTGCTCGTTCACGTCTTGCTCCAGCACCATA AGCCTTGGCCTCAATTCCA
JG233	CFIm68 RS1 S-to-A 3 Fwd	Fragment for annealing to make RS1 S-to-A mutant	CAAGAGAGAGGGACCATGCTAGAGCACGAGAAAAGGCTCGA CGTCATAAATCCCGTAGTAGA
JG238	CFIm68 RS1 S-to-A 3 Rvs	Fragment for annealing to make RS1 S-to-A mutant	ATCGTCATGACGGTCTCTACTACGGGATTTATGACGTCGAGC CTTTTCTCGTGCTCTAGCAT
JG234	CFIm68 RS1 S-to-A 4 Fwd	Fragment for annealing to make RS1 S-to-A mutant	GACCGTCATGACGATTATTACAGAGAGAGAAGCAGAGAACG AGAGAGGCACCGGGATCGTGA
JG239	CFIm68 RS1 S-to-A 4 Rvs	Fragment for annealing to make RS1 S-to-A mutant	CGGTCACGGTCTCGGTCACGATCCCGGTGCCTCTCTCGTTCTC TGCTTCTCTCTCTGTAATA
JG235	CFIm68 RS1 S-to-A 5 Fwd	Fragment for annealing to make RS1 S-to-A mutant	CCGAGACCGTGACCGAGAGCGTGACCGAGAGCGCGAATATC GTCATCGTTAGC
JG240	CFIm68 RS1 S-to-A 5 Rvs	Fragment for annealing to make RS1 S-to-A mutant	TCGAGCTAACGATGACGATATTCGCGCTCTCGGTCACGCTCT

Appendix 3

CFIm25/CFIm68 vertebrate protein sequence alignments

FASTA sequence files were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/) for the vertebrates indicated (*Homo sapiens, Pan troglodytes, Bos taurus, Mus musculus, Gallus gallus, Python bivittatus, Xenopus tropicalis, Latimeria chalumnae, Danio rerio, Callorhinchus milii*) and aligned using the Clustal Omega tool (http://www.ebi.ac.uk/Tools/msa/clustalo/).

CFIm25

```
Homo            MSVVPPNRSQTGWPRGVTQFGNKYIQQ-TKPLTLERTINLYPLTNYTFGT 49
Pan             MSVVPPNRSQTGWPRGVTQFGNKYIQQ-TKPLTLERTINLYPLTNYTFGT 49
Bos             MSVVPPNRSQTGWPRGVTQFGNKYIQQ-TKPLTLERTINLYPLTNYTFGT 49
Mus             MSVVPPNRSQTGWPRGVNQFGNKYIQQ-TKPLTLERTINLYPLTNYTFGT 49
Gallus          MSVVPPNRSQTGWPRGVNQFGNKYIQQ-TKPLTLERTINLYPLTNYTFGT 49
Python          MSVVPPNRSQTGWPRGVNQFGNKYIQQ-SKPLTLERTINLYPLTNYTFGT 49
Xenopus         MSVLPPNRSQTGWPRGVNQFGNKYIQQ-TKPLTLERTINLYPLTNYTFGT 49
Latimeria       MSVVPPNRSQTGWPRGVTQFGNKYIQQ-SKPLTLERTINLYPLTNYTFGT 49
Danio           MSVVPPNRSSTGWPRGVNQFGNKYITQATKPLTLERTINLYPLTNYTFGT 50
Callorhinchus   MSVVPPSRSQAGWPRGVSQYGNKYLQ--AKPLTLERTINLYPLTNYTFGT 48
                ***:**.**.:******.*:****:   :*********************

Homo            KEPLYEKDSSVAARFQRMREEFDKIGMRRTVEGVLIVHEHRLPHVLLLQL 99
Pan             KEPLYEKDSSVAARFQRMREEFDKIGMRRTVEGVLIVHEHRLPHVLLLQL 99
Bos             KEPLYEKDSSVAARFQRMREEFDKIGMRRTVEGVLIVHEHRLPHVLLLQL 99
Mus             KEPLYEKDSSVAARFQRMREEFDKIGMRRTVEGVLIVHEHRLPHVLLLQL 99
Gallus          KEPLYEKDSSVAARFQRMREEFDKIGMRRTVEGVLIVHEHRLPHVLLLQL 99
Python          KEPLYEKDSSVAARFQRMREEFDKIGMRRTVEGVLIVHEHRLPHVLLLQL 99
Xenopus         KEPLYEKDSSVAARFQRMREEFDKIGMRRTVEGVLIVHEHRLPHVLLLQL 99
Latimeria       KEPLYEKDSSVAARFQRMREEFDKIGMRRTVEGVLIVHEHRLPHVLLLQL 99
Danio           KEPLYEKDSSVAARFQRMREEFEKIGMRRTVEGVLIVHEHRLPHVLLLQL 100
Callorhinchus   KEPLYEKDSSVAARFQRMREEFEKIGMRRTVEGVLIVHEHRLPHVLLLQL 98
                **********************:**************************

Homo            GTTFFKLPGGELNPGEDEVEGLKRLMTEILGRQDGVLQDWVIDDCIGNWW 149
Pan             GTTFFKLPGGELNPGEDEVEGLKRLMTEILGRQDGVLQDWVIDDCIGNWW 149
Bos             GTTFFKLPGGELNPGEDEVEGLKRLMTEILGRQDGVLQDWVIDDCIGNWW 149
Mus             GTTFFKLPGGELNPGEDEVEGLKRLMTEILGRQDGVLQDWVIDDCIGNWW 149
Gallus          GTTFFKLPGGELNPGEDEVEGLKRLMTEILGRQDGVQQDWVIDDCIGNWW 149
Python          GTTFFKLPGGELNPGEDEVEGLKRLMTEILGRQDGVLQDWVIDDCIGNWW 149
Xenopus         GTTFFKLPGGELNPGEDEVEGLKRLMTEILGRQDGVQQDWVIDDCIGNWW 149
Latimeria       GTTFFKLPGGELNPGEDEVEGLKRLMTEILGRQDGVLQDWVIDDCIGNWW 149
Danio           GTTFFKLPGGELNPGEDEVEGLKRLMTEILGRQDGVKQDWVIDDCIGNWW 150
Callorhinchus   GTTFFKLPGGELNPGEDEVEGLKRLMTEILGRQDGVQQDWVIDDCIGNWW 148
                ************************************ ************

Homo            RPNFEPPQYPYIPAHITKPKEHKKLFLVQLQEKALFAVPKNYKLVAAPLF 199
Pan             RPNFEPPQYPYIPAHITKPKEHKKLFLVQLQEKALFAVPKNYKLVAAPLF 199
Bos             RPNFEPPQYPYIPAHITKPKEHKKLFLVQLQEKALFAVPKNYKLVAAPLF 199
Mus             RPNFEPPQYPYIPAHITKPKEHKKLFLVQLQEKALFAVPKNYKLVAAPLF 199
Gallus          RPNFEPPQYPYIPAHITKPKEHKKLFLVQLQEKALFAVPKNYKLVAAPLF 199
Python          RPNFEPPQYPYIPAHITKPKEHKKLFLVQLQEKALFAVPKNYKLVAAPLF 199
Xenopus         RPNFEPPQYPYIPAHITKPKEHKKLFLVQLQEKALFAVPKNYKLVAAPLF 199
Latimeria       RPNFEPPQYPYIPAHITKPKEHKKLFLVQLQEKALFAVPKNYKLVAAPLF 199
Danio           RPNFEPPQYPYIPAHITKPKEHKKLFLVQLQEKALFAVPKNYKLVAAPLF 200
Callorhinchus   RPNFEPPQYPYIPAHITKPKEHKKLFLVQLQEKALFAVPKNYKLVAAPLF 198
                **************************************************
```

```
Homo            ELYDNAPGYGPIISSLPQLLSR-FNFIYN 227
Pan             ELYDNAPGYGPIISSLPQLLSR-FNFIYN 227
Bos             ELYDNAPGYGPIISSLPQLLSR-FNFIYN 227
Mus             ELYDNAPGYGPIISSLPQLLSR-FNFIYN 227
Gallus          ELYDNAPGYGPIISSLPQLLSR-FNFIYN 227
Python          ELYDNAPGYGPIISSLPQLLSR-FNFIYN 227
Xenopus         ELYDNAPGYGPIISSLPQLLSR-FNFIYN 227
Latimeria       ELYDNAPGYGPIISSLPQLLSRTFGPPAA 228
Danio           ELYDNAPGYGPIISSLPQLLSR-FNFIYN 228
Callorhinchus   ELYDNAPGYGPIISSLPQLLSR-FNFIYN 226
                ********************** *.
```

CFIm68

```
Homo            MADGVDHIDIYADVGEEFNQEAEYGGHDQIDLYDDVISPSANNGDAPEDR 50
Pan             MADGVDHIDIYADVGEEFNQEAEYGGHDQIDLYDDVISPSANNGDAPEDR 50
Bos             MADGVDHIDIYADVGEEFNQEAEYGGHDQIDLYDDVISPSANNGDAPEDR 50
Mus             MADGVDHIDIYADVGEEFNQEAEYGGHDQIDLYDDVISPSANNGDAPEDR 50
Gallus          MADGVDHIDIYADVGEEFNQEAEYGGHDQIDLYDDVISPSANNGDAPEDR 50
Python          MADGVDHIDIYADVGEEFNQEAEYGGHDQIDLYDDVISPSANNGDAPEDR 50
Xenopus         MADGVDHIDIYADVGEEFNQEAEYGAHEQIELYEDVISPSANNGDAPEDR 50
Latimeria       MADGVDQIDIYADVGEEFNQEAEYGGHEQIDLYDDVISPSANNGDAPEDR 50
Danio           MADGVDHIDIYADVEEEFNQESDYPVHDQIDLYDDVISPSANNGDAPEDR 50
Callorhinchus   MADGVDHIDIYADVGEEFNQESEYSGHDQIDLYDDVISPSANNGDAPEDR 50
                ******:******* ******::*  *:**:**:****************

Homo            DYMDTLPPTVGDDVGKG---AAPNVVYTYTGKRIALYIGNLTWWTTDEDL 97
Pan             DYMDTLPPTVGDDVGKG---AAPNVVYTYTGKRIALYIGNLTWWTTDEDL 97
Bos             DYMDTLPPTVGDDVGKG---AAPNVVYTYTGKRIALYIGNLTWWTTDEDL 97
Mus             DYMDTLPPTVGDDVGKG---AAPNVVYTYTGKRIALYIGNLTWWTTDEDL 97
Gallus          DYMDSLPPSVGDDVGKG---AAPNVVYTYTGKRIALYIGNLTWWTTDEDL 97
Python          DYMDSLPPSVGDDVGKG---SAPNVVYTYTGKRIALYIGNLTWWTTDEDL 97
Xenopus         DYMDNLATSVGDDVVKG---SVPNIVYTYTGKRIALYIGNLTWWTTDEDL 97
Latimeria       DYMDTLPPSVGDDVGKG---AAPNVVYTYTGKRVALYIGNLTWWTTDEDL 97
Danio           DYLDSLPAPGGNEGSKG---APANVVYTYNGKRIALYIGNLTWWTTDEDL 97
Callorhinchus   DYMDSLPQSVTDDGGKSSNSMGPNIVYTYTGKRIALYIGNLTWWTTDEDL 100
                **:*.*      ::  *.     *:****.***:***************

Homo            TEAVHSLGVNDILEIKFFENRANGQSKGFALVGVGSEASSKKLMDLLPKR 147
Pan             TEAVHSLGVNDILEIKFFENRANGQSKGFALVGVGSEASSKKLMDLLPKR 147
Bos             TEAVHSLGVNDILEIKFFENRANGQSKGFALVGVGSEASSKKLMDLLPKR 147
Mus             TEAVHSLGVNDILEIKFFENRANGQSKGFALVGVGSEASSKKLMDLLPKR 147
Gallus          TEAVHSLGVNDILEIKFFENRANGQSKGFALVGVGSEASSKKLMDLLPKR 147
Python          TEAVHSLGVNDILEIKFFENRANGQSKGFALVGVGSEASSKKLMDLLPKR 147
Xenopus         TDAVHSLGVNDILEIKFFENRANGQSKGFALICVSSESSSKKLMDLLPKR 147
Latimeria       IEIVHSLGVNDILEIKFFENRANGQSKGFALICVGSDMSSKKLMDLLPKR 147
Danio           TDAIRSIGINDVLEIKFFENRANGQSKGFALVCVGSDSSSRKLMDLLSKR 147
Callorhinchus   TEALHSMGISDVLEIKFFENRANGQSKGFALICVGSEASSKKLLDMLPKR 150
                 : ::*:*:.*:******************: *.*: **:**:*:* **
```

```
Homo            ELHGQNPVVTPCNKQFLSQFEMQSRKT-TQSGQMSGEGKAGPPGGSSRAA 196
Pan             ELHGQNPVVTPCNKQFLSQFEMQSRKT-TQSGQMSGEGKAGPPGGSSRAA 196
Bos             ELHGQNPVVTPCNKQFLSQFEMQSRKT-TQSGQMSGEGKAGPPGGSSRAA 196
Mus             ELHGQSPVVTPCNKQFLSQFEMQSRKT-TQSGQMSGEGKAGPPGGSSRAA 196
Gallus          ELHGQNPVVTPCNKQFLSQFEMQSRKT-TQSGQMSGEGKAGPPGGSSRAA 196
Python          ELHGQNPVVTPCNKQFLSQFELQSRKT-TQSGQMSGEGKAGPPGGSSRAT 196
Xenopus         EMHGQKPIVTPCNKQFLSQFEMQSRKTATQAGQMSGEGKAGPPGVNARLP 197
Latimeria       ELHGQNPVVTPCNKQFLSQFEMQSRKS-TQSGQMSGEGKAGPPGANSRAA 196
Danio           ELHGQNPIVTPCNKQSLSQFEMQSRKS-TQSGQMSGEGKAGPPGSGSRGG 196
Callorhinchus   DMHGQNPIVTPCNKQSLSQFEMQSRKS-TQSGQMSGEGKAGPPGASSRLA 199
                ::***.*:******* *****:****: **:************** .:*

Homo            FPQGGRGRGRFPGAVPGGDRFPGPAGPGGPPPP----------------F 230
Pan             FPQGGRGRGRFPGAVPGGDRFPGPAGPGGPPPP----------------F 230
Bos             FPQGGRGRGRFPGAVPGGDRFPGPAGPGGPPPP----------------F 230
Mus             FPQGGRGRGRFPGAVPGGDRFPGPAGPGGPPPP----------------F 230
Gallus          FPPSNRGRGRFPGAIPGGDRFPGPAGPGGPPPP----------------F 230
Python          FPPSNRGRGRFPGALPGGDRFPGPAGPGGPPPP----------------F 230
Xenopus         FPPGGRGRGRFPTSGPGGDGFPGPAGPGGPPPP----------------F 231
Latimeria       FPPGGRGRGRFPAAGPGGDRFPGPVGPGGPPPP----------------F 230
Danio           GFPPGKGQRRFPGPPGQGDRFPGPVGPGGPPPH----------------F 230
Callorhinchus   FSQSGRGRGRFPPNGPGGDRFPGPPGPGGPPPFPAPFTRGSVPERLIHHI 249
                    .:*: ***     ** **** *******                :

Homo            PAGQTPPRPPLGPPGPPGPPGPPPPGQVLPPPLAGPPNRGDRPPPPVLFP 280
Pan             PAGQTPPRPPLGPPGPPGPPGPPPPGQVLPPPLAGPPNRGDRPPPPVLFP 280
Bos             PAGQTPPRPPLGPPGPPGPPGPPPPGQVLPPPLAGPPNRGDRPPPPVLFP 280
Mus             PAGQTPPRPPLGPPGPPGPPGPPPPGQVLPPPLAGPPNRGDRPPPPVLFP 280
Gallus          PAGQTPPRPPLGPPGPPGPPGPPPPGQVLPPPLAGPPNRGDRPPPPVLFP 280
Python          PAGQTPPRPPLGPPGPPGPPGPPPPGQVLPPPLTGPPNRGDRPPPPVLFP 280
Xenopus         AAGMTPPRPPMCPPGPPGPPGPPPPGQPLPPPLPGPPNRGDRPPPPVMFP 281
Latimeria       P-GQNPPRPPLGPPGPPGPPGPPPPGQALPPPIAGPPNRGDRPPPPVLFP 279
Danio           PGMQGPPRLPSGPPGPLGPPGPPPPGQGLPPPLGGPPNRGDRPPPPVLFP 280
Callorhinchus   FSGQGPPRPPLGPPGPPGPPGPPPPGQGLPPPLGGPPNRGDRPPPPVMFP 299
                     *** *  **** ********** ****: **************:**

Homo            GQPFGQPPLGPLPPGPPPPVPGYGPPPGPPPPQQGPPPPPGPFPPRPPGP 330
Pan             GQPFGQPPLGPLPPGPPPPVPGYGPPPGPPPPQQGPPPPPGPFPPRPPGP 330
Bos             GQPFGQPPLGPLPPGPPPPVPGYGPPPGPPPPQQGPPPPPGPFPPRPPGP 330
Mus             GQPFGQPPLGPLPPGPPPPVPGYGPPPGPPPPQQGPPPPPGPFPPRPPGP 330
Gallus          GQPFGQPPLGPLPPGPPPPVPGYGPPPGPPPPQQGPPPPPGPFPPRPPGP 330
Python          GQPFGQPPLGPLPPGPPPPVPGYGPPPGPPPPQQGPPPPPGPFPPRPPGP 330
Xenopus         GQPYGQPPIVPLPPGPPPPV--YGPPPGPPPPQQGPPPPPGPFPPRPPG- 328
Latimeria       GQPFGQPPLGPLPPGPPPPVPGYGPPPGPPPPQQGPPPPPGPFPPRPPGP 329
Danio           GQ-FGQPPMGPMPPGPPP--PGYGPPPGPPPPQQGPP-PPGPFPPRPPGP 326
Callorhinchus   GQPFGQPPLGPLPPGPPPPVPGYGPPPGPPPPQQGPPPPPGPFPPRPPGP 349
                ** :****: *:******    *************** ***********

Homo            LGPPLTLAP-PPHLPGPPPGAPPPAPHVNPAFFPPPTNSGMPTSDSRGPP 379
Pan             LGPPLTLAP-PPHLPGPPPGAPPPAPHVNPAFFPPPTNSGMPTSDSRGPP 379
Bos             LGPPLTLAP-PPHLPGPPPGAPPPAPHVNPAFFPPPTNSGMPTSDSRGPP 379
Mus             LGPPLTLAP-PPHLPGPPPGAPPPAPHVNPAFFPPPTNSGMPTSDSRGPP 379
Gallus          LGPPLTLAP-PPHLPGPPPGAPPPAPHVNPAFFPPPANSGIPTSDSRGPP 379
Python          LGPPLTLAP-PPHLPGPPPGAPPPAPHVNPAFFPPPANSGIPTSDSRGPP 379
Xenopus         --PPMALGP-PPHLPGPPPGGPPPAPHVNPNFFPPPGNAGMPSSDSRGPP 375
Latimeria       LGPPLTLAP-PPHLPGPPPGAPPPAPHVNPAFFPPPANSGIPTSDSRGPP 378
Danio           LGPPLGLAP-PPHMQGPPPGGPPPAPHVNPAFFPPPGN-NMPSSDGRGPP 374
Callorhinchus   IGPPINLAPPPPHLPGPPPGAPPPAPHVNPAFFPPPANSVVPSSDSRGPP 399
                  **: *.* ***: *****.********* ***** *  :*:**.****
```

```
Homo            PTDPYGRPPPYDRGDYGPPGREMDTARTPLSEAEFEEIMNRNRAISSSAI 429
Pan             PTDPYGRPPPYDRGDYGPPGREMDTARTPLSEAEFEEIMNRNRAISSSAI 429
Bos             PTDPYGRPPPYDRGDYGPPGREMDTARTPLSEAEFEEIMNRNRAISSSAI 429
Mus             PTDPYGRPPPYDRGDYGPPGREMDTARTPLSEAEFEEIMNRNRAISSSAI 429
Gallus          PTDPYGRPPPYDRGDYGPPGREMDAARTPLSEAEFEEIMNRNRAISSSAI 429
Python          PSDPYGRPPPYDRGDYGPPGREIDAARTPLSEAEFEEIMNRNRAISSSAI 429
Xenopus         PSDPYGRPPPYERGDYGPPGRDMDVVRAPLSEAEFEEIMNRNRAISSSAI 425
Latimeria       PSDPYGRPPPYDRGDYGPSNREMEAARTPLSETEFEEIMNRNRAISSSAI 428
Danio           PGDPYGRPPPYDRDF--PGGRDMDASRTPLSEAEFEEIMNRNRAISSSAI 422
Callorhinchus   PSDPYGRPPPYDRGDYGPAGRDMDSSRTPLSEAEFEEIMNRNRAISSSAI 449
                * *********:*.   * .*:::  *:****:*****************


Homo            SRAVSDASAGDYGSAIETLVTAISLIKQSKVSADDRCKVLISSLQDCLHG 479
Pan             SRAVSDASAGDYGSAIETLVTAISLIKQSKVSADDRCKVLISSLQDCLHG 479
Bos             SRAVSDASAGDYGSAIETLVTAISLIKQSKVSADDRCKVLISSLQDCLHG 479
Mus             SRAVSDASAGDYGSAIETLVTAISLIKQSKVSADDRCKVLISSLQDCLHG 479
Gallus          SRAVSDASAGDYGSAIETLVTAISLIKQSKVSADDRCKVLISSLQDCLHG 479
Python          SRAVSDASAAEYGSAIETLVTAISLIKQSKVSADDRCKVLISSLQDCLHG 479
Xenopus         SRAVSDASAGDYGSAIETLVTAISLIKQSKVSADDRCKVLISSLQDCLHG 475
Latimeria       SRAVSDASAGDYGSAIETLVTAISLIKQSKVSADDRCKVLISSLQDCLHG 478
Danio           SRAVSDASAADYGSAIETLVTAISLIKQSKVSADDRCKVLISSLQDCLHG 472
Callorhinchus   SRAVSDASAGDYGSAIETLVTAISLIKQSKVSADDRCKVLISSLQDCLHG 499
                *********.:****************************************

Homo            IESKSYGSGSR-RERSRERDHSRSREKSRRHKS--RSRDRHDDYYRERSR 526
Pan             IESKSYGSGSR-RERSRERDHSRSREKSRRHKS--RSRDRHDDYYRERSR 526
Bos             IESKSYGSGSR-RERSRERDHSRSREKSRRHKS--RSRDRHDDYYRERSR 526
Mus             IESKSYGSGSR-RERSRERDHSRSREKSRRHKS--RSRDRHDDYYRERSR 526
Gallus          IESKSYGSGSR-RERSRERDHSRSREKSRRHKS--RSRDRHDDYYRERSR 526
Python          IESKSYGSGSRRRERSRERDHSRSREKSRRHKS--RSRDRHDDYYRERSR 527
Xenopus         IESKSYGSGSR-RERSRERDHSRSREKSRRHKS--RSRDRHDDYYRERSR 522
Latimeria       IESKSYGSGSRRRERSRERDHSRSREKSRRHKS--RSRDRHDDYYRERSR 526
Danio           IESKSYGSVAGRRERSRERDHSRSREKSRRHKS--RSRDRHEDYYRERSR 520
Callorhinchus   IESKSYGSGSRKRERSRERDHSRSREKSRRHKERSRSRDRHDDYYRERSR 549
                ******** :  ********************.  ******:********

Homo            ERERHRDRDRDRDRERDREREYRHR 551
Pan             ERERHRDRDRDRDRERDREREYRHR 551
Bos             ERERHRDRDRDRDRERDREREYRHR 551
Mus             ERERHRDRDRDRDRERDREREYRHR 551
Gallus          ERERHRDRDRDRDRERDREREYRHR 551
Python          ERERHRDRDRDRDRERDREREYRHR 552
Xenopus         ERERHRDRERDRDRERDREREYRHR 547
Latimeria       ERERHRDRERDRDRERDREREYRHR 551
Danio           ERDRHRERDRDRDRDREREREYRHR 545
Callorhinchus   ERERHRDRDRD--RERDREREYRHR 572
                **:***:*:**  *:*:********
```