RECOGNITION OF NONCANONICAL POLY(A) SITES IN HUMAN GENES

BY

ZHIQIAO WANG

DEPARTMENT OF BIOCHEMISTRY

UNIVERSITY OF OXFORD

& ST EDMUND HALL, OXFORD

THESIS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

TRINITY 2016

Recognition of noncanonical poly(A) sites in human genes

Zhiqiao Wang

Department of Biochemistry, University of Oxford & St Edmund Hall, Oxford

Submitted for the degree of Doctor of Philosophy, Trinity 2016

Abstract

Cleavage and polyadenylation are essential pre-mRNA maturation reactions for almost all eukaryotic mRNAs. Noncanonical poly(A) sites are frequently present at proximal sites during alternative cleavage and polyadenylation. These sites are involved in the regulation of gene expression. This project aims to elucidate the mechanism of noncanonical polyadenylation.

Two noncanonical poly(A) sites *JUND* and ARAF have been investigated. Analysis of the sequence requirement of the AGUAAA poly(A) site in JunD reveals that, besides the core sequence element, a U-rich element (adU-rich) is required for optimal usage and the distance between AGUAAA and this element is important; Similar to *MC1R*, the *ARAF* poly(A) site has a degenerated poly(A) signal and it has been suggested to require a distant auxiliary element, possibly G-rich enhancer, for optimal poly(A) site usage.

With regard to trans-acting factors, WDR33 appears to be important for the recognition of sub-optimal JunD poly(A) sites, whereas CPSF30 and CPSF160 preferentially recognise the canonical poly(A) site over a weak poly(A) site in JunD-MC4R 3' end formation. Contrary to the current models, depletion of CFIm68 does not favour proximal noncanonical poly(A) site usage in our JunD-MC4R Mini-gene system, but instead reduces the usage of the distal site and generates read through transcripts. Finally, we find that whilst WDR33 and CPSF30 affect poly(A) site selection in JunD-MC4R plasmids, they globally do not show general preference for either canonical or noncanonical sites.

In summary, we investigate the architecture of ARAF and JunD poly(A) sites. We conclude that in both cases the strength of the poly(A) site is determined by the complex architecture of the poly(A) sites rather than the core sequences. Recognising the wider architecture of poly(A) sites will be essential to fully understand what constitutes a strong poly(A) site and how this can be modulated, in particular in the context of alternative cleavage and polyadenylation.

Acknowledgements

I would like to thank my wife Na Wu, my father Jun Wang, my mother Jihong Chu, my father-in-law Wenyong Wu and my mother-in-law Baolan Zhang. Without your support, it would not have been possible for me to complete the DPhil project.

I would like to thank my supervisor Andre Furger, without whom, this project would not have come to fruition. Also, I would like to thank all the members in Furger Lab: Jonathan Neve, Chaitanya Vuppusetty, John Butterworth, Radhika Patel and Alastair Louey.

I would also like to thank our collaborators: Prof. Bin Tian, Prof. Chris Norbury, Jessica Hardy and Dr James Anderson.

Table of Contents

List of Figures

Abbreviations

3’ READS ___3’ region extraction and deep sequencing

A _____Adenine

adU-rich ____Adjacent uracil-rich

APA _____Alternative cleavage and polyadenylation

ATP _____Adenosine triphosphate

bp _____Base pairs

C _____Cytosine

CBC _____Cap binding protein complex

cDNA _____Complementary DNA

CE _____Capping enzyme

CE-RGt ___RNA guanylyltransferase

CE-RMt ___RNA (guanine-7)-methyltransferase

CE-RTp ___RNA 5’-triphosphatase

CFIm _____Mammalian cleavage factor I

CFIIm ____Mammalian cleavage factor II

CLIP _____Cross-linking immunoprecipitation

COX-2 ____Cyclo-oxygenase 2

CPSF _____Cleavage and polyadenylation specificity factor

CR _____Coding region

CRISPR ___Clustered regularly interspaced short palindromic repeats

CstF _____Cleavage stimulatory factor

CTD _____Carboxyl-terminal domain

DMEM ____Dulbecco's Modified Eagle's medium

DNA _____Deoxyribonucleic acid

DNase ____Deoxyribonuclease

dNTP _____Deoxynucleotide

DSE _____Downstream sequence element

dsRNA ____Double-stranded RNA

eIF _____Eukaryotic initiation factor

EST _____Expressed sequence tag

E.coli _____*Escherichia coli*

FBS _____Fetal bovine serum

Fig _____Figure

G _____Guanine

GFP _____Green fluorescent protein

GMP _____Guanosine monophosphate

GTF _____General transcription factor

GTP _____Guanosine-5'-triphosphate

h _____Hour(s)

HEK 293 ___Human Embryonic Kidney 293 cells

hFip1 _____Human factor interacting with poly(A) polymerase

HIV _____Human immunodeficiency virus

HSP _____Heat shock protein

HTLV _____Human T-cell leukemia virus

IgM _____Immunoglobulin heavy chain M

Inr _____Initiator element

K _____Keto

kb _____Kilo base pairs

kDa _____Kilo-Daltons

M _____Molar

MC1R _____Melanocortin 1 receptor

MC4R _____Melanocortin 4 receptor

MEME _____Multiple EM for Motif Elicitation

min _____Minute(s)

miRNA ____micro RNA

mRNA ____messenger RNA

mRNP ____messenger ribonucleoprotein

Mut _____Mutant

N _____Any nucleotide

NGS _____Next-generation sequencing

NMD _____Nonsense-mediated mRNA decay

nt _____Nucleotide(s)

OH _____Hydroxyl group

Oligo ____Oligonucleotides

ORF _____Open-reading frame

PABPN1 __Poly(A) binding protein found in the nucleus

PAP _____Poly(A) polymerase

PAS _____Poly(A) signal

PCR _____Polymerase chain reaction

Poly(A) ____Polyadenylation

Pre-mRNA __Premature messenger RNA

PTB _____Polypyrimidine tract binding protein

P-TEFb ____Positive transcription elongation factor b

QGRS _____Quadruplex forming G-rich Sequences Mapper

R _____Purine

RNA _____Ribonucleic acid

RNA Pol I __RNA polymerase I

RNA Pol II __RNA polymerase II

RNA Pol III ___RNA polymerase III

RNAi _____RNA interference

RNase _____Ribonuclease

RNA-seq _____RNA sequencing

RP _____RNase protection

RRM _____RNA recognition motif

rRNA _____Ribosomal RNA

RT-PCR _____Reverse transcriptase polymerase chain reaction

S _____Strong (C and G)

s _____Second(s)

SELEX _____Systematic evolution of ligands by exponential enrichment

Ser _____Serine

SF1/mBBP _____Branch point bridging protein

siRNA _____Small interfering RNA

snoRNA _____Small nucleolar RNA

SNP _____Single-nucleotide polymorphism

snRNA _____Small nuclear RNA

snRNP _____Small ribonucleoprotein

SS _____Splicing site

T _____Thymine

TFIID _____Transcription factor II D

tRNA _____transfer RNA

U _____Uracil

u _____Unit

U2AF _____U2 auxiliary factor protein

USE _____Upstream sequence element

UTP _____Uridine triphosphate

UTR _____Untranslated region

UV _____Ultraviolet

V _____Volt(s)

W _____Watt(s)

WDR33 _____Tryptophan-aspartic acid dipeptide repeat domain 33 subunit

WT _____Wildtype

Y _____Pyrimidine

ZF _____Zinc finger

ZK _____Zinc knuckle

Chapter 1

Introduction

Chapter 1

Transcription is an important process to generate RNA molecules based on the sequence of DNA. 3’ end formation is a fundamental processing step for the maturation of mRNAs in all eukaryotes. All protein-encoding primary transcripts are cleaved at their 3’ end and, with the exception of replication-dependent metazoan histone genes, subsequently subjected to polyadenylation. The cleavage and polyadenylation reactions are directed by a large multi-protein complex which in humans can constitute more than eighty proteins. In mammals, assembly of the 3’ end processing complex is initiated by the cooperative interaction of Cleavage and Polyadenylation Specific Factor (CPSF) and Cleavage Stimulation Factor (CstF) with specific core sequences on the pre-mRNA. This core poly(A) sequence constitutes a conserved sequence element which in 70% of human genes is defined by the so called “canonical” hexamers AAUAAA or AUUAAA. These hexameric motifs are recognised by CPSF. In addition to the hexamers, a functional poly(A) site further requires a somewhat less defined so called U or GU-rich region (DSE) which is positioned downstream of the cleavage site and is recognised by the CstF component of the cleavage and polyadenylation machinery. About 30% of human genes do not contain canonical poly(A) signals (Neve, Burger et al. 2016). Interestingly, if several poly(A) sites are present in a gene, as is the case for alternatively cleaved and polyadenylated transcripts, different transcripts will be generated. These different transcripts will not only result in different levels of the protein but also different types of the protein. In addition, a large proportion, around 70% of human genes undergo alternative cleavage and polyadenylation (APA)

(Hoque, Ji et al. 2012). Thus, APA is an important regulatory step during gene expression. Interestingly, the noncanonical poly(A) sites are used preferentially at the proximal site (Wang, Dowell et al. 2013). Unlike the canonical poly(A) sites, the recognition of human noncanonical 3' end processing sites by the cleavage and polyadenylation machinery is poorly understood due to the limited number of examples that have been analyzed at the biochemical level to date.

This project is aimed at the very heart of this question by elucidating the mechanisms and identifying the components that are required for the recognition of poly(A) sites of the AGUAAA type and poly(A) sites that lack any recognisable hexamer. If we understand how the cleavage and polyadenylation machinery is able to recognise these types of noncanonical sites, we will not only further our understanding of the fundamental process of cleavage and polyadenylation but also be able to unravel how noncanonical site recognition can contribute to APA and the regulation of gene expression in eukaryotes. Therefore, understanding how these sites are recognised is of profound importance. We began this investigation using the poly(A) site of individual gene examples *JunD* and then expanded our research to a transcriptome level.

1.1 Eukaryotic gene expression

1.1.1 Deoxynucleic acid and genes

The early experiment in 1940s indicated that deoxyribonucleic acid (DNA) was demonstrated to function as the reservoir to store the genetic information (Avery, MacLeod et al. 1944, Hershey and Chase 1952). Afterwards, James D. Watson and Francis Crick proposed the model of the structure of DNA (Fig 1) which is

the classical double stranded helix with nucleotide base paired based on the X-ray crystallography experiment (Watson and Crick 1953). This model lays the foundation of the central dogma of molecular biology, which explains the flow of copying genetic information from DNA to RNA to protein (Crick 1970).

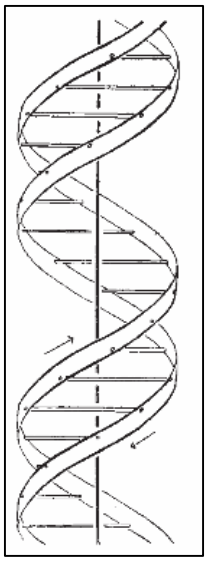


Figure1 Diagram depicting the double stranded DNA structure (Watson and Crick 1953)

The double stranded DNA structure predicted by James D. Watson and Francis Crick (Watson and Crick 1953). Two DNA chains are indicated by two ribbons. Horizontal represents the hydrogen bonds holding two chains together. The vertical line is the fibre axis.

1.1.2 The process of eukaryotic gene expression

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. These products can be proteins, but in non-protein coding genes such as transfer RNA (tRNA) or small nuclear RNA (snRNA) genes, where the product is a functional RNA. The first step of the gene expression is transcription (Fig 2). The genetic messages are transcribed

by RNA polymerase. Unlike prokaryotic cells that use one polymerase, eukaryotes have three different types of polymerases (RNA Pol I, II and III) to transcribe different classes of genes. All three RNA polymerases contain the conservative core part and some type-specific subunits (Archambault and Friesen 1993). RNA Pol I transcribes most of the ribosomal genes (28s, 18s and 5.8s ribosomal RNAs (rRNAs)). RNA Pol III transcribes 5s rRNA gene, tRNA genes, U6 snRNA and some small nucleolar RNAs (snoRNAs) (Paule and White 2000, Orphanides and Reinberg 2002). The messenger RNAs (mRNAs) and some small nuclear RNAs are transcribed by RNA Pol II. RNA Pol II transcribes all protein-encoding genes. These genes are transcribed as precursor RNAs and undergo several steps to become mature products. The initiation of the transcription is a complicated process. Unlike the RNA polymerase of the bacteria that can recognise the promoters in the gene, eukaryotic polymerases depend on the binding of general transcription factors (GTFs) to promoters along the DNA sequence or formation of a large transcription pre-initiation complex to activate transcription (Orphanides, Lagrange et al. 1996, Woychik 2000, Woychik and Hampsey 2002, Reese 2003, Shilatifard, Conaway et al. 2003). For example, the TATA box found in the promoter region with the core sequence “5’-TATAAA-3’” is usually recognised by transcription factor II D (TFIID). Then the recruitment of several other transcription factors takes place to form a protein complex to further recruit RNA Pol II to start the transcription (Yang, Bolotin et al. 2007). Similarly to the function of the TATA box, the Initiator elements (Inrs) can be recognised by components of TFIID (Xi, Yu et al. 2007). TATA box and Inrs are usually mutually exclusive. Followed by the transcription initiation, the elongation

process starts by the recruitment of the elongation factor. The nascent transcripts are processed into mature mRNA through three key reactions that occur mostly co-transcriptionally: 5’ capping, splicing and 3’ end formation.

The mature protein-encoding mRNAs are then exported to cytoplasm to create a chain of amino acids in a process called translation (Fig 2). The nuclear export of mRNA is a highly regulated process and involves three major steps with extensive mechanisms coupling each other (Köhler and Hurt 2007). Initially, the mRNAs are loaded with proteins turning them into ribonucleoprotein complexes (mRNPs), which are then able to “dock” onto the nuclear pore complexes in the nuclear envelope, then translocate through the pore and are finally released into cytoplasm (Carmody and Wente 2009). A mature mRNA has a cap, followed by a 5’ untranslated region (5’ UTR), the open reading frame (ORF), a 3’ untranslated region (3’ UTR) and a stretch of adenosines which constitute the poly(A) tail. Within the ORF resides the information to translate the genetic information into a protein product according to the rules of the genetic code (Lengyel, Speyer et al. 1961). The code defines the codon which consists of consecutive nucleotide triplets that specify which amino acids will be added during protein synthesis (Sonenberg 1994, Nirenberg 2004). The amino acid chains are extended at the ribosome in three steps: translation initiation, elongation and termination. During translation process, the ribosome translocates from start codon (AUG) to stop codon (UAA, UGA or UAG) in a sequence dependent manner. This means that the sequence of the mRNA is vital to the subsequent gene expression process. A single mutation in a gene may lead to the change of the amino acid constitution in the protein.

Whilst the flow of information is generally from DNA to RNA to protein, in cases such as retroviruses and retrotransposons, RNA can be reversely transcribed to DNA. These reverse transcribed DNA elements can then be transcribed into new RNA and subsequently translated (Kalendar, Vicient et al. 2004).

Almost all the steps in eukaryotic gene expression can be modulated and regulated. Gene regulation is vital to the life of a single organism through its control of cellular differentiation and versatility. In addition, regulation of gene expression plays a key role in evolution since controlling the timing, the location, and the amount of a particular protein or an individual RNA can have a profound effect on the phenotype and function of a single cell or a multicellular organism.

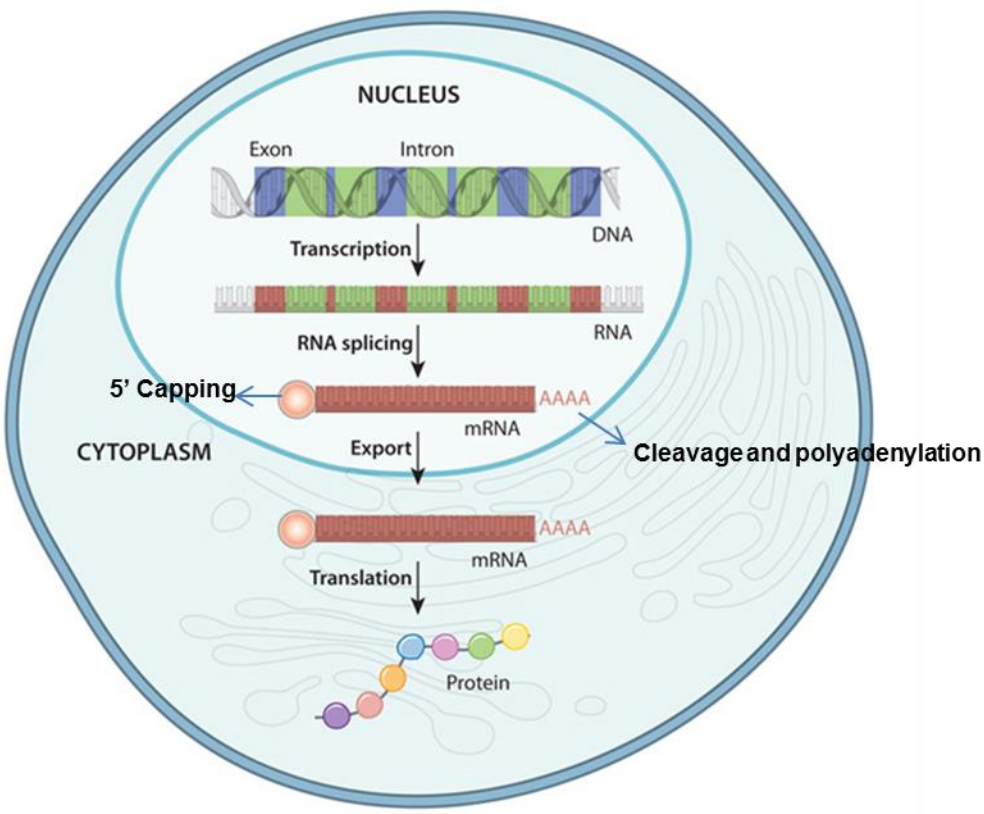


Figure 2 Overview of the gene expression flow in eukaryotes (modified from Nature education [W1])

The process of the canonical eukaryotic gene expression from a gene to a functional protein. This process contains several essential steps including transcription, pre-mature RNA 5’ capping, splicing, 3’ end formation (cleavage and polyadenylation), mRNA export and translation.

1.2 Pre-mRNA processing

A unique feature of eukaryotic gene expression is the extensive modification of the protein-encoding primary transcripts into mature export and translation competent mature mRNAs. These modifications include capping of the 5’ end, removing of non-coding intervening intron sequences in process, which called splicing, and the adding of a characteristic tail of adenosines at the end of each transcript.

1.2.1 5’ Capping

Capping is a biochemical reaction unique to RNA Pol II transcribed transcripts and is defined as the addition of a 5'(m7Gp)(ppN)[pN]n cap structure to the 5’ end (Shatkin 1976, Schoenberg and Maquat 2009). Before transcription initiation, the capping enzyme complex containing capping enzymes (CEs), RNA 5’-triphosphatase (CE-RTp), RNA guanylyltransferase (CE-RGt) and RNA (guanine-7)-methyltransferase (CE-RMt) binds to the RNA Pol II phosphorylated carboxy terminus domain (CTD) (McCracken, Fong et al. 1997). The capping process begins as soon as the 5’ end of the new transcript emerges (Cho, Takagi et al. 1997, Hirose and Manley 1998, Ho, Lehman et al. 1999, Fabrega, Shen et al. 2003) using a free GTP as a substrate. In the first

instance, the gamma (ɣ) phosphate of the 5’ terminal nucleotide is removed (by CE-RTp), leaving a bisphosphate group (i.e. 5'(ppN)[pN]n); subsequently CE-RGt catalyzes the fusion of guanylate (GMP) through an unusual 5’-5’ triphosphate linkage, producing 5'(Gp)(ppN)[pN]n. Then CE-RMt methylates the 7-nitrogen of guanine, using the demethylation of S-adenosyl-L-methionine to S-adenosyl-L-homocysteine as the source for the methyl group. This complex reaction results in the final cap structure: 5'(m7Gp)(ppN)[pN]n which is referred to as “cap-0” (Shuman and Schwer 1995, Proudfoot, Furger et al. 2002). Further modification can occur involving the methylation of the first and second nucleotides adjacent to the cap, which are typical to other types of RNAs, for example the snRNAs) (Fechter and Brownlee 2005, Meyer and Jaffrey 2014). Cap-1 has a methylated 2'-hydroxy group on the first ribose sugar and cap-2 has methylated 2'-hydroxy groups on the first two ribose sugars.

The cap-binding protein complex (CBC) which is composed of cap-binding protein 20 (CBP20) and cap-binding protein 80 (CBP80) then binds to the 5’ end of the capped RNA (Izaurralde, Lewis et al. 1994). The CBC can regulate the nuclear RNA export by its recognition with the nuclear pore complex (Lewis and Izaurralde 1997). Once the RNA-protein complex arrives in the cytoplasm, the CBC is replaced by the translation initiation factors eIF-4F (comprising eIF-4E and G) (Shatkin and Manley 2000). Interestingly, in the yeast model, nuclear export does not require the cap (Fresco and Buratowski 1996, Dower and Rossh 2002). Also, Picornaviruses can initiate translation via an internal ribosome entry segment instead of cap dependent mechanism (LÓPez-Lastra, Rivas et al. 2005). Apart from export and translation initiation, important additional functions are associated with capping. During early elongation in

transcription, both NELF and DSIF bind to RNA Pol II after the synthesis of a short stretch of pre-mRNA, and CTD Ser5 phosphorylation stimulates capping via association with Spt5, a subunit of DSIF (Chiba, Yamamoto et al. 2010). The promoter-proximal pausing of RNA Pol II induced by NELF and DSIF is associated with capping and is thought to allow time for efficient capping reaction (Adelman and Lis 2012, Laitem, Zaborowska et al. 2015). In an early study based on a Mini-gene, the cap structure was shown to enhance proximal intron excision (Ohno, Sakamoto et al. 1987). In mammals, the connection between CBC and alternative splicing through P-TEFb has been established (Lenasi, Peterlin et al. 2011). The cap also promotes the cleavage reaction (Hart, McDevitt et al. 1985).

Capping prevents 5’ degradation in two ways. First, degradation of the mRNA by 5’ exonucleases is prevented by appearing functionally similar to a 3’ end. Second, the access of decapping enzymes to the cap is inhibited by CBC and eIF-4E. This RNA 5’ end protection feature increases the half-life of the mRNA assuring enough time for eukaryotic RNA nuclear export and translation (Izaurralde, Lewis et al. 1994, Muhlrad, Decker et al. 1994). The decapping complex competes with eIF-4E binding to the cap. This means the 5’ cap is a marker of an actively translating mRNA (Parker and Sheth 2007). A capped and polyadenylated RNA can be circularized by eIF-4E/eIF-4G/Pab1p complex (Wells, Hillner et al. 1998). The translation process is further stimulated by capping through the interaction of ribosomal subunits with cap-binding factors (Lewis and Izaurralde 1997). .

1.2.2 Pre-mRNA splicing

The open reading frame (ORF) of >90% of eukaryotic genes are interrupted by the noncoding sequences， the so called introns. To generate mRNA with continuous ORFs for the subsequent translation process, the pre-mRNA needs to undergo a process called pre-mRNA splicing which is a two-step reaction that includes the removal of the introns and the subsequent ligation of the exons. This reaction relies on a large complex, the spliceosome, that is constituted of around 100 proteins and 5 snRNAs that make up the spliceosome. Components of the spliceosome interact with specific consensus sequences in the pre-mRNA to identify the positions where splicing occurs.

In higher eukaryotes, the consensus elements contain a donor site (5' end of the intron) 5'-AG^GURAGU-3' (^ marks the splice site), an acceptor site (3' end of the intron) 5'-YAG^RNNN-3' and a branch point (positioned upstream of the 3' splice site) 5'-CURAY-3' which are required for splicing (Proudfoot, Furger et al. 2002). The splice donor site includes a highly conserved GU at the 5' end of the intron. The splice acceptor site defines the 3' end of the intron with the conserved AG sequence. Many introns contain a pyrimidine rich region (C or U) between the donor site and the branch point. There is a highly conserved adenine nucleotide essential for the lariat formation. In summary, the consensus sequence for an intron is: 5'-AG^GURAGU (donor site) - intron sequence - CURAY (branch point) – Y rich region - YAG^RNNN-3' (acceptor site).

The splicing reaction is catalysed by the spliceosome, a large protein-RNA complex. It contains five small nuclear RNAs (snRNA U1, U2, U4, U5 and U6)

and a large number of proteins. The respective snRNA-protein complexes are called small ribonucleoproteins (snRNPs).

The assembly of the spliceosome involves a stepwise highly complex and dynamic interaction of the snRNPs with the pre-mRNA substrate. The first recognition of pre-mRNAs involves the snRNA U1 base pairing to the 5' splice site. The 3’ splice site is bound by $U2AF_{35}$, one of the subunits of U2 snRNP auxiliary factor (U2AF). Another subunit, $U2AF_{65}$, recognises the Y-rich region upstream of the acceptor site. Then U2 snRNP can be recruited to the branch region by U2AF with the help of branch point bridging protein (SF1/mBBP). U2 then base pairs to the branch point except the conserved adenosine, bulges out the branch adenosine specifying it as the nucleophile for the first transesterification reaction (Query, Moore et al. 1994). U5 snRNP, U6 snRNP and U4 snRNP are recruited as a “tri-snRNP” complex (Makarova, Makarov et al. 2001). After the displacement of U1 snRNP from the donor site, U6 base pairs with the 5’ end splice site. U6 subsequently interacts with U2 to bring the 5’ splice site and the conserved adenosine in the branch point to a close distance and to create a catalytic conformation that enables the first nucleophilic attack between the 2’OH of the BP adenosine and the very 5’ nucleotide of the intron (Madhani and Guthrie 1992, Sun and Manley 1995). The U5 snRNP interacts with sequences at the 5' and 3' exon via the invariant loop of U5. This can bring the two exons to a close distance to promote the second transesterification reaction (Newman, Teigelkamp et al. 1995).

The biochemical mechanism of spliceosome mediated splicing is a two-step transesterification reaction. First, the 2’OH of the conserved branch-point adenosine performs a nucleophilic attack on the phosphodiester bond of the 5’

splice site, forming the lariat structure intermediate. Second, the 3’OH of the released 5' exon performs another nucleophilic attack at the phosphodiester linkage at the 3' splice site, thus resulting in the releasing of intron lariat and ligation of two exons (Moore and Sharp 1993, Proudfoot, Furger et al. 2002, Fica, Tuttle et al. 2013).

The length of the intron is an important factor for how the correct splice sites are identified. In pre-mRNA with long introns and short exons, the 5’ and 3’ splice sites are identified by interactions of proteins across the exons (Fig 3) (Berget 1995). SR proteins bind to exonic splicing enhancers (ESE) to recruit U1 and U2AF and further recruit U2. This mode of recognition, which is called exon definition model, is believed to be the predominant form in vertebrate genes.

In the recognition of short introns protein-protein interactions spanning the intron are considered to be critical for identification of the splice sites. Intron recognition model is the default pathway in fission yeast genes with short introns (Fig 3) (Berget 1995). The binding of U1 to the upstream 5’ splice site and U2AF and U2 to the downstream polypyrimidine tract and branch site, respectively, of the same intron.

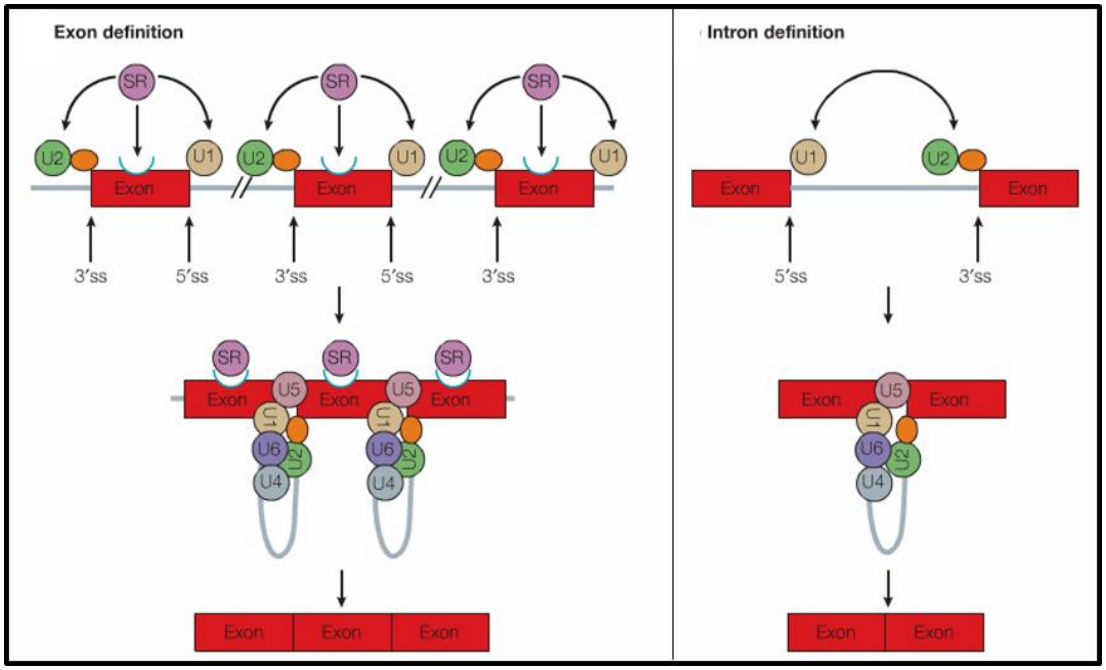


Figure 3 Exon definition and intron definition (Ast 2004)

Diagram depicting the procedure for exon definition splicing and intron definition splicing. SR proteins bind to exonic splicing enhancers to promote exon definition (Cartegni, Chew et al. 2002). SR proteins: purple; ESE: blue; U2AF: orange.

A single gene can create varieties of mRNA isoforms by changing the exon composition during a process called alternative splicing. Thus, different mRNAs can then be translated to distinct proteins. Alternative splicing can occur in many ways such as exon extension, exon skipping or intron retention. The extent of alternative splicing in vertebrates has been highlighted by the onset of high throughput mRNA sequencing technology. It is now estimated that approximately 95% of transcripts from mammalian genes undergo alternative splicing. Alternative splicing is an integral mechanism to establish tissue-specific transcriptomes (Pan, Shai et al. 2008, Wang, Sandberg et al. 2008) producing different mRNA isoforms and expression levels across tissues and cells (Eksi, Li et al. 2013, Li, Menon et al. 2014). Alternative splicing is controlled

by a large number of splice factors that can either promote or inhibit the recognition of specific splice sites by the spliceosome (Jain, Morgan et al. 2016). In addition to the tissue specific effect, the secondary structure of the pre-mRNA transcript also plays a role in regulating alternative splicing, by blocking the access of the binding site for a splicing factor or by narrowing the distance between two splicing elements (Warf and Berglund 2010, Lim, Ferraris et al. 2011). The complexity of the regulation of alternative splicing is considerable and includes chromatin modifications (Luco, Pan et al. 2010, Wang, Liu et al. 2014).

1.2.3 3’ end formation

3’ end formation is characterised by two distinctive catalytic steps: (i) the endonucleolytic cleavage of the pre-mRNA and (ii) the non-templated addition of an adenosine tail by the poly(A) polymerase (PAP) (Millevoi and Vagner 2010). Cleavage and polyadenylation is an essential process for the maturation of the nascent mRNA in all eukaryotes except replication-dependent metazoan histone genes (Proudfoot, Furger et al. 2002). The poly(A) tail added to an RNA during polyadenylation reaction can protect the mRNA molecule from RNase degradation and support transcription termination, export of the mRNA from the nucleus and translation (Guhaniyogi and Brewer 2001). The details of 3’ end formation will be expanded in the following sector.

1.2.4 Histone mRNA processing

Histone proteins are the major protein components of chromatin, which not only packs DNA but also controls gene expression and prevents DNA damage (Redon, Pilch et al. 2002). The 3' end formation process of the histone genes is unique. Stem-loop structures instead of a poly(A) tail are formed at 3' end of these mRNAs. Histone 3' end processing is achieved by a cleavage reaction only and assembly of the processing machinery on the pre-mRNA is dependent on a purine-rich sequence (usually named histone downstream element) and its association with the U7 snRNP (Marzluff, Gongidi et al. 2002).

The canonical histone pre-mRNAs contain the conserved stem loop sequence that binds stem loop binding protein (SLBP) followed by the histone downstream element (HDE), which base-pairs with U7 snRNA.

CPSF73, CPSF100, Symplekin, and possibly hFip1, which are the same protein involved in normal mRNA 3' end processing, together with some unknown factors form a histone cleavage complex to cleave pre-mRNA. CPSF73 is the endonuclease and possess the same cleavage function in normal 3' end processing complex. The cleavage reaction occurs five nucleotides downstream of the stem loop and upstream of the HDE. LSM11 is a component of U7 snRNP, interacting with ZFP100, to further bridge the SLBP-stem-loop complex and the histone cleavage complex (Marzluff, Wagner et al. 2008).

1.3 The general cleavage and polyadenylation signal

1.3.1 Discovery of the polyadenylation signal

In the early 1970s, several research groups found the presence of long stretches of adenosines after the digestion of mRNAs by pancreatic RNase and RNaseT1 (Lim and Canellakis 1970, Edmonds, Vaughan et al. 1971). Later, an enzyme that could potentially be responsible for the synthesis of these A-stretches, a poly(A) polymerase was identified and purified (Winters and Edmonds 1973, Winters and Edmonds 1973). It was the first protein to be characterised among the 3' end forming machinery. The AAUAAA sequence was subsequently discovered during early sequencing experiments by Nicholas Proudfoot (Proudfoot 2011). The development of recombinant DNA technology subsequently enabled the characterisation of the features of 3' end of mRNAs further. Using this technology, mutagenesis experiments revealed that the AAUAAA hexamer sequence is required for 3' end formation in a gene of the SV40 virus (Fitzgerald and Shenk 1981). Later, additional close variants such as AUUAAA/ UAUAAA that function like AAUAAA hexamer were identified (Zhao, Hyman et al. 1999). In addition to the hexamer, a less well defined GU rich region downstream of the cleavage site was also found to be essential for cleavage (McLauchlan, Gaffney et al. 1985, Gil and Proudfoot 1987). Additionally, so called auxiliary upstream elements (USEs) and the composition of the actual site of cleavage were also found to be important for 3' end processing (Carswell and Alwine 1989, DeZazzo, Kilpatrick et al. 1991). A statistical analysis based on EST sequences from yeast indicated that the USE-

hexamer-DSE constellation is highly conserved amongst eukaryotes (Graber, Cantor et al. 1999).

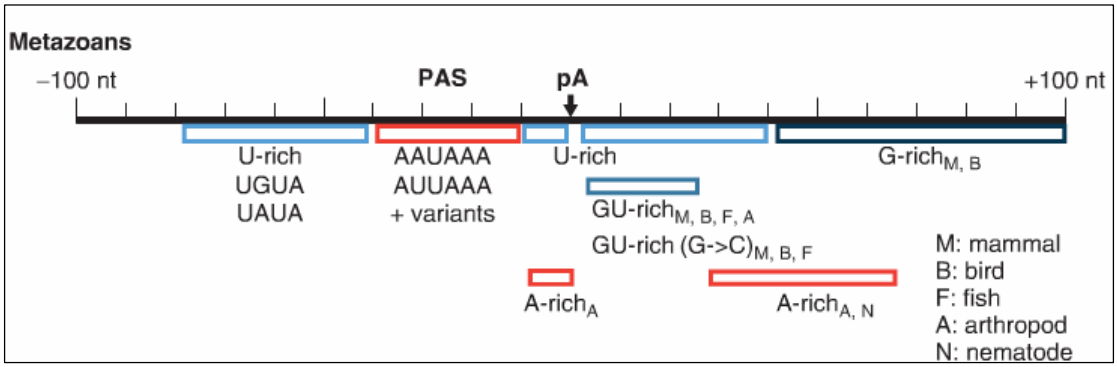


Figure 4 Overview of the poly(A) signals in metazoans (Tian and Graber 2012)
The Poly(A) signals specifically appear in mammals (M), birds (B), fish (F), arthropod (A) and nematode (N) are labelled. A-rich elements are red boxes. U-rich, G/U-rich and UAUA elements are blue boxes. G-rich elements are black boxes.

1.3.2 *Cis*-acting elements

The location of 3' end formation is determined by specific sequence elements, the *cis*-acting elements (*cis* elements) that are present on the pre-mRNAs.

i) Hexanucleotide sequence

Arguably, a key *cis* element is the conserved hexanucleotide sequences AAUAAA or AUUAAA (canonical hexamer). The global analysis shows that about 70% of all the poly(A) sites contain either of these two poly(A) signals (Neve, Burger et al. 2016). Furthermore, a study based on SV40 late RNA showed that the mutation of any of the nucleotide of the AAUAAA signal into any other three nucleotides could prevent the cleavage and poly(A) addition process except for the AUUAAA which retains 77% polyadenylation and 66% cleavage activity (Sheets, Ogg et al. 1990). Global transcriptome wide studies

also reveals that the hexamer is on average 20 nucleotides upstream of the cleavage site (Jun, Lutz et al. 2005).

ii) Downstream sequence element (DSE)

The less well defined element downstream (DSE) of the poly(A) site usually contains U-rich and GU-rich elements, with GU-rich elements generally positioned closer to the poly(A) site than U-rich elements (Jun, Lutz et al. 2005). The U-rich elements typically contain a short run of uracils. The classic forms of GU-rich elements are GUGU or UGUG and other functional variants like GUCU, CUGU, UCUG, and UGUC have also been identified (Salisbury, Hutchison et al. 2006, Hutchins, Murphy et al. 2008). Some poly(A) sites have been found to fully function without one or both of the U-rich or GU-rich motifs (Tian and Graber 2012). In contrast to the hexamers, generally the mutation or deletion of several nucleotides in the DSE are required to render a poly(A) site non-functional.

iii) Cleavage site

It is demonstrated that the relative position of the AAUAAA and the U-rich elements to one another define the approximate site where cleavage will occur (Chen, MacDonald et al. 1995). It is reported that CA and UA are globally the two most frequent dinucleotides at the cleavage site (Li and Du 2013). Mutagenesis of the SV40 poly(A) site revealed that the optimal nucleotide composition at the cleavage site is in the order of A>U>C>G. However, whilst most mRNAs indeed feature a “CA” dinucleotide at the cleavage site, the “CA”

is often irrelevant for the efficiency (Sheets, Ogg et al. 1990, Chen, MacDonald et al. 1995). In contrast, a prominent SNP occurring at a 1% frequency in the Caucasian population altering the dinucleotide of cleavage site of the prothrombin gene from “CG” to “CA”, increases the cleavage efficiency of the pre-mRNA significantly and this change has been linked to a higher tendency towards thrombophilia (Gehring, Frede et al. 2001).

iv) Auxiliary elements

Apart from the core sequence elements, many other auxiliary sequences were found to be critical for 3’ end formation in different mRNAs.

For example, the UGUAN element located upstream of the hexamer is recognised by CFIm (see below) and is important to direct the 3’ end formation in the absence of a canonical hexamer in a number of genes (Brown and Gilmartin 2003).

In many viral (adenovirus, adeno-associated virus and HIV-1) poly(A) sites and some mammalian mRNAs (for example C2 complement gene), an upstream U-rich region (USE) is present and critical for 3’ end processing of these genes (Valsamakis, Zeichner et al. 1991, Valsamakis, Schek et al. 1992, Brackenridge, Ashe et al. 1997, Moreira, Takagaki et al. 1998, Qiu, Nayak et al. 2004). There is no standard consensus for such USEs but they can be efficiently cross-linked to splicing factors such as the polypyrimidine tract binding protein (PTB). The presence of USEs can also increase the binding affinity of 64 kDa subunit of the cleavage stimulatory factor (CstF64) to the polyadenylation substrates (Moreira, Takagaki et al. 1998). In addition, *in vivo* polyadenylation assays demonstrate that cyclo-oxygenase 2 (COX-2) proximal polyadenylation signal

includes three “triple-U” USEs. Mutation in any of the three triple-Us (to GAU) substantially decreased usage of the poly(A) site (Hall-Pogar, Zhang et al. 2005). Another type of USE identified in the adenovirus L3 poly(A) site which is binding hFip1, one of the subunits of cleavage and polyadenylation specific factor (CPSF) and this interaction is critical for the activity of the poly(A) polymerase (Kaufmann, Martin et al. 2004). Similar poly(A) sites containing hFip1 binding sites have been identified in mammals where they appear to be critical for the regulation of alternative polyadenylation involving these sites (Lackford, Yao et al. 2014). An A-rich region upstream of the hexamer has been shown to be essential for 3’ end formation in the melanocortin receptor 4 (*MC4R*) gene and function like a hexamer when the original hexamer was deleted (Nunes, Li et al. 2010). In the melanocortin receptor 1 (*MC1R*) gene, which lacks a defined DSE, a G-Rich region 30 nucleotide downstream of the poly(A) site, was demonstrated as an essential enhancer. An additional G-rich region, 440 nucleotides downstream of the cleavage site, was further required for optimal usage of this poly(A) site (Dalziel, Nunes et al. 2006).

v) RNA secondary structures and cleavage and polyadenylation

The structural environment of the regions flanking or encompassing the poly(A) sites can contribute and facilitate the recognition of poly(A) sites. For example, a G-Quadruplex structure has been shown to be critical for the 3’ end formation of the human p53 pre-mRNA in stress induced cells. Several guanine tetrads downstream of the poly(A) site interact with hnRNP H/F which facilitates 3’ end processing of the *P53* pre-mRNA circumnavigating a global shut down of

normal 3’ end formation after DNA damage (Decorsière, Cayrel et al. 2011). In addition, several stem-loop structures have been found to be important for 3’ end formation and usually function by shortening the distance of two elements for their interaction (Ahmed, Gilmartin et al. 1991, Gilmartin, Fleming et al. 1992). The secondary structure of the USE in human the HIV-1 RNA and the human T-cell leukemia virus (HTLV) RNA rather than their actual nucleotide composition have been shown to create the space for CPSF to interact with poly(A) sites (Ahmed, Gilmartin et al. 1991, Graveley, Fleming et al. 1996, Graveley, Fleming et al. 1996).

vi) Noncanonical poly(A) signal

Apart from the canonical hexamers AAUAAA or AUUAAA, a number of hexamers so called noncanonical poly(A) signals have been identified in mammalian poly(A) sites. Recent whole transcriptome based analysis identified AGUAAA as the most frequent noncanonical hexamer which appears in 3.7% of mRNAs (Fig 5) (Neve, Burger et al. 2016). The noncanonical poly(A) sites are generally considered to be weak as their affinity to CPSF complex is compromised and weaker compared to the canonical variants (Sheets, Ogg et al. 1990). However, it is important to note that individual strong noncanonical poly(A) sites have been found by our group (Nunes, Li et al. 2010). Whether the recognition of noncanonical sites share the same mechanistic features and factors as those responsible to process canonical sites remains to be determined.

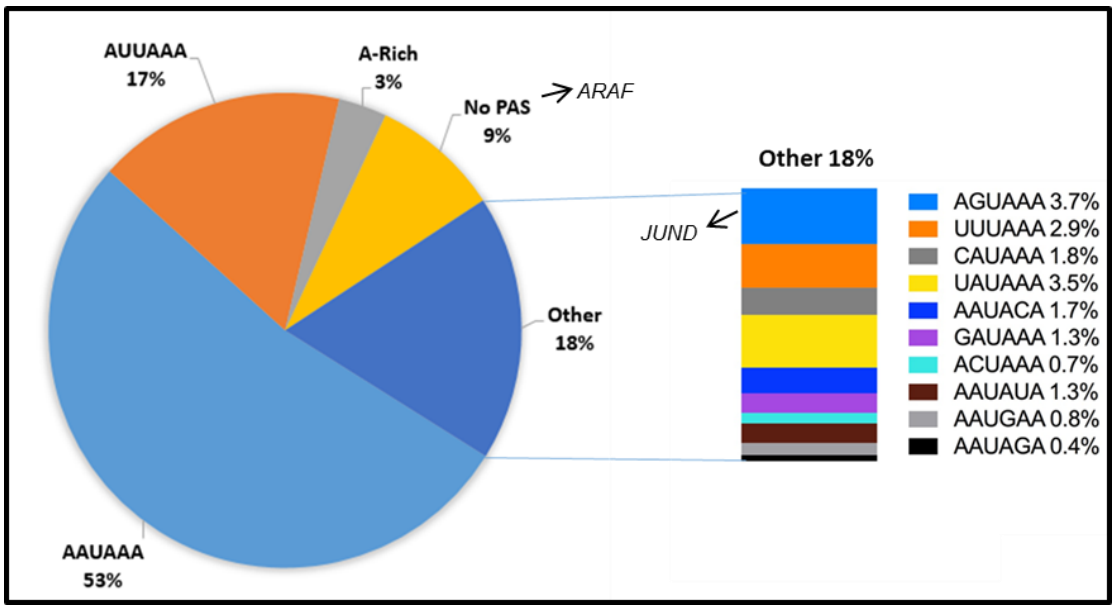


Figure 5 Distribution of canonical and noncanonical polyadenylation signals

Summary of the poly(A) signal frequencies independent of read number in the 0- to -40-nt region of all mapped cleavage and polyadenylation sites in nuclear fraction of HEK 293 cells by 3' READS sequencing method (Hoque, Ji et al. 2012, Neve, Burger et al. 2016). *ARAF* is in the "no PAS" group and *JunD* is in the "AGUAAA" group.

1.4 The importance of noncanonical polyadenylation signals

Global transcriptomics analysis showed that up to 30% of the human genes do not possess canonical poly(A) signal (Zarudnaya, Kolomiets et al. 2003, Zhang, Lee et al. 2005, Neve, Burger et al. 2016). Despite their relative high frequency, we currently have a poor understanding of how these noncanonical poly(A) sites are recognised. This is even more surprising, since noncanonical poly(A) sites have been strongly linked to regulated alternative cleavage and polyadenylation (APA), a process that affects around 70% of human genes

(Tian, Hu et al. 2005, Derti, Garrett-Engele et al. 2012, Miura, Shenker et al. 2013). Most strikingly, noncanonical poly(A) signals tend to be more frequently present in the proximal sites (Fig 6A) which during alternative polyadenylation are often somehow “activated” when shorter 3’ UTRs are generated such as observed when APA between embryonic skin stem cells and their corresponding daughter cell lineages are compared (Fig 6B) (Wang, Dowell et al. 2013). We are only just beginning to understand some of the underlying mechanisms, and a recent global RNA-protein interaction analysis suggests a role of CstF64/64tau in regulating usage of different types of poly(A) sites. In this study, depletion of CstF64/64tau, forced a switch from the proximal to distal poly(A) site usage in 241 genes. Interestingly, the affected proximal poly(A) sites showed an enrichment for noncanonical poly(A) sites and also appear to have a characteristic GUKKU (K=U or G) type DSE element (Hwang, Park et al. 2016).

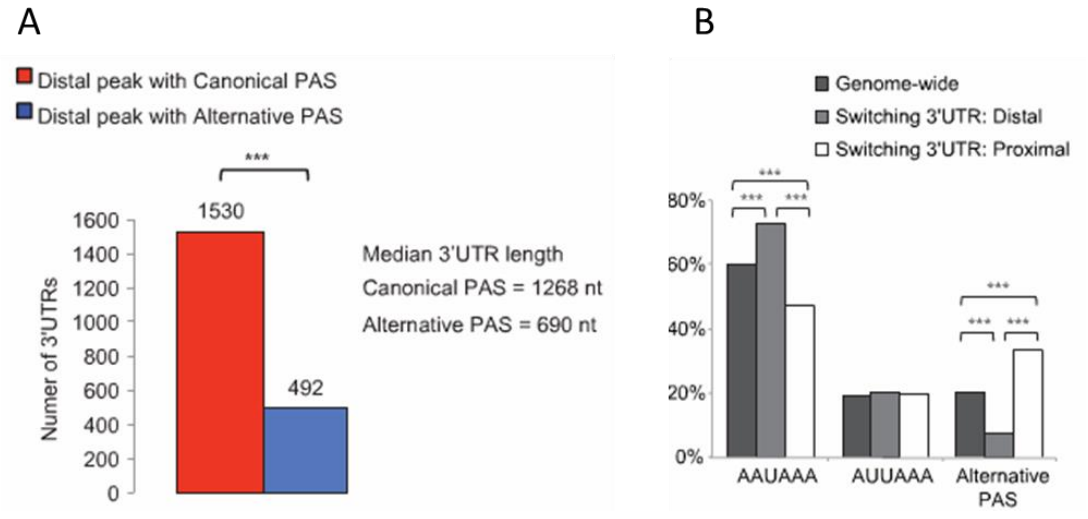


Figure 6 Preference for canonical and alternative (noncanonical) poly(A) signal by distal and proximal 3′ UTRs. (Wang, Dowell et al. 2013)

This analysis is based on an RNA sequencing experiment from mouse skin stem cells. (A) The number of distal 3′ ends that use canonical AAUAAA/AUUAAA hexamer (red)

is significantly larger than that of the distal 3′ ends that use noncanonical poly(A) signal (blue). (B) Differential 3’ UTR switching is identified based on the type of the hexamer signal. There is a significant proximal switching trend for noncanonical poly(A) site, (Alternative poly(A) signal (PAS), white bar). Increased preference for the noncanonical poly(A) signal was mainly balanced by the decreased usage of AAUAAA rather than AUUAAA.

1.5 The general cleavage and polyadenylation machinery

In human cells, there are more than eighty proteins that make up the cleavage and polyadenylation complex (Shi, Giammartino et al. 2009) of which around 20 are considered core factors (Mandel, Bai et al. 2008). In mammals, there are four protein subunits which are highly conserved. These include the cleavage and polyadenylation specificity factor (CPSF), the cleavage stimulation factor (CstF), mammalian cleavage factors I and II (CFIm and CFIIm) and poly(A) polymerase (PAP). The process of 3’ end formation is initiated by the recognition of the respective *cis* elements in the nascent transcripts by CPSF and CstF (Zarudnaya, Kolomiets et al. 2003). These RNA protein interactions are believed to be of a cooperative nature (Bienroth, Keller et al. 1993).

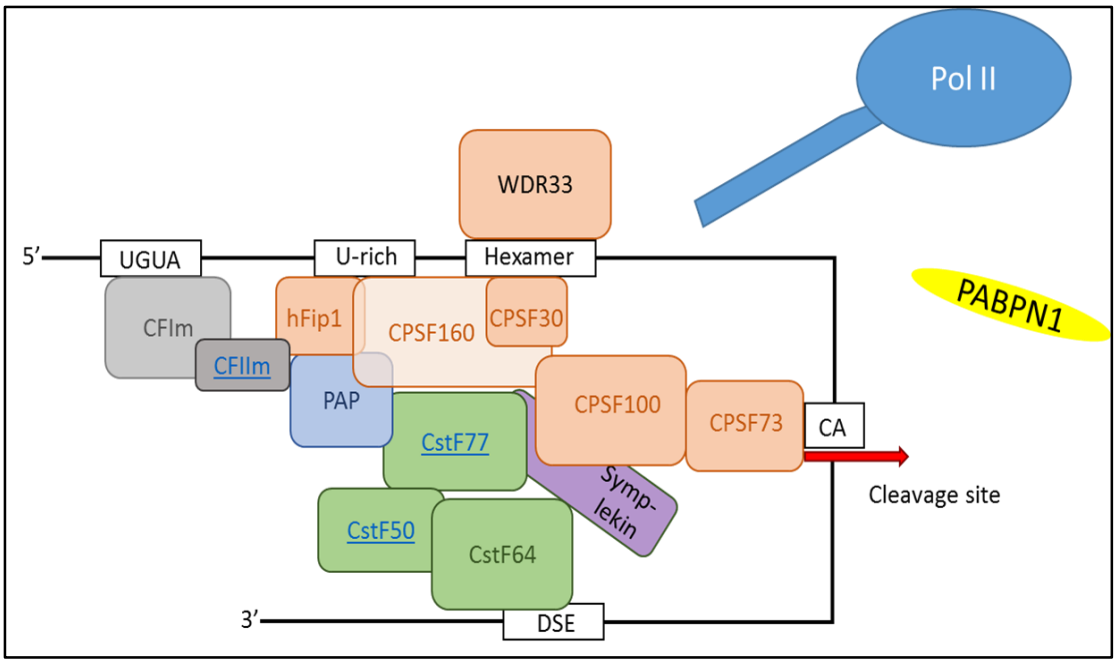


Figure 7 The 3' end processing complex in human

The black line is the pre-mRNA under 3' end processing. The UGUA, U-rich region, hexamer, CA (cleavage site) and DSE are the *cis* elements shown in the black box. The general protein machinery sub complexes are defined by different colours according to different sub-complexes.

1.5.1 Cleavage and polyadenylation specificity factor (CPSF)

CPSF specifically recognises the conserved hexamer signal AAUAAA and CPSF is essential for both the cleavage and the polyadenylation reaction (Murthy and Manley 1995). The CPSF protein complex features six protein subunits WDR33, CPSF30, CPSF160, hFip1, CPSF100 and CPSF73 (Shi, Giammartino et al. 2009), but four subunits: CPSF160, CPSF30, WDR33 and hFip1 proved to be sufficient to direct the polyadenylation *in vitro* (Schönemann, Kühn et al. 2014). The structure and roles of the CPSF components are as follows:

i) WD repeat domain 33 (WDR33) subunit:

WDR33 is a large 180 kDa protein that has recently been shown to directly interact with the AAUAAA hexamer (Chan, Huppertz et al. 2014). Interestingly, its homologue in yeast, the Pfs2p protein is an essential protein for 3' end formation (Ohnacker, Barabino et al. 2000). The mammalian WDR33 contains not only the conserved WD40 repeat domain but also various additional domains that are absent in Pfs2p (Chan, Choi et al. 2011, Xiang, Tong et al. 2014). The WD40 domains mediate protein-protein interaction (Stirnimann, Petsalaki et al. 2010) by forming β-propeller structures but as evidenced by recent proteomics based analysis can also participate in RNA interaction (Castello, Fischer et al. 2012, Kwon, Yi et al. 2013). As previously mentioned, the WDR33 subunit in reconstituted CPSF is required for polyadenylation *in vitro* (Schönemann, Kühn et al. 2014) and this crucial role is supported by a global protein-RNA interaction study utilizing PAR-CLIP that shows WDR33, *in vivo,* strongly and specifically interacts upstream of the sites of cleavage, peaking at -16 to -18 nt, which is the likely position of the hexamer in the transcript. The structural and biochemical principle of its interaction with the hexamer however needs to be further elucidated (Schönemann, Kühn et al. 2014).

ii) Cleavage and polyadenylation specificity factor 30 kDa (CPSF30) subunit:

Several *in vitro* studies show that CPSF30 binds directly to the hexamer (Chan, Huppertz et al. 2014, Schönemann, Kühn et al. 2014). The mammalian CPSF30 contains five zinc finger ZF domains and one zinc knuckle ZK domain.

It has been suggested that the CPSF30-RNA interaction is primarily mediated by its ZF2 and ZF3 domains and the binding of CPSF30 to the RNA is required for the entire CPSF complex to interact with RNA (Bienroth, Keller et al. 1993). In addition, CPSF30 has been shown to have a core role in poly(A) site selection in plants (Delaney, Xu et al. 2006, Thomas, Wu et al. 2012). Thus, the CPSF30-RNA interaction is likely to act as a key regulator role in mRNA 3' processing. Like WDR33, CPSF30 has been shown to directly associate with the AAUAAA hexamer and is required for mammalian 3' end processing (Chan, Huppertz et al. 2014).

iii) Cleavage and polyadenylation specificity factor 160 kDa (CPSF160) subunit:

Several early studies suggested that the major component that mediates the CPSF-AAUAAA interaction resides in the CPSF160 subunit. This was indeed supported by the fact that the 160 kDa protein can be specifically UV cross-linked to AAUAAA containing RNA substrates (Keller, Bienroth et al. 1991) and the Yhh1/Cft1p, which is the yeast homolog of CPSF160, has also been shown to be able to bind RNA (Dichtl, Blank et al. 2002); Furthermore, pull-down assays confirmed that recombinant CPSF160 does interact with AAUAAA containing RNAs (Murthy and Manley 1995). However, since it is now clear that WDR33 and CPSF30 are the key interacting partners between CPSF and AAUAAA, the precise role that the subunit CPSF160 plays during the complex RNA recognition phase needs to be re-evaluated. CPSF does play a critical role for the interaction between CPSF and CstF complex as it specifically interacts with the CstF77 subunit and this interaction may be integral to the

cooperative nature of poly(A) site recognition. CPSF160 also interacts with the poly(A) polymerase (Murthy and Manley 1995).

iv) Human factor interacting with poly(A) polymerase (hFip1):

This protein was discovered based on the sequence similarity to the yeast polyadenylation factor Fip1p. hFip1 interacts with PAP and can bind to the U-rich regions that are located either up- or downstream of the hexamer on pre-mRNA. The RNA-binding activity of hFip1 lies within the arginine-rich C-terminus of the protein (Kaufmann, Martin et al. 2004). hFip1 has been shown to play a critical role in alternative cleavage and polyadenylation during stem cell self-renewal (Lackford, Yao et al. 2014) .

v) Cleavage and polyadenylation specificity factor 100 kDa and 73 kDa subunit (CPSF100 and CPSF73):

The crystal structure study shows that both CPSF100 and CPSF73 contain metallo-β-lactamase and β-CASP domains. The active site of CPSF73 is dependent on a zinc ion located between these two domains. It has been demonstrated *in vitro* that CPSF73 possesses RNA endonuclease activity and that the disruption of the zinc binding activity abolishes this activity (Mandel, Kaneko et al. 2006). Thus, CPSF73 is considered to harbour the endonuclease activity that executes cleavage during 3' end formation. This is further supported by the finding that the conserved residues from the metallo-β-lactamase domains of both CPSF73 and CPSF100 are also required for the assembly of the endonuclease activity that executes cleavage at histone pre-mRNAs (Kolev, Yario et al. 2008).

1.5.2 Cleavage stimulation factor (CstF)

CstF is the component of the cleavage and polyadenylation machinery that specifically recognises the less well defined DSE region of poly(A) sites. It is a heterotrimeric protein complex containing the protein subunits CstF77, CstF64 and CstF50 (Takagaki, Manley et al. 1990, Gilmartin and Nevins 1991). The CstF complex is essential for the cleavage reaction but it is not required for the polyadenylation step (Gilmartin and Nevins 1989, Takagaki, Ryner et al. 1989).

i) Cleavage stimulation factor 64 kDa subunit (CstF64):

The interaction of CstF complex with pre-mRNA is achieved by the CstF64 subunit evidenced by SELEX experiments (Takagaki and Manley 1997). The structural analysis shows that the RNA-binding of CstF64 contains an RNA recognition motif (RRM) augmented by N- and C-terminal helices. CstF64 specifically recognises U/GU-rich regions in pre-mRNA approximately 30nt 3' to the poly(A) sites (Tian and Manley 2013). Consecutive "U"s are required for strong CstF-GU interactions between the protein and the substrate RNA. Interestingly, RNA binding also induces structural changes that may be critical to facilitate assembly of the larger complex (Cañadillas and Varani 2003). Recent global analysis suggests a role of CstF in modulating the selection of noncanonical poly(A) sites during alternative cleavage and polyadenylation (Yao, Biesinger et al. 2012, Hwang, Park et al. 2016).

ii) Cleavage stimulation factor 77 kDa subunit (CstF77):

CstF77 bridges the CstF64 and CstF50 subunits (Takagaki and Manley 1994) and thus plays a key role in the formation of the CstF complex. CstF77 appears

to be an elongated dimer suggesting that CstF may be a hexameric complex (Bai, Auperin et al. 2007). CstF77 also appears to be critical for the establishment of protein interactions between CstF and other complexes. Cross-linking experiments demonstrate a strong interaction between CPSF160 and CstF77, which may provide the basis for the observed CPSF-CstF cooperative RNA binding during the assembly process (Murthy and Manley 1995).

iii) Cleavage stimulation factor 50 kDa subunit (CstF50):

As for CstF77 it appears that CstF50 forms a homodimer (Moreno-Morcillo, Minvielle-Sébastia et al. 2011) providing further evidence for a hexameric nature of the functional CstF complex. CstF50 contains seven WD40 repeats, a motif that can mediate protein-protein interaction (Zhao, Hyman et al. 1999) and enables contacts between CstF77 and breast cancer 1 (BRCA1) associated RING domain 1 protein (BARD1). The latter interaction provides a link between mRNA 3' processing and DNA repair and tumour suppression (Kleiman and Manley 2001). CstF50 also interacts with the CTD of RNA Pol II and this interaction may be established via the N-terminal part of CstF50 (Fong and Bentley 2001).

1.5.3 Cleavage factor Im (CFIm)

CFIm contains four polypeptides CFIm25, CFIm68, CFIm59 and CFIm72. It has been suggested that different forms of CFIm complex exist *in vivo* including CFIm25-68, CFIm25-59 and possibly CFIm25-72. Evidenced by SELEX experiments (Brown and Gilmartin 2003) and site directed mutagenesis

(Venkataraman, Brown et al. 2005), the protein recognises the UGUAN motif in RNA. Functionally, addition of CFIm into in cleavage reaction *in vitro* stabilizes the binding of CPSF complex to the pre-mRNA (Rüegsegger, Beyer et al. 1996, Rüegsegger, Blank et al. 1998). CFIm has also been shown to play a critical role in the recognition of poly(A) sites that lack an AAUAAA or AUUAAA hexamer (Venkataraman, Brown et al. 2005).

Several individual groups showed that the knockdown of CFIm enhances the usage of the proximal poly(A) site either in individual gene examples or at a global scale (Kubo, Wada et al. 2006, Gruber, Martin et al. 2012, Martin, Gruber et al. 2012, Masamha, Xia et al. 2014, Hwang, Park et al. 2016). These findings suggest an important role for CFIm in alternative poly(A) sites selection. Interestingly, the crystal structure of CFIm25-CFIm68 complex shows that two CFIm68 monomers wrap around opposite sides of the CFIm25 dimer. These two CFIm68s interact with CFIm25 dimers simultaneously by their RNA recognition domains (RRM). These two RRMs are reported to enhance RNA binding and RNA looping which suggests a possible mechanism how this complex may regulate alternative polyadenylation (Yang, Coseno et al. 2011).

1.5.4 Cleavage factor IIm (CFIIm)

The CFIIm complex is the least well characterised core component of the cleavage and polyadenylation machinery and to date is only partly purified. It is required for the cleavage step only and mammalian CFII can be separated into two fractions: CFIIAm and CFIIBm (de Vries, Rüegsegger et al. 2000). CFIIAm fraction contains the essential components for the cleavage reaction while

CFIIBm is not required but has a stimulatory role for cleavage. hClp1, one of the 15 polypeptides associated with CFIIAm, has been discovered to interact with CFIm and CPSF by immunoprecipitation experiments, suggesting that it bridges these two during the cleavage process (de Vries, Rüegsegger et al. 2000). An additional component of CFIIm has been proved to be hPcf11 (protein 1 of human cleavage factor I), which has been shown to be critical in controlling RNA Pol II transcription termination by mediating the degradation of the polymerase associated 3' RNA after cleavage at the poly(A) site product (West and Proudfoot 2008). Human Pcf11 has also been implicated in alternative cleavage and polyadenylation. It binds directly to the pre-mRNA and enhances poly(A) site shortening (Li, You et al. 2015).

1.5.5 Poly(A) polymerase (PAP)

The monomer poly(A) polymerase is responsible for the addition of the 3' polyadenosine tail to a newly synthesized pre-mRNA molecule (Raabe, Bollum et al. 1991). This enzyme catalyzes the chemical reaction ATP + RNA-3'OH -- pyrophosphate + RNApA-3'OH. The two substrates of this enzyme are ATP and RNA (Wahle 1991, Wahle, Martin et al. 1991). PAP is recruited to the 3' end processing complex by CPSF through CPSF160 and hFip1 (Murthy and Manley 1995, Kaufmann, Martin et al. 2004). PAP is phosphorylated by mitosis-promoting factor, a key regulator of the cell cycle. High phosphorylation levels decrease PAP activity (Colgan, Murthy et al. 1996). There are several human PAP isoforms which are separated into canonical PAPs and noncanonical

PAPs. Most RNAs are polyadenylated by canonical PAPs. Noncanonical PAPs, amongst other functions, are associated with mitochondrial mRNA adenylation (hmtPAP), cytoplasmic polyadenylation (hGld2) and miRNA biogenesis. A prominent noncanonical PAP called Star-PAP, is controlling the polyadenylation of a subset of mRNAs that encode for proteins that are associated with DNA damage induced apoptosis and stress responses (Laishram 2014).

1.5.6 Poly-A Binding Protein found in the nucleus (PABPN1)

PABPN1 has been shown to contain a single RNA recognition motif and an arginine rich carboxy terminal domain (CTD) (Deo, Bonanno et al. 1999). The rate at which PAP adds adenine nucleotides is dependent on the presence of PABPN1. The first few nucleotides added by PAP are added very slowly, but the short polyadenosine tail is then bound by PABPN1, which accelerates the rate of adenine addition by PAP. Whilst PAP adds 200-250 adenine nucleotides to the 3' end of the mRNAs (Kühn, Gündel et al. 2009), the median length of cellular mRNAs however is much shorter and is between 50 to 100 nucleotides long (Chang, Lim et al. 2014). In addition, PABPN1 is implicated in controlling alternative cleavage and polyadenylation (Jenal, Elkon et al. 2012, Li, You et al. 2015).

1.5.7 Symplekin

The Symplekin protein forms a high-molecular weight complex with components of the 3’ end processing machinery including CPSF and CstF complex. It is suggested to serve as a scaffold role for recruiting other factors to the cleavage and polyadenylation complex (Takagaki and Manley 2000). It also participates in 3' end maturation of histone mRNAs, which do not undergo polyadenylation. Symplekin has been found to form a complex with heat shock transcription factor (HSF1) after stress treatment. This complex may be critical for the recruitment of the 3’ end processing machinery to the heat shock protein (HSB) pre-mRNA during stress and so safeguard its expression (Xing, Mayhew et al. 2004).

1.5.8 RNA polymerase II carboxy terminal domain (RNA Pol II CTD, RPB1)

The CTD of large subunit of RNA Pol II (RPB1) is necessary for the efficient cleavage reaction and serves as the key platform that interconnects pre-mRNA processing to transcription (Hirose and Manley 1998). The human RNA Pol II CTD features 21 consensus YSPTSPS and 31 non-consensus heptad repeats (Zaborowska, Egloff et al. 2016). This characteristic nature organises and presents a dynamic interaction hub for a variety of pre-mRNA processing protein factors including cleavage and polyadenylation factors such as Pcf11, CstF77 and CstF50.

Table1 Proteins have been shown to be involved in 3' end formation.

Protein Name	Accession No.	Motifs on Functions	Relevant RNA region	Ref
CPSF complex				
CPSF160	NP_037423	SFT1		(Keller, Bienroth et al. 1991, Murthy and Manley 1995)
CPSF30	NP_006684	Zinc finger	AAUAAA	(Chan, Huppertz et al. 2014)
WDR33	NP_060853	WD repeats	AAUAAA	(Chan, Huppertz et al. 2014, Schönemann, Kühn et al. 2014)
hFip1	NP_112179	Arginine-rich RNA binding motif	U rich	(Kaufmann, Martin et al. 2004, Lackford, Yao et al. 2014)
CPSF73	NP_057291	metallo-β-lactamase domain and β-CASP domain	Cleavage site	(Dominski, Yang et al. 2005, Mandel, Kaneko et al. 2006, Kolev, Yario et al. 2008)

CPSF100	NP_059133.1	metallo-β-lactamase domain and β-CASP domain		(Dominski, Yang et al. 2005, Kolev, Yario et al. 2008)
CstF complex				
CstF64	NP_001316	RRM	U/GU rich	(Takagaki and Manley 1997, Cañadillas and Varani 2003)
CstF77	NP_001317	HAT		(Takagaki and Manley 1994, Bai, Auperin et al. 2007)
CstF50	NP_001315	WD repeats		(McLauchlan, Gaffney et al. 1985, Moreno-Morcillo, Minvielle-Sébastia et al. 2011)
CFIm complex				
CFIm72				(Brown and Gilmartin 2003)
CFIm68	NP_008938	RRM	UGUA	(Venkataraman , Brown et al. 2005)
CFIm59	NP_079087	RRM		(Brown and Gilmartin 2003)
CFIm25	NP_008937	RRM	UGUA	(Venkataraman , Brown et al. 2005)
CFIIm complex				

CFIIAm				(de Vries, Rüegsegger et al. 2000)
CFIIBm				(de Vries, Rüegsegger et al. 2000)
hClp1	NP_001136069.1			(de Vries, Rüegsegger et al. 2000)
hPCF11	NP_056969.2	CID, Zinc finger		(Li, You et al. 2015, Yang, Hsu et al. 2016)
Symplekin	NP_004810			(Takagaki and Manley 2000)
PAP	NP_075045	RRM	Poly(A)	(Raabe, Bollum et al. 1991, Wahle 1991)
PABPN1	NP_004634	RRM	Poly(A)	(Deo, Bonanno et al. 1999)
RNA Pol II associated factors				
POLR2A	NP_000928.1	Polymerase		(Hirose and Manley 1998, Lian, Karpikov et al. 2008)
Splicing factors				
SF3B1	NP_036565.2			(Kyburz, Friedlein et al. 2006)
SF3B3	NP_036558.3			(Kyburz, Friedlein et al. 2006)

SF3B4	NP_005841.1			(Kyburz, Friedlein et al. 2006)
U2AF$_{65}$	NP_009210.1	RRM	Polypyrimidine	(Millevoi, Loulergue et al. 2006, Danckwardt, Kaufmann et al. 2007)
U2AF$_{35}$	NP_006749.1	RRM	USE	(Danckwardt, Kaufmann et al. 2007)
Transcription factors				
TFIID	NP_003185.1		Inr and E-box	(Dantonel, Murthy et al. 1997)
PC4	NP_006704.2			(Calvo and Manley 2001)

1.6 Mechanism of 3' end formation

The mechanism of 3' end formation has turned out to be more complex than previously anticipated. It is currently believed that the initiating step of the assembly of the 3' end processing complex is the coordinated recognition of the hexamer signal AAUAAA and the DSE on the nascent RNA by CPSF and CstF respectively. This initial recognition is likely to be assisted by additional factors including CFIm and probably also CFIIm. CPSF constitutes the core processing complex required for both the cleavage and the subsequent polyadenylation reaction. It binds the RNA directly by the association of WDR33

and CPSF30 with the hexamer. However, whether these two subunits bind to pre-mRNAs together or in a time dependent or poly(A) site type dependent manner remains unknown. Interestingly, earlier studies suggest that the size of the CPSF complexes can vary and range from 350 kDa to 500 kDa (Bienroth, Wahle et al. 1991, Murthy and Manley 1992). A possible explanation for this observation may be that different types of CPSF complexes are assembled to accommodate processing of a variety of transcripts that differ in their *cis* elements. Unlike CPSF, the CstF complex is only essential for the cleavage reaction and associates with the RNA by binding via the RNA recognition motif that are present in CstF64 to di-uridine pockets in the U/GU rich DSE. The interaction between CstF64 and the DSE is enhanced by CstF77 (Colgan and Manley 1997) (Fig 7).

With CPSF and CstF complexes anchored to the poly(A) site, additional factors are recruited to assemble the functional 3' end processing machinery. This machinery can subsequently catalyse a two-step reaction which entails endonucleolytic cleavage at the cleavage site and the polyadenylation of the 3' half of the cleaved pre-mRNA. The cleavage occurs between the hexamer and the DSE, generally within a window of 20 nucleotides, either side of the core sequences.

The speed of assembly of a functional cleavage and polyadenylation complex is dependent on the strength of the poly(A) site but is generally achieved within about 10 seconds (Chao, Jamil et al. 1999). However, the absolute strength of a given poly(A) signal is complicated to determine. It depends on the sequence architecture of the poly(A) site (see above) and relative availability of core and auxiliary elements.

The poly(A) polymerase is recruited by CPSF complex and is responsible for the addition of the 3' polyadenosine tail. The length of the poly(A) tail is measured by PABPN1. Once the tail length reaches threshold (usually 250-300 nucleotide), PABPN1 will inhibit the CPSF-PAP interaction.

1.7 Pre-mRNA processing and transcription are interconnected

To date, it is widely accepted that 5’ capping, splicing and 3’ end processing occur co-transcriptionally in a highly regulated manner (Cook 1999, Maniatis and Reed 2002, Proudfoot, Furger et al. 2002). TFIID complex bound to the core promoter has been shown to recruit CPSF (Dantonel, Murthy et al. 1997). This highlights that a link between transcription initiation and elongation by RNA Pol II and processing of the 3’ end of mRNA is established at an early step during gene activation. Additional evidence for co-transcriptional pre-mRNA 3’ end processing is revealed by the interaction of CstF complex with the transcriptional co-activator PC4 (Calvo and Manley 2001). Furthermore, cleavage/polyadenylation and transcription termination are tightly linked. Mutagenesis based approaches show that the mutation of the AAUAAA hexamer not only abolishes 3’ end formation but also compromises transcription termination (Connelly and Manley 1988).

1.7.1 The 3’ end formation step connects with 5’ capping

The connection between capping and 3’ end processing was discovered in the 1980s. The level of cleavage of uncapped compared to capped RNA templates

was significantly reduced in *in vitro* cleavage reactions. In addition, cleavage of capped RNA templates was inhibited if the reaction mix was supplemented with high concentration of m7GpppG (10µM). In contrast, supplementing the reaction mix with a non-methylated cap had no effect on cleavage (Hart, McDevitt et al. 1985). Another *in vitro* 3' end processing reaction suggests that the interaction between the cap-binding complex (CBC) and splicing factors could inhibit cleavage and polyadenylation (Cooke and Alwine 1996).

1.7.2 The link between 3' end formation and splicing

Early studies based on viral genes revealed that the recognition of the poly(A) sites in introns are somehow inhibited by the presence of the U1snRNP interacting with a downstream 5' SS (Levitt, Briggs et al. 1989, Zhao, Hyman et al. 1999). A detailed analysis of this concept in the bovine papilloma virus showed that the 70 kDa protein component of U1snRNP inhibits the activity of PAP (Furth, Choe et al. 1994, Gunderson, Polycarpou-Schwarz et al. 1998). Furthermore, U1snRNP mediated inhibition of upstream poly(A) sites proved instrumental for the expression of genomic RNA and mRNA in HIV-1 (Ashe, Griffin et al. 1995, Ashe, Furger et al. 2000, Fortes, Cuevas et al. 2003). The U1 snRNP mediated inhibition of intronic poly(A) sites is not restricted to viruses and it has later been found to be an essential process that prevents the activation of the numerous cryptic poly(A) sites located in introns (Kaida, Berg et al. 2010).

However, splicing components can also have a stimulatory effect on 3' end formation. The positive influence of splicing factors on 3' end formation was

reported in the 1990s and since then has been widely documented. The earliest experiments showed that the recognition of the 3' SS of the terminal intron significantly promotes the usage of the downstream poly(A) site (Niwa and Berget 1991, Niwa, MacDonald et al. 1992, Dye and Proudfoot 1999). This stimulation is achieved by the interaction between components of the splicing machinery and poly(A) factors most notably between the 65 kDa subunit of U2AF and PAP. $U2AF_{65}$ has also been found to interact and recruit CFIm59 to the poly(A) site (Vagner, Vagner et al. 2000, Millevoi, Geraghty et al. 2002, Millevoi, Loulergue et al. 2006). In addition, physical interaction of CPSF with several subunits of U2snRNP including SF3b155, SF3b130 and SF3b49 have been found. Mutations of U2 snRNA binding sites significantly reduces the cleavage efficiency in reporter genes (Kyburz, Friedlein et al. 2006). The stimulatory effect is mutual as mutations of the hexamer can inhibit terminal exon removal (Niwa, Rose et al. 1990) and the depletion of CPSF100 has a negative impact on splicing in a coupled splicing and 3' end cleavage assay (Kyburz, Friedlein et al. 2006). Thus the CPSF and U2snRNP interaction contributes to the coupling of splicing and 3' end formation (Kyburz, Friedlein et al. 2006).

1.8 The role of alternative cleavage and polyadenylation (APA)

In higher eukaryotes, many genes can generate different isoforms of mRNAs. This is achieved not only by alternative splicing events which are key to drive protein complexity and diversity, but also by alternative cleavage and

polyadenylation (Black 2003, Johnson, Castle et al. 2003, Wang, Sandberg et al. 2008, Chen and Manley 2009).

APA events can broadly be divided into two distinct classes. 1) UTR-APA are isoforms that are generated by the use of different poly(A) sites that are located downstream of the terminal exon. The resulting APA mRNA isoforms code for identical proteins but the length of their UTRs differs. 2) Coding region APA (CR-APA) summarises events where alternative poly(A) sites are located in upstream introns or exons and their activation, or skipping produces APA mRNA isoforms that can encode different proteins or produces transcripts that are funnelled into nonsense mediated decay (NMD). Depending on the localization of the terminal exon, APA events involving poly(A) sites in introns can be further classified into two subtypes: composite terminal exons and skipped terminal exons (Fig 8) (Tian and Manley 2013).

When APA was discovered, it was suggested that poly(A) site selection was random and was the result of failure for the cleavage and polyadenylation machinery to recognise the proximal site, and then to capture a distal site instead (Denome and Cole 1988). In essence, APA was considered a fail-safe feature. However, later many examples were found where APA was regulated and had a physiological impact (Sandberg, Neilson et al. 2008, Ji, Lee et al. 2009, Mayr and Bartel 2009). Whilst UTR-APA have identical coding capacity, the longer 3' UTRs potentially possess more regulatory elements such as miRNA binding sites which can regulate mRNA stability or translation. A classic example of regulated CR-APA occurs during the differentiation from B cells into plasma cells. During this differentiation process, the immunoglobulin heavy chain M (IgM) switches from a membrane bound type to a secreted type. This

is achieved by an APA process where an intronic poly(A) site is activated resulting in transcripts that omit the membrane binding domains (Takagaki, Seipelt et al. 1996). This APA event is tightly linked to splicing, and the competition of splicing and 3' end formation is regulated by the different levels of CstF64 between B cells and plasma cells (Takagaki and Manley 1998). A similar competition event has also been observed in mRNA encoding the calcitonin gene related peptide due to the usage of the proximal poly(A) site (Zhao, Hyman et al. 1999). Many individual APA events have also been linked with different diseases. In these cases, pathogenesis is generally the results of mutations of *cis* elements in one specific gene that affect its 3' end formation. These alterations can result in either loss of function, such as in thalassemia and IPEX syndrome (Orkin, Cheng et al. 1985, Whitelaw and Proudfoot 1986, Bennett, Brunkow et al. 2001, Yasuda, Shabbeer et al. 2003), or gain of function, as exemplified in thrombophilia and mantle cell lymphoma (Wiestner, Tehrani et al. 2007).

The emergence of whole transcriptomics technology allowed the investigation of APA at a global scale. Since then, many studies have reported that eukaryotes generate a large number of APA isoforms with wide ranging coding and regulatory consequences. A large fraction of mammalian genes (approximately 70%) (Derti, Garrett-Engele et al. 2012, Hoque, Ji et al. 2012), and approximately 50% of genes in flies (Smibert, Miura et al. 2012), worms (Jan, Friedman et al. 2010) and zebra fish (Ulitsky, Shkumatava et al. 2012) have been reported to undergo APA. However, when the conservation of individual poly(A) sites between human and mouse are considered, only around 500 genes contain conserved tandem poly(A) sites. Consequently, it has been

suggested that alternative poly(A) sites are species-specific (Ara, Lopez et al. 2006). Another global analysis based on five species including human, rhesus, dog and rat has shown that only 0.6% of the poly(A) sites in human were found to have poly(A) sites within the equivalent 30 base region in all four other mammals (Derti, Garrett-Engele et al. 2012). Furthermore, the comprehensive global study reveals that APA have tissue specific characteristics. Blood cells and testis prefer the usage of the proximal poly(A) sites, while neuronal cells prefer distal ones and the 3' UTR lengthening is a characteristic and widespread phenomenon in neuronal tissues (Zhang, Lee et al. 2005, Lianoglou, Garg et al. 2013, Miura, Shenker et al. 2013). Recently, single cell studies have started to expose the heterogeneity of poly(A) site usage within a population of homogenous mouse stem cells (Velten, Anders et al. 2015). Individual cells have been found to differ significantly in their choice of poly(A) site usage and this variability seen has the potential to contribute to the functional cell-to-cell heterogeneity (Velten, Anders et al. 2015). In addition, several studies have observed shifts to proximal poly(A) sites upon carcinogenesis (Mayr and Bartel 2009, Singh, Alley et al. 2009, Fu, Sun et al. 2011, Lin, Li et al. 2012, Morris, Bos et al. 2012, Masamha, Xia et al. 2014, Xia, Donehower et al. 2014). It is thought that shifts to the proximal poly(A) sites in cancer functions to exclude instability elements residing in the alternative UTR and therefore avoid particular regulatory networks such as miRNA mediated destabilisation of specific transcripts. Nevertheless, it is important to note that in the cancer cells where general 3' UTR shortening is observed, there are still a significant number of genes in which 3' UTR lengthening is seen (Singh, Alley

et al. 2009, Fu, Sun et al. 2011, Lin, Li et al. 2012). This clearly highlights that APA patterns in transformed cells are highly complex.

How APA is regulated however is still poorly understood. The global studies often show a relationship between the 3' UTR length and the expression levels of several cleavage and polyadenylation factors (Ji, Lee et al. 2009). For example, the depletion of CFIm68 results in global 3' UTR shortening events (Kim, Yamamoto et al. 2010, Martin, Gruber et al. 2012, Masamha, Xia et al. 2014). Co-depletion of CstF64 and CstF64tau leads to APA shifts in a small number of genes primarily to the distal poly(A) site, which is thought to be reflective of the stronger *cis* elements surrounding distal poly(A) sites (Hwang, Park et al. 2016). hFip1 globally influences APA events by blocking the CstF binding site of the proximal poly(A) sites and causes APA lengthening (Lackford, Yao et al. 2014, Li, You et al. 2015). In addition to poly(A) factors, a number of other RNA binding proteins have been implicated in the regulation of APA often by competing with poly(A) factors for binding to the poly(A) signals and thereby inhibiting usage of these particular sites. For example, PTB competes with CstF64 inhibiting poly(A) sites (Millevoi and Vagner 2010). Other protein factors have been shown to interact with pre-mRNA surrounding the poly(A) site and act as an enhancer such as αCP and FUS (Ji, Wan et al. 2013, Masuda, Takeda et al. 2015) or inhibitor role such as Cirbp and Rbm3 (Liu, Hu et al. 2013).

Not only protein factors but also some features that modulate gene transcription have been globally linked to APA. Slow transcription rates result in a longer time between the transcription of the proximal and distal sites which was shown to cause increased usage of the proximal site (Neve, Burger et al. 2016). Chromatin status also regulates poly(A) site choice. Global 3' UTR shortening

has been observed with an “open chromatin” state as measured by high H3K4me3 levels in spermatids compared to spermatocytes (Li, Park et al. 2016). Cells under stress conditions such as exposed to arsenite or viruses tend to enhance the usage of intergenic poly(A) sites and generate 3’ extended transcripts (Vilborg, Passarelli et al. 2015, Hollerer, Curk et al. 2016).

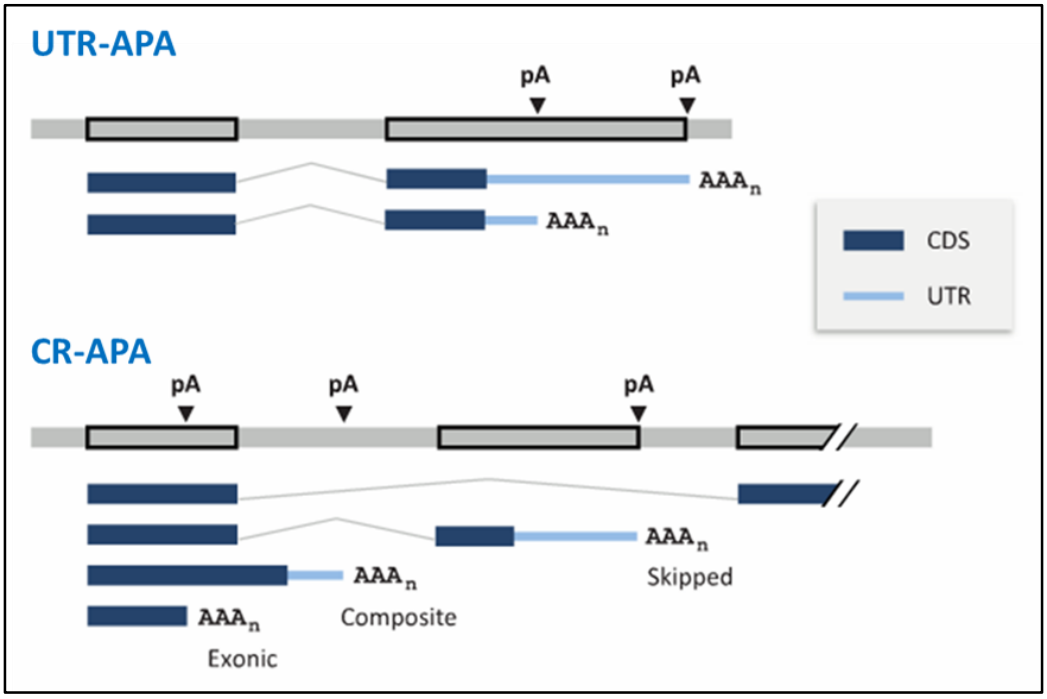


Figure 8 The mRNA diversity caused by alternative polyadenylation (Tian and Manley 2013)

The top panel is 3’ UTR APA (UTR-APA). The bottom panel is coding region APA (CR-APA) which contains skipped APA, composite APA and exonic APA. Different mRNA sequences can be generated by any of these events.

1.9 The importance of next generation sequencing (NGS) approaches for the global analysis of 3' end formation.

The high demand for cost-efficient sequencing has motivated the growth of high-throughput sequencing (or next-generation sequencing) technologies that parallelize the sequencing process, generating thousands or millions of sequences concurrently (Church 2006, Hall 2007). The current popular sequencing platforms are Illumina, Ion Torrent and Pacific Biosciences. The Illumina platform, including MiSeq and HiSeq, is based on reversible dye-terminators that enable the identification of single bases as they are introduced into DNA strands. The Ion Torrent is based on the detection of hydrogen ions that are released during the polymerisation of DNA by semiconductor based system. The Pac Bio utilizes single molecule real time sequencing method enables extra-long read length 14,000 bases average. These high throughput technologies give huge advantages to researchers in assessing biological events including the 3' end formation process on a global scale.

The global mRNA isoforms and mRNA level can be determined by RNA-seq (based on cDNA library) mapped to the reference genome (Ozsolak and Milos 2010). Mapping the 3' end of the transcripts can be accomplished by the identification of the 3' clusters of adenosines. APA usage can be quantified by the ratio of isoform specific read density (Wang, Sandberg et al. 2008). The poly(A) specific sequencing method usually contains the first step of oligo dT priming to mRNA. The cDNA libraries are constructed by RT-PCR (Mangone, Manoharan et al. 2010, Fox-Walsh, Davis-Turak et al. 2011, Fu, Sun et al. 2011, Shepard, Choi et al. 2011, Derti, Garrett-Engele et al. 2012). Different

modifications based on this principle aims to improve the library fidelity and sequencing quality. Other methods including 3' region extraction and deep sequencing (3’ READS) (Hoque, Ji et al. 2012) and direct RNA sequencing (DRS) (Ozsolak, Kapranov et al. 2010) were designed to avoid the PCR process reducing the potential bias introduced during cDNA amplification. But still, the major shortcoming of oligo dT is that an internal A stretch sequence, not necessarily the poly(A) tails, can cause priming. Several programmes have been developed to filter out the internal priming sites (Mangone, Manoharan et al. 2010, Derti, Garrett-Engele et al. 2012). Another sequencing method avoiding oligo dT priming is the 3P-seq (Jan, Friedman et al. 2010). However, the enzymatic steps introduced during library preparation step could generate bias (Hafner, Renwick et al. 2011). The above methods are all based on the RNA from whole cell lysate. The sequencing method based on fractionated RNA has been designed to avoid the bias introduced from the cytoplasmic fraction (Neve, Burger et al. 2016).

Global investigations of RNA-protein interactions are based on the same sequencing technology. The RNA-protein interaction is “frozen” by UV cross-linking and the RNA protein complex is subsequently immune precipitated. Several methods have been developed including CLIP-seq, iCLIP and PAR-CLIP (Ule, Jensen et al. 2003, Ule, Jensen et al. 2005, Licatalosi, Mele et al. 2008, Hafner, Renwick et al. 2011, Konig, Zarnack et al. 2011). The RNA binding sites of several 3’ end processing factors have been mapped globally (Martin, Gruber et al. 2012) and many results from individual poly(A) sites analysis have been confirmed by the global interaction studies.

1.10 Aim of this study

Noncanonical poly(A) sites represent a conundrum, as hexamer mutation of any nucleotide in genes with canonical poly(A) sites abolishes 3’ end formation. Naturally noncanonical sites are relatively frequent and these sites are efficiently processed. Furthermore, in genes with multiple poly(A) sites, the noncanonical sites are overrepresented at the proximal position and during APA are often activated. How noncanonical poly(A) sites are recognised and activated is thus an important question. This question is at the heart of this project and I aim to elucidate how noncanonical cleavage and polyadenylation sites are recognised by the cleavage and polyadenylation machinery. To address these questions, a combination of reporter gene based biochemical methods and whole genome based approaches were used. In the individual gene study, we use the model gene JunD to investigate the 3’ end processing of AGUAAA genes. Also, the model gene ARAF is selected to investigate the 3’ end formation of no PAS genes.

Chapter 2

Materials and Methods

Chapter 2

2.1 Bacteria strains

The *Escherichia coli* (*E.coli*) DH5α competent cells were used for plasmid cloning.

2.1.1 Preparation of competent *E.coli* cells

A single colony (*E.coli* DH5α) was picked and grown at 37℃ overnight in LB media (10g/L Trypton, 5g/L Yeast extract, 10g/L NaCl, pH7.0) without antibiotics until OD_{600} reached 0.5. The bacterial pellet was collected by centrifugation and resuspended in 100mL 0.1M $CaCl_2$ on ice. The cells were collected and resuspended again in 0.1M $CaCl_2$ /15% glycerol and incubated overnight at 4℃. The competent cells were aliquoted into sterile Eppendorf tubes and stored in -80℃.

2.2 Plasmid design, manipulation and cloning

2.2.1 Primers

Primers were designed based on the data from Primer-BLAST or manually designed for specific purpose and then ordered from MWG Eurofins.

2.2.2 Genomic DNA isolation

Human embryonic kidney 293 cells (HEK 293) were washed twice with PBS and resuspended in PBS. Approximately 1,000 cells were diluted in 10µL PBS and 90µL H_2O. The mixture was heated at 95℃ for 10min and then kept on ice. 1µL of 100µL reaction mix was used as the template for 25µL PCR reaction.

2.2.3 Kinase reaction

The kinase enzyme catalyzes the transfer and exchange of Pi from the ɣ position of ATP to the 5'- hydroxyl terminus of polynucleotides (double and single stranded DNA and RNA) and nucleoside 3-monophosphates. The kinase reaction of primers was done in the reaction mix containing 1x NEB kinase buffer and 1 unit of NEB kinase (M0201S). The mixture was incubated at 37℃ for 30min and inactivated at 65℃ for 20min.

2.2.4 Ligation

The 20µL DNA ligation reaction was mixed with a suitable amount of DNA fragment A, DNA fragment B, 2µL 10x T4 DNA ligation buffer, 1µL NEB T4 DNA ligase (M0202S) and H_2O up to 20µL were mixed together and incubated at 20℃ for 2h to overnight and inactivated at 65℃ for 20min.

2.2.5 Digestion

Digestion reactions were performed according to manufacturers' instructions. In brief, a 25µL reaction was set up containing DNA, appropriate 10X Buffer, 1µL restriction enzymes and BSA if required. Reactions were incubated at 37℃ for 2 hours and inactivated at 65℃ or 80℃ for 20min. Some restriction enzymes cannot be heat inactivated.

2.2.6 Blunting ends by 3' overhang removal and 3' recessed end fill-in

DNA was dissolved in 1X NEB buffer 1/2/3/4 or T4 DNA ligase reaction buffer supplemented with 33µM of each dNTP. One unit of NEB DNA Polymerase I, Large Fragment (Klenow) (M0210S) per µg DNA was added and incubated for 15min at 25℃. The reaction was stopped by adding EDTA to a final concentration of 10mM and heated at 75℃ for 20min.

2.2.7 Polymerase chain reaction (PCR)

NEB *Taq* polymerase (18038042) was used for normal PCR. When high fidelity amplification was required, NEB Q5 polymerase (M0491S) was used. PCR was performed according to manufacturer's instructions. The annealing temperature was adjusted for each reaction according to the melting temperature of the pair of primers. If the GC content in either of the primer exceeded 60%, 0.2X GC

Enhancer (NEB) was added in the reaction. All the PCR reactions were executed by Biometra T3000 Thermocycler.

2.2.8 1% Agarose gel electrophoresis

DNA fragments were run on an agarose gel for separation and analysis. Gels were made with 1% agarose (Sigma), 1X TBE buffer (100mM Tris, 100mM boric acid, 1mM EDTA) and 0.5µg/mL ethidium bromide. Samples were premixed with 6X DNA loading dye prior to electrophoresis. Gels were run in 1X TBE buffer in constant 110V. DNA fragments were observed under $UV320_{nm}$ light and pictured in Kodak gel camera system.

2.2.9 Transformation of plasmids to competent cells

The transformation of plasmids was achieved using the heat shock approach. The whole ligation mix or 1µg of plasmid DNA was added to 50µL of *E.coli* DH5α competent cells. The mixture was incubated for 15min on ice, then heat shocked at 42℃ for 30s before another 15min incubation on ice. 250µL LB buffer was added to the mix in 15mL Falcon tubes and incubated in 37℃ shaker for 1h. The mixture was spread in LB agar plates and incubated overnight in a 37℃ incubator. A single bacterial colony was picked the next day and incubated in suitable volume (5mL for mini-prep and 60mL for midi-prep) of 1X LB buffer overnight for plasmid DNA isolation.

2.2.10 Plasmid DNA isolation

A single colony or 500µL of liquid from overnight mini-prep bacteria culture mix was incubated in 60mL of 1X LB medium overnight for midi-prep (Qiagen, 12145) with slight modifications from the manufacturer's instructions. Bacterial cells were harvested by centrifugation at 4℃ for 15min at 6,000g. The pellet was resuspended in 4mL Buffer P1. 4mL Buffer P2 was added and mixed gently by inverting 4-6 times and incubated at room temperature for 5min. 4mL of Buffer P3 was added and mixed immediately but gently by inverting 4-6 times and incubated on ice for 15-20min. A glass funnel with two layers of tissue was used to filter the liquid. In the meantime, Qiagen-tip 100 was equilibrated by applying 4mL of Buffer QBT. The filtered liquid was applied to Qiagen-tip. The tip was washed by adding two times 10mL Buffer QC. DNA was eluted with 5mL Buffer QF. Then DNA was precipitated by adding 3.5mL room temperature isopropanol to the eluted DNA. The mixture was spun immediately at 12,000g for 30min at 4℃. DNA pellet was washed with 5mL of room temperature 70% ethanol and spun at 12,000g for 10min. Supernatant was carefully discarded without disturbing the pellet. The pellet was air-dried for 5-10min and dissolved in a suitable volume of H_2O. The concentration of the DNA was measured by Thermo NanoDrop 1000.

2.2.11 Reporter plasmid construction

We started from the original wildtype MC4R reporter plasmid (Nunes, Li et al. 2010), which contains a CMV promoter, the 5' UTR from *MC1R* gene, a GFP ORF and 1.3kb MC4R fragment origination from downstream of the stop codon.

Then we deleted 50 core sequences including from the A-rich element to DSE of the first MC4R poly(A) site. This was achieved by a two-step PCR strategy. The first PCR step was the amplification of the 3' UTR upstream of the core sequences of the first poly(A) site from the *XbaI* site and the amplification of the fragment from downstream of the DSE of the first poly(A) site until a *HindIII* site. A *XhoI* restriction site was created by introducing the sequence into the reverse primer amplifying the upstream fragment and the forward primer amplifying the fragment downstream of the DSE. These two fragments were ligated and a second PCR step was applied to amplify the fusion fragment. The fusion fragment was cloned into the *XbaI* and *HindIII* site in the original plasmid.

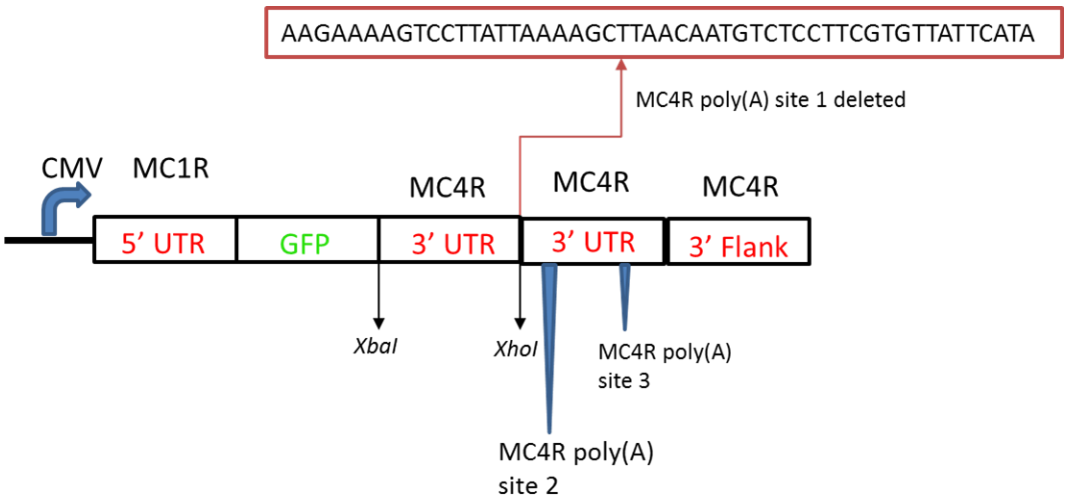


Figure 9 Schematic representation of the MC4R plasmid after the first poly(A) site deletion

The JunD-MC4R constructs were cloned by inserting the JunD fragments into the *XhoI* site as described in Figure 9. The shortest wildtype JunD Mini-gene (containing hexamer and 36 nucleotides downstream) and various mutated JunD Mini-genes including hexamer mutation, DSE mutation, upstream U-rich

region mutation and the upstream U-rich region mutation plus canonical hexamer mutation, were cloned into the reporter at *XhoI* restriction site by the same approach. The sequence of the 42-nucleotide JunD insert was designed in a pair of long primers (46 nucleotides including the sticky ends of *XhoI* site 5'-TCGA-JunD-3'). The forward and reverse primers were annealed in 1X NEB buffer 2 at 95℃ for 5min and then kept on ice. The MC4R plasmid (Fig 9) was linearized by *XhoI* digestion and gel purified by Qiagen gel purification kit. JunD insert was ligated into the linearized plasmid catalyzed by Roche DNA ligase (10481220001). The correct JunD-MC4R construct was then isolated by *XhoI* digestion. In addition, the Zip4 Mini-genes were also cloned into the reporter construct by the same method. The initial plasmid verification was done by restriction enzyme digestion with *XhoI* (destroyed) and *BamHI* (present in the plasmid described in Fig 9) resulting in a linearized plasmid without any short cut. The orientation verification was done by PCR amplification with forward primer in MC4R plasmid and reverse primer in JunD/Zip4 insert. Final verification was achieved by sequencing.

The long JunD Mini-genes (100bp, 150bp and 200bp) were cloned by a slightly different strategy. JunD fragments were amplified first by forward and reverse primers introducing the *XhoI* restriction site. The MC4R plasmid was linearized by *XhoI* digestion. The wildtype JunD inserts were then digested and ligated into the plasmids. The plasmid verification step was identical as previously described.

The ARAF fragments were inserted into *XbaI* and *XhoI* site in the MC4R plasmid (Fig 9). The wildtype ARAF fragment was amplified from genomic DNA by a forward primer containing *XbaI* site and a reverse primer containing *XhoI*

site. The MC4R plasmid was digested by *XbaI* and *XhoI* restriction enzymes. Then ARAF fragments were inserted by Roche ligase. The mutation of putative poly(A) signal in ARAF fragment was introduced by the sequence on the primer achieved by two-step PCR strategy. In the first step, two fragments of ARAF one with the mutation and another with sequence downstream of the first one were amplified. The reverse primer of the first PCR and the forward primer of the second PCR was modified by kinase. Then two PCR fragments were ligated and amplified by a forward primer containing *XbaI* site and a reverse primer containing *XhoI* site. The MC4R plasmid was digested by *XbaI* and *XhoI* restriction enzymes. Then ARAF fragments were inserted by Roche ligase. The first verification step was done by *XhoI* and *XbaI* digestion. The second verification was done by sequencing.

2.2.12 DNA sequencing

The DNA sequencing of PCR products and plasmids was carried out by the Source Bioscience Sequencing Service. The sequencing result was aligned by ApE plasmid editing software.

2.2.13 Plasmid transfection

(i) Cell culture: HEK 293 cells were incubated in Dulbecco's Modified Eagle's medium (DMEM, Sigma, D5546) supplemented with 10% fetal bovine serum (FBS, Sigma, 12003C), 1% L-glutatmine (Sigma, G3126), and 1% penicillin-streptomycin (Sigma, P4333) at 37℃, 5% CO_2.

(ii) Plasmid transfection: Cells were transfected with plasmid at 60% confluence. The transfection reaction complex contained 3µg plasmid, 150µL DMEM medium and 25µL Polyfect (Qiagen, 301105) and was incubated for 20min at room temperature. The reaction mix was added to the cells after washing the cells twice with PBS (Sigma, D8662) and replacing with the fresh cell culture medium. RNA was isolated 24 hours later.

2.3 RNA manipulation and analysis

2.3.1 RNA Extraction

(i) RNA subcellular fractionation

Cells were incubated in one 225 cm^2 flask till 80% confluence. Fractionated RNA was extracted using TRIzol (ThermoFisher, 15596026). Cells were washed by ice-cold PBS twice and then trypsinized by adding 3mL Trypsin (Sigma, T1426). Cells were then spun down in 10mL DMEM medium for 5min at 1000rpm at 4℃. Cells were then washed twice in ice-cold PBS and spun down at 1000rpm for 5min at 4℃. The cell pellet was resuspended in 1mL Lysis Buffer B (10mM Tris-HCl pH8, 140mM NaCl, 1.5mM $MgCl_2$, 0.5% NP40) to a 1.7mL Eppendorf tube. If required, 150µL was removed for the whole cell fraction and added to 150µL of TRIzol and kept on ice; 50µL was removed for protein fraction analysis and added to 16.7µL of 4X Laemmli buffer (40% Glycerol, 240mM Tris/HCl pH6.8, 8% SDS, 0.04% bromophenol blue, 5% β-mercaptoethanol). Cells were then spun at 1000g for 3min at 4℃. The supernatant was transferred to a new 1.7mL tube as the unpurified cytoplasmic

fraction. This fraction was spun at 10000rpm for 1min at 4℃ and the supernatant was transferred to a new tube and kept on ice to purify as the cytoplasmic fraction. The pellet in the original 1.7mL Eppendorf tube was the nuclear fraction. The pellet was resuspended in 1mL Lysis Buffer B and transferred to 10mL round-bottom Falcon tube. 100µL of Detergent stock (3.3% (w/v) Sodium deoxycholate, 6.6% (v/v) Tween 40) was added drop-by drop under slow vortexing and then incubated on ice for 5 min. The suspension was then transferred to a new 1.7mL tube and spun at 1000g for 3min at 4℃. The supernatant was discarded and the pellet was the nuclear fraction. The nuclear pellet was resuspended in 1mL of Lysis Buffer B and spun at 1000g for 3min at 4℃. The pellet represented intact nuclei and it was resuspended in 1mL of TRIzol. The suspension was passed through a 21-gauge needle until no lumps were present. The cytoplasmic fraction was added to 500µL of TRIzol. The TRIzol mix of all three fractions was incubated at room temperature for 5min. Cytoplasmic, nuclear and whole-cell fraction were added to 100µL, 200µL and 30µL of chloroform respectively and vortexed vigorously for 15s. All the fractions were spun at 12000g for 15min at 4℃. The aqueous fractions of cytoplasmic, nuclear and whole-cell fractions were transferred to 750µL, 580µL or 230µL of isopropanol. The isopropanol mixtures were precipitated at room temperature for 10min and then spun at 12000g for 10min at 4℃. The RNA pellet was resuspended in 87µL of RNase-free H_2O. DNase treatment was applied by adding 10µL 10X DNase buffer, 2µL DNaseI (Roche, 04716728001) and 1µL RNaseOUT (Invitrogen, 10777019) and incubating at 32℃ for 30min. Phenol chloroform mix (1:1) was then added to the DNase reaction. The mixture was vortexed and spun at 13000g for 5min at room temperature. The

supernatant was transferred to a new 1.7mL Eppendorf tube and phenol chloroform extraction was repeated. RNA was precipitated in 250µL 100% ethanol and 10µL of 3M NaOAc for a minimum of 2h at -20℃. The RNA pellet was collected by spinning at 13,000rpm at 4℃ for 30min and resuspended in RNase free H_2O. The concentration of the RNA was measured by Thermo NanoDrop 1000.

(ii) Whole-cell RNA isolation – Hot phenol

Whole-cell RNA isolation was used for the polyadenylation signal analysis (RT-PCR and RNase protection assay) after MC4R plasmid transfection. Cells were incubated in one 75cm^2 flask and grown to 80% confluency. Cells were washed with ice-cold PBS twice and then scraped off using a cell scraper. Cells were then spun down in 10mL PBS for 5min at 1000rpm at 4℃. If required, 100µL was removed for protein fraction and added to 33.3µL of 4X Laemmli buffer before cell pellet collection. Hot phenol solution (500µL Phenol (Sigma, P1037), 450µL 10X NTE buffer (5M NaCl, 100mM Tris-HCl pH7.5, 10mM EDTA pH8) and 50µL 10% SDS) was heated at 90℃ simultaneously. The cell pellet was collected and added to hot phenol solution in a 1.7mL screw-cap tube. The cells were vortexed vigorously for 3min and then spun at 13000g for 5min at room temperature. The aqueous phase was transferred to 500µL phenol chloroform mix (1:1). 1250µL of ethanol was added to the phenol chloroform purification product and were precipitated at -20℃ for 4h. RNA was spun down at 12,000g for 30min at 4℃. The RNA pellet was resuspended in 87µL of RNase-free H_2O.

The RNA pellet was subjected to DNase treatment as described in section 2.3.1(i).

2.3.2 Reverse transcriptase polymerase chain reaction (RT-PCR)

The property of RNA transcripts was qualitatively measured by RT-PCR. Total or fractionated RNA (1µg) was primed by oligo dT or gene-specific reverse transcription primer at 75℃ for 5min and then reversely transcribed to cDNA by M-MLV reverse transcriptase (ThermoFisher, 28025013) in 400µM dNTP and 1X RT Buffer at 37℃ for 60min. Subsequently, 1µL cDNA was used as the template for PCR.

2.3.3 Electrophoresis of RNA in denaturing Formaldehyde/1% agarose gel

The quality of the fractionated RNA was analyzed on denaturing formaldehyde/1% agarose gels. The gel was made with 1X formaldehyde gel running buffer (8mM sodium acetate, 1mM EDTA PH8, 20mM MOPS), 3.3% formaldehyde and 1% agarose. RNA samples were premixed with formamide, formaldehyde running buffer and formaldehyde and then denatured at 65℃ for 15min. Formaldehyde gel loading buffer (50% glycerol, 1mM EDTA PH8, 0.25% bromophenol blue, Xylene cyanol FF) and ethidium bromide (EtBr) were added to the samples prior to electrophoresis. Denaturing formaldehyde gels were run

in 1X formaldehyde gel running buffer. RNA fragments were visualised under UV_{320nm} light.

2.3.4 Polyacrylamide gel electrophoresis

Radiolabelled RNA fragments were separated and analyzed on polyacrylamide gels. The gel was made with 50% urea, 6% acrylamide, 1X TBE, 0.06% TEMED and 0.06% APS. RNA samples were premixed with RNA loading buffer (80% formamide, 10mM EDTA, 0.25% bromophenol blue, 0.25% Xylene cyanol FF) and boiled for 15 minutes prior to electrophoresis. These gels were run in 1X TBE buffer. Films were developed with a Compact X4 developer or in Fujifilm Phosphorimager.

2.3.5 RNase protection assay

(i) *In vitro* transcription

The linearized pGEM4 plasmid containing T7 polymerase site was used as the template in *in vitro* transcription. The sense sequence of the targeting regions in JunD-MC4R reporter (spanning JunD poly(A) site, MC4R poly(A) site 2 and MC4R poly(A) site 3) was amplified. *EcoRI* and *HindIII* restriction sites were generated at each end during the PCR reaction by the additional sequence in forward or reverse primers. After digestion with *EcoRI* and *HindIII* restriction enzyme, the targeting regions were inserted into pGEM4 plasmid downstream of T7 polymerase site. The sequence of the targeting region in pGEM4 was confirmed by sequencing to be mutation-free. 5µg of JunD-MC4R pGEM4

plasmid was linearized by *EcoRI* digestion and purified by agarose gel. Antisense RNase protection riboprobes were *in vitro* transcribed in the 20µL reaction complex containing linearized JunD-MC4R pGEM4 plasmid (0.5µg-1µg), 0.25mM (ACG)TP, 5µM UTP, 5mM DTT, 1X transcription buffer, 1u/µL RNaseOUT, [α-32P]-UTP (400Ci/mmol) and 50u/µL T7 polymerase (NEB, M0251S) at 37℃ for one hour.

(ii) Probe purification

The *in vitro* transcription mix was added to 80% formamide RNA loading dye and boiled for 5min. 6% polyacrylamide gel was run in 35W for 2h. Then the radioactive gel was exposed in the film for 2min and developed. The probe was cut and incubated in 350µL cracking buffer at room temperature for 1h. Then it was purified by phenol chloroform extraction. The RNA probe was precipitated by adding 30µL NaOAc and 750µL ethanol. Then the RNA pellet was dissolved in R-Loop buffer. The volume of R-Loop buffer was adjusted based on the radioactive counts of the pellet.

(iii) Hybridisation of target RNA and probe

The hybridisation reaction mix (30µL in total) contained 200 counts of probe normally around 10µL and RNA (3µg-50µg) in R-Loop Buffer. It was denatured at 80℃ for 5min and then hybridised overnight at 55℃.

(iv) RNA digestion and protection

The hybridised RNA mix was added to 300µL RNAse A (Qiagen, 19101)/ RNase T1 (Sigma, R1003) mix and was incubated at 30℃ for 1h. Then the reaction mix was added to 30µL Proteinase K (Sigma, P6556) mix and was incubated at 37℃ for 45min. The sample was purified by phenol chloroform extraction. The RNA was precipitated by adding 20µg tRNA and 1mL of ethanol. The pellet was resuspended in 15µL 80% formamide loading dye and run in 6% polyacrylamide gel at 40W for 30min and 35W for 90min. The gel was dried at 80℃ for 2h and then exposed at room temperature overnight.

(v) Quantification of RNase protection data

The RNase protection data was collected and analyzed in a Fujifilm Fluorescent Image Analyser FLA-3000. The strength of the band was quantified by ImageJ software by calculating the grey scale of the bands. The usage ratio of the poly(A) site or read through was calculated by the formula (specific poly(A) site or read through usage)/(JunD poly(A) site usage + MC4R poly(A) site usage 2 + MC4R poly(A) site usage 3+ read through usage). The standard deviation of the data and the related chart was calculated and drawn by Prism Graphpad. The statistical comparison between two usage ratio was performed by Excel in a two-tailed t-Test.

2.3.6 RNAi transfection

Transient protein knockdown was achieved by RNAi transfection. CPSF30 (5'-ACGCCAACAGAUGCACCAAAGTT, 5'-CUUUGGUGCAUCUGUUGGCGUTT) and hFip1 (5'-GUGCUGAUCUUUCUGAUUATT, 5'-UAAUCAGAAAGAUCAGCACTT) siRNA was ordered from MWG Eurofins according to the sequence from the Tian Lab (Li, You et al. 2015). WDR33, CPSF160 and SAM68 silencer select siRNAs were ordered from Life Technology. The siRNA was transfected in a reaction mix containing Opti-MEM medium (Sigma, 31985062), Lipofectamine RNAiMAX (ThermoFisher, 13778075) when the cells reached 30% confluence. siRNA containing media was replaced by normal DMEM media 24 hours after the first transfection. Cells were transfected again 24 hours after the replacement of the normal media. RNA and protein was extracted 24 hours after the second transfection. The knockdown efficiency was measured by RT-PCR or Western Blotting. In RNAi based sequencing experiment, negative control siRNA (Sigma SIC001) was transfected by the same method.

2.4 Protein analysis

2.4.1 SDS PAGE

The protein isolation method in the fractionation experiment is stated above. The protein samples were collected by dissolving cells in Laemmli buffer before hot phenol RNA extraction. Protein samples were separated in SDS PAGE. These gels contained resolving gel part (10% Acrylamide, 400mM Tris PH8.8,

0.1% SDS, 0.1% APS, 0.1% TEMED) and stacking gel part (5% Acrylamide, 125mM Tris PH6.8, 0.1% SDS, 0.1% APS, 0.25% TEMED). Samples were boiled in Laemmli buffer for 10min prior to electrophoresis. The gels were run in 1X SDS running buffer (25mM Tris, 192mM glycine, 0.1% SDS).

2.4.2 Western blotting assay

A semi-dry transfer method was used if the size of the targeting protein was under 80 kDa. The transfer was done by Bio-Rad Trans-Blot® SD semi-dry transfer cell. The SDS gel was attached to PVDF blotting membrane (GE Healthcare, 10600021) sandwiched by four layers of blotting paper each side in 1X transfer buffer (25mM Tris-HCl, 192mM glycine, 20% methanol). Alternatively, wet transfer method was used if the size of the relevant protein was above 80 kDa. Proteins were transferred to PVDF membrane in wet electroblotting systems (Bio-Rad). Subsequently, the membrane was pre-blocked in 5% blocking powder in TBST buffer (50mM Tris, 150mM NaCl, 0.05% Tween20) overnight. Then the membrane was incubated in first antibody (1:500-1:5000) for 1h-overnight. The membrane was incubated in second IgG antibody (1:5000) after three washing steps. Proteins were detected by enhanced chemiluminescence (ECL) method. The blots were developed in Kodak films and analyzed with ImageJ software.

2.5 Sequencing and data analysis

2.5.1 RNA 3’ end sequencing

Nuclear RNA was used to generate libraries, since total whole-cell RNA mainly consists of the cytoplasmic fraction, which cannot fully represent the nucleic polyadenylation event because mRNAs may be subjected to degradation after the nuclear cytoplasmic transport (Neve, Burger et al. 2016). Each isolation was subjected to quality control where the quality of all RNA fractions were assessed by (i) monitoring the nuclear long noncoding RNA MALAT1 (Tripathi, Ellis et al. 2010) by RT-PCR and (ii) visualising the precursor ribosomal RNA (Henras, Plisson‐Chastang et al. 2015) in a denaturing RNA gel. A small fraction of the cells was also collected in each case to verify the respective knockdown efficiency by Western blot.

Nuclear RNA that passed the fractionation quality control step was subsequently used to prepare sequencing libraries. Library preparation was completed with Lexogen QuantSeq 3' mRNA-Seq Library Prep Kit (012.24A/B) based on RT-PCR principle. Lexogen’s QuantSeq kit provides a library preparation protocol designed to generate Ion Torrent compatible libraries from polyadenylated RNA. The QuantSeq protocol generates only one sequencing fragment per transcript, resulting in extremely accurate gene expression values as determined by RNA-seq. The kit is designed to yield sequences close to the 3’ end of the transcripts which makes it ideal to study polyadenylation at a global scale. Library generation was started with 500ng of nuclear RNA or 1.5µg of cytoplasmic RNA. Reverse transcription process was initiated with oligo dT priming containing Ion Proton compatible linker sequences. Second strand

synthesis was performed with random priming and a DNA polymerase. Random primers contained Ion Proton compatible linker sequences. Second strand synthesis was followed by a magnetic bead-based purification step. The library was amplified by 13-cycle PCR reaction and the complete sequences required for colony formation on Ion Torrent platforms were introduced. Another magnetic bead-based purification step was employed.

The quality of the library was measured by Agilent 2100 Bioanalyzer in Agilent DNA high sensitivity chip. This chip accommodated sample wells, gel wells and a well for an external standard (ladder). During chip preparation, the micro-channels were filled with gel dye mix. Then, eleven samples and one ladder with marker were loaded in each well and applied in Bioanalyzer instrument. The DNA as charged biomolecules were driven by a voltage gradient. Smaller DNA fragments migrated faster than larger ones and were detected by laser-induced fluorescence. The data was translated into simulated gel images and electropherograms (peaks).

The sequencing platform used was the Ion Proton (ThermoFisher). The sequencing chip was loaded by Ion Chef System. The raw data was processed through the algorithm developed by Tian Lab, our collaborator and was described in (Hoque, Ji et al. 2012) to generate the raw Excel file with the information (including reads per million, localisation, poly(A) signal, name of the gene) of poly(A) sites and genes as well as the figure of the distribution of the significant APA events.

Ion Proton Sequencing

Ion semiconductor sequencing (by ThermoFisher) is a method of DNA sequencing based on the detection of hydrogen ions that are released during the polymerization of DNA. This is an approach of "sequencing by synthesis", during which a complementary strand is built based on the sequence of a template strand. A micro-well containing a template DNA strand to be sequenced is flooded with a single species of deoxyribonucleotide triphosphate (dNTP). If the introduced dNTP is complementary to the leading template nucleotide, it is incorporated into the growing complementary strand. This causes the release of a hydrogen ion that triggers an ion-sensitive field-effect transistor (ISFET) ion sensor, which indicates that a reaction has occurred. If homopolymer repeats are present in the template sequence, multiple dNTP molecules will be incorporated in a single cycle. This leads to a corresponding number of released hydrogens and a proportionally higher electronic signal.

2.5.2 Bioinformatics analysis

(i) Gene category definition

Two gene categories (“AGTAAA” and “All noncanonical”) were filtered from the raw excel file to investigate the global influence of the protein factors on noncanonical poly(A) sites. In “AGTAAA” gene category, globally all the genes with tandem poly(A) sites, one directed by AGUAAA hexamer signal another directed by AAUAAA/AUUAAA signal, and both were located in 3’ UTR of the gene were included. In “All noncanonical” gene category, globally all the genes

with tandem poly(A) sites, one directed by AAUAAA/AUUAAA signal another directed without AAUAAA/AUUAAA signal, and both were located in 3’ UTR of the gene were included.

(ii) Global noncanonical usage analysis

The noncanonical poly(A) site usage ratio was calculated in each individual gene with the formula (Noncanonical poly(A) signal)/ (Noncanonical + Canonical poly(A) signal). The strength of the poly(A) signal was measured by reads per million reads. The preference of the protein factors on noncanonical poly(A) site in each gene was determined by the comparison of the protein knockdown data to the control data. Global influence was determined by the change of the distribution of noncanonical poly(A) site usage ratio after the protein factor knockdown.

(iii) Motif identification

The sequence including AGUAAA hexamer and 20 nucleotides downstream from the up- or downregulated genes in response to the protein factor knockdown was applied in MEME programme for the motif analysis.

Chapter 3

The Investigation of the recognition of the JunD poly(A) site as a representative of cleavage and polyadenylation at an AGUAAA noncanonical site

Chapter 3

3.1 Introduction

Pre-mRNA 3' end formation is a fundamental processing step for the maturation of mRNAs in all eukaryotes (Colgan and Manley 1997). All protein-encoding primary transcripts are cleaved at their 3' end and, with the exception of replication-dependent metazoan histone genes, are subsequently subjected to addition of a string of adenosines resulting in mature transcripts with characteristic poly(A) tails (Pandey and Marzluff 1987). The co-transcriptional cleavage and polyadenylation reaction is directed by a large multi-protein complex, which in humans can constitute more than eighty proteins (Shi, Giammartino et al. 2009). The mammalian cleavage and polyadenylation reaction is cooperatively initiated by the interaction of CPSF, CFIm and CstF complexes with the poly(A) signals (PAS) on the pre-mRNA transcripts. The majority of the core 3' processing factors are well conserved. The PAS is defined by multiple cis elements. For canonical mammalian poly(A) signals, these cis elements include an AAUAAA or AUUAAA hexamer, which comprise 70% of the poly(A) sites (Neve, Burger et al. 2016), a less well defined DSE, and other auxiliary sequence elements including the U-rich upstream stimulatory element (USE) (Tian and Graber 2012).

Recognition of the so called noncanonical poly(A) sites, which lack the canonical hexamer signals, are poorly understood. In this chapter, we focus on the recognition of AGUAAA containing poly(A) sites, the most frequent noncanonical poly(A) site. We started from the RNA secondary structure prediction amongst AGUAAA poly(A) sites in collaboration with Dr James

Anderson. We then applied Mini-gene and mutagenesis based approaches to explore the *cis* elements important for JunD 3' end formation, which we chose as a representative of AGUAAA poly(A) site gene. Finally, different protein factors were depleted to evaluate their role on canonical and AGUAAA poly(A) sites 3' end processing.

3.2 RNA secondary structure prediction of AGUAAA containing poly(A) sites

The secondary structure of a nucleic acid molecule refers to the base-pairing interactions within a single molecule or set of interacting molecules, and can be represented as a list of bases which are paired in a nucleic acid molecule. The secondary structures of biological RNAs tend to be different from DNA (Mathews, Moss et al. 2010). Biological DNA mostly exists as fully base paired double helices. Due to the fact that much biological RNA is single stranded, it often forms complicated base pairing interactions because of its increased ability to form hydrogen bonds stemming from the extra hydroxyl group in the ribose sugar. The GU wobble base pair, with two hydrogen bonds, frequently occurs in RNA (Doty, Boedtker et al. 1959, Fresco, Alberts et al. 1960). Several RNA secondary structures including stem-loops and G-Quadruplexes have been shown to influence the 3' end formation (Ahmed, Gilmartin et al. 1991, Gilmartin, Fleming et al. 1992, Decorsiere, Cayrel et al. 2011).

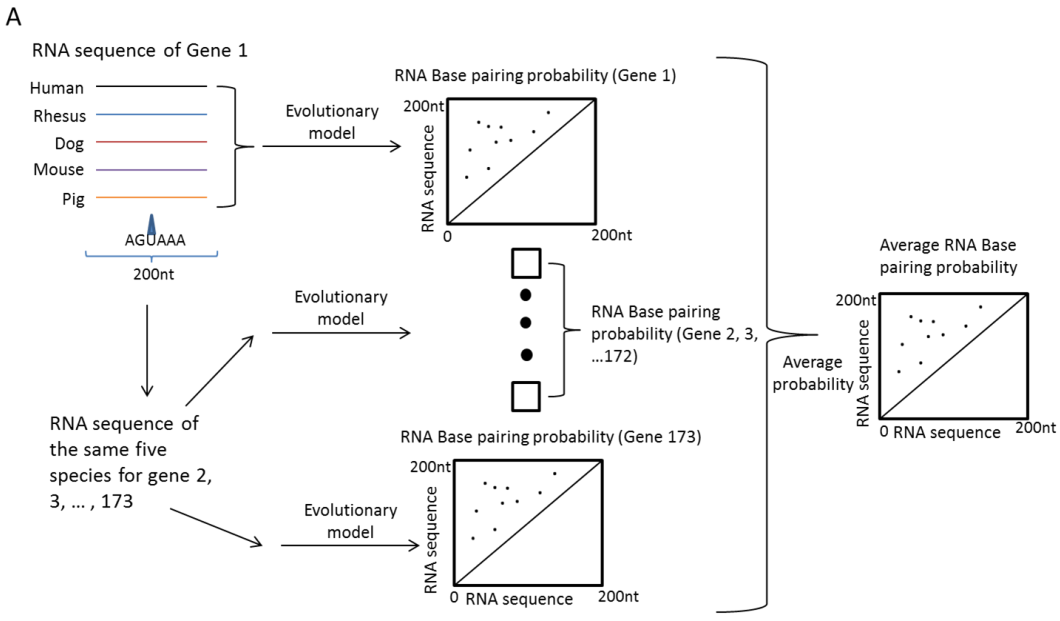
In order to investigate whether, globally, AGUAAA containing poly(A) sites have typical stem loop structures surrounding their 3' end processing signals, we

used the Oxfold algorithm to predict RNA secondary structure. It is an evolutionary model based on iterative algorithm in the framework of stochastic context-free grammar developed by JWJ Anderson, our collaborator in Department of Statistics, Oxford. This algorithm is the first ever model to consider both RNA folding kinetics and evolutionary information (Anderson, Haas et al. 2013). By integrating biological principle, this approach improves considerably on existing grammatical models. Additionally, the model compares favourably to non-kinetic thermodynamic models (Hofacker, Fontana et al. 1994, Markham and Zuker 2008). Significant structures like double helices or stem loops can be predicted by the model.

To perform the model, we interrogated 200-nucleotide long sequences (97nt up- and downstream of AGUAAA) of all the 173 genes that have a single AGUAAA poly(A) site (filtered from the 3' READS data (Neve, Burger et al. 2016)). Since the algorithm is based on an "evolutionary" model, sequences from five species per gene including human, rhesus, dog mouse and pig were used to predict the structure. The programme first analyzed the probability of each base pairing within the given nucleotides in a single gene context in the evolutionary model, from five species, and then calculated the average probability of each base pairing within the given sequences in 173 genes (Fig 10A). The result can be seen in Figure 10B. 173 groups of random sequences with or without AGUAAA hexamer in the mid-point were generated and calculated by the programme as the control (Fig 10C, D). The darker a particular dot appearance is, the higher its probability of being a base pairing point is. And continuous dark dots represent a helical structure with high probability. Compared to the control (Fig 10 C, D), the *in silico* analysis

suggests the appearance of several helices (Fig 10 B, blue circles) that could represent common secondary structure elements associated with AGUAAA poly(A) sites.

We conclude that AGUAAA poly(A) sites have structural similarities that may be important for their recognition. To further investigate the importance of specific secondary structures to the AGUAAA poly(A) site 3' end formation, biochemical based analysis on individual poly(A) site examples need to be applied.



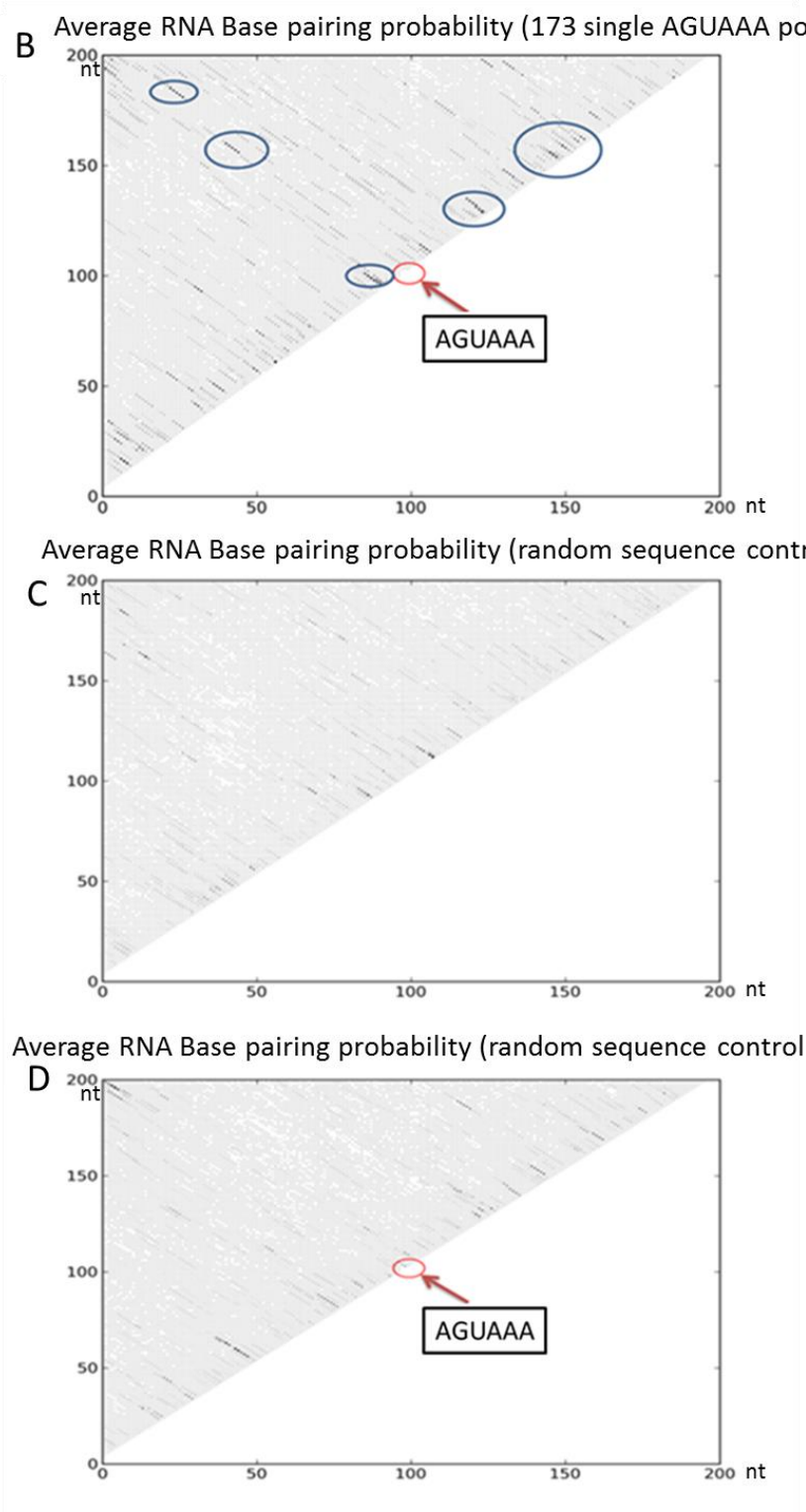


Figure 10 The RNA secondary structure prediction

A) Work flow of the secondary structure predicting. B) Normalised averaged cubed RNA base-pair probabilities of 173 AGUAAA RNA sequences with 200nt length calculated by Oxfold. Individual AGUAAA RNA sample was analyzed first with different five species (human, rhesus, mouse, dog and pig) in the “evolutionary” model. Each

dot represents the interaction between two nucleotides (location showed in X-axis and Y-axis). A larger intensity dot indicates higher base-pairing probability. Position 0 indicates the first nucleotide in the sequence, position 200 the last, the same for X-axis and Y-axis. Continuous dots in the blue circles represent the possible helical structures present in the 173 AGUAAA containing genes. C) Normalised averaged cubed base-pair probabilities of 173 random sequences with 200nt length calculated by Oxfold. Individual RNA sample was analyzed first with five random sequences simulating different species in the “evolutionary” model. D) Normalised averaged cubed base-pair probabilities of 173 random sequences with 200nt length with AGUAAA motif produced in the middle. Individual RNA sample was analyzed first with five random sequences simulating different species in the “evolutionary” model.

3.3 *JUND* is an ideal candidate to study noncanonical AGUAAA poly(A) site recognition

The protein encoded by the intronless *JUND* gene is a member of the JUN family, and a functional component of the AP1 transcription factor complex (Nicolaides, Correa et al. 1992). This protein has been proposed to protect cells from p53-dependent senescence and apoptosis (Weitzman, Fiette et al. 2000). JUND has been shown to interact with ATF3, MEN1, DNA damage-inducible transcript 3 and BRCA1 (Nilsson, Toftgård et al. 1995, Agarwal, Guru et al. 1999, Hu and Li 2002). According to our sequencing data using HEK 293 RNA and the 3’ READS method (Fig 11) (Neve, Burger et al. 2016), *JUND* has only one poly(A) site that is defined by the noncanonical hexamer AGUAAA, which represents the most frequent noncanonical hexamer (Fig 5). As *JUND* is an intronless gene, it is perfectly suited to study how its noncanonical poly(A) site

is recognised by the cleavage and polyadenylation machinery without the complication of splicing dependent 3' end formation.

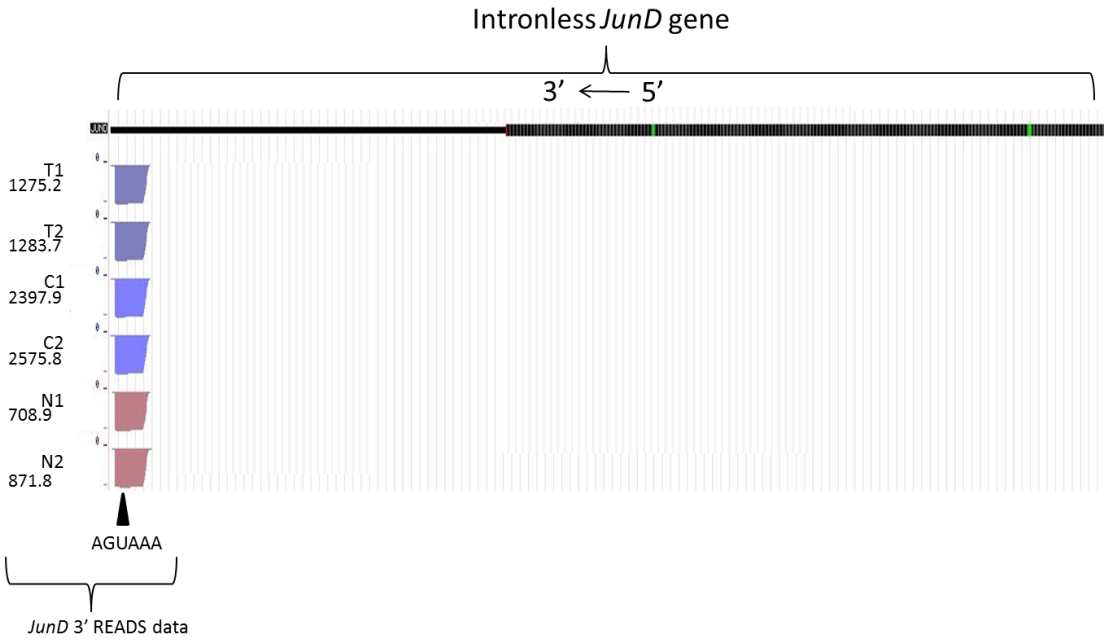


Figure 11 RNA 3' READS sequencing data of *JUND* in HEK 293 cells

RNA 3' READS deep sequencing data (Neve, Burger et al. 2016) of *JUND* mRNA in HEK 293 cells using whole cell (T1, T2), cytoplasmic (C1, C2) and nuclear (N1, N2) fractions. The only poly(A) signal in each sample is detected 14 nucleotides downstream of the only hexamer signal AGUAAA. The same cleavage and polyadenylation location occurs in all six samples (replicates). The data is visualised in UCSC. Reads per million in all the six samples are shown.

We then revisited the data generated by Oxfold algorithm described in Figure 10B to explore the potential secondary structures in the *JUND* gene poly(A) site. As can be seen in Figure 12A, the AGUAAA hexamer is between the 98th and 103rd nucleotide in X- and Y- axis. The dark dots in blue and burgundy circles show a high probability of base pairing which leads relevant nucleotides to fold into helical structures. As an example, the structure in the blue circle is

visualised in Figure 12B. The impact of the predicted secondary structure on *JUND* cleavage and polyadenylation was subsequently analyzed by mutagenesis analysis.

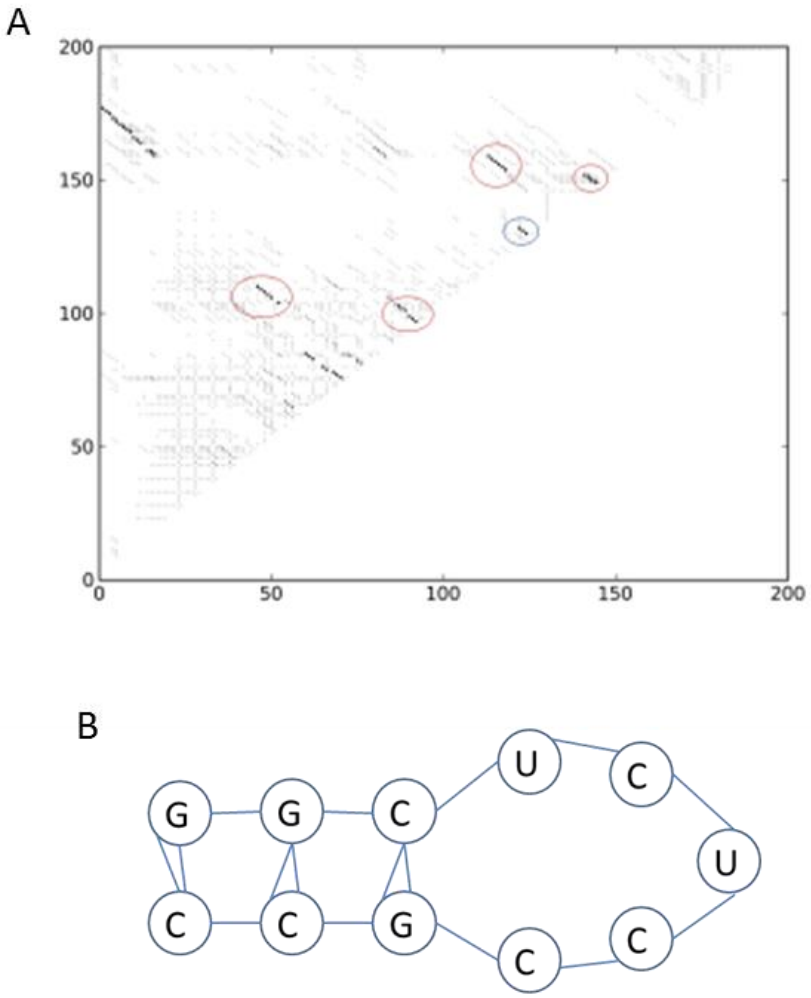


Figure 12 The RNA secondary structure prediction of *JUND* pre-mRNA

A) Dot-plot of the base-pair probabilities cubed produced by Oxfold. Five 200-nucleotides RNA sequences of *JUND* from human, rhesus, mouse, dog and pig were used to perform the model. The AGUAAA hexamer was located from the 98th nucleotide to the 103rd nucleotide, both in X- and Y-axis. A larger intensity dot indicates higher base-pairing probability. Position 0 indicates the first nucleotide in the sequence, position 200 the last, the same for X-axis and Y-axis. Dark dots in the blue and

burgundy circles are helical structures with high probability. B) Visualisation of the helical structure predicted by Oxfold in the blue circle in (A).

3.4 The sequence requirement of the JunD poly(A) site

3.4.1 The identification of the minimal *cis* elements that can drive JunD 3' end formation in a Mini-gene context

The initial series of experiments focused on establishing a Mini-gene based system that used the JunD poly(A) site for 3' end formation to determine the minimal cis elements that are required for function. To do this, the MC4R reporter plasmid, as described in Figure 9, was used to insert the JunD poly(A) site upstream of the MC4R poly(A) site 2. To create this poly(A) competition vector, the function of the alternative poly(A) site needed to be confirmed after the deletion of the MC4R poly(A) site 1 in the plasmid described in Figure 9. This plasmid was transfected into HEK 293 cells. After 24 hours, whole cell RNA was isolated by the hot phenol method (see Materials and Methods). cDNA was generated by Oligo dT annealing and reverse transcription. The subsequent PCR reaction was directed by a forward primer "F" in the GFP region upstream of the JunD insert and Oligo dT (Fig 13A). In the RT-PCR analysis, a strong band corresponding to the size of the MC4R poly(A) site 2 was detected (Fig 13B MC4R poly(A) site 2, lane 2). However, a further downstream band and an upstream band were also detected. All these three bands were purified and sequenced. The band close to MC4R poly(A) site 2 was confirmed to be the polyadenylation site of MC4R poly(A) site 2, directed by AUUAAA with the exact cleavage site slightly different from (15nt

downstream of) the reported site (Nunes, Li et al. 2010). The proximal poly(A) site named “Cryptic poly(A) site” was possibly directed by an AUAAAA hexamer signal. The distal poly(A) named “MC4R poly(A) site 3” site was potentially directed by another canonical hexamer AUUAAA. These two sites are somehow activated after the deletion of the major poly(A) site. From this RT-PCR analysis, we verified the functional poly(A) sites in the major poly(A) site deletion MC4R plasmid and this plasmid can be further modified to generate the poly(A) competition plasmid.

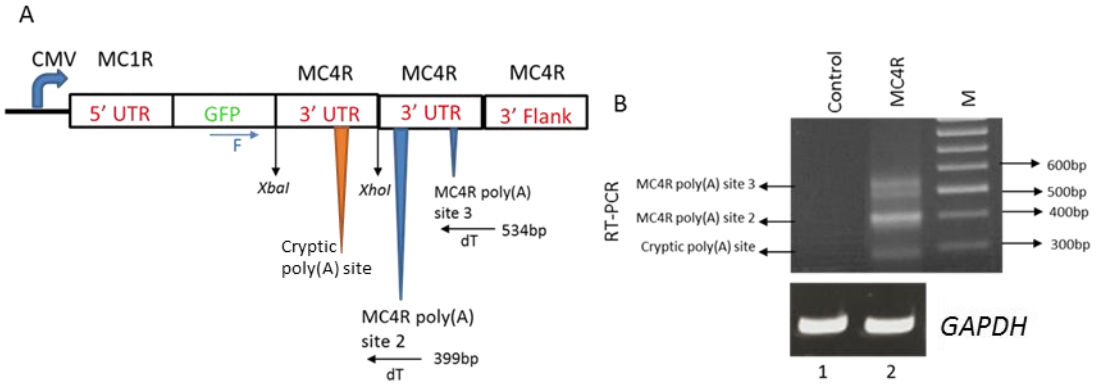


Figure 13 The functional poly(A) sites in the MC4R plasmid after the major poly(A) site deletion

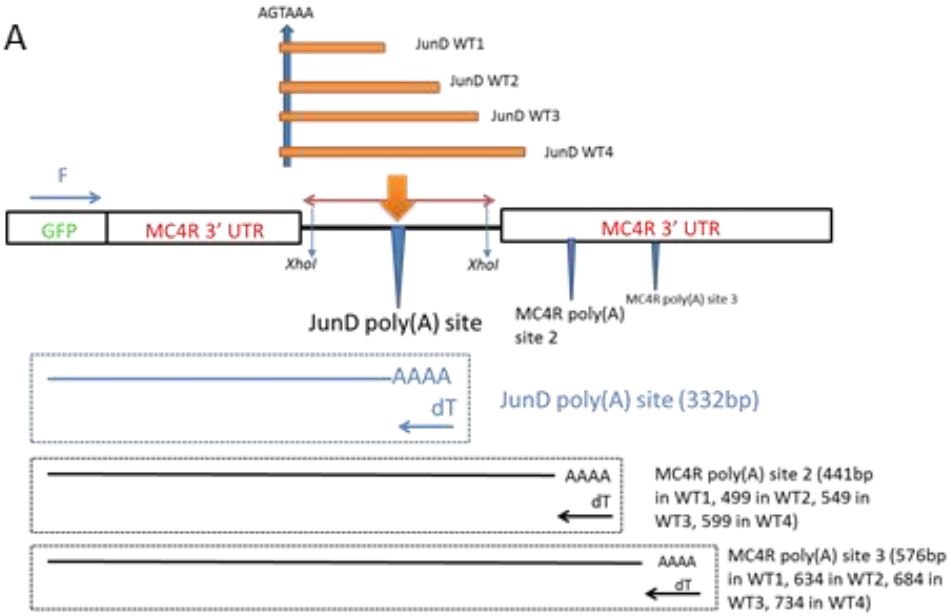
(A) Schematic depicting major poly(A) site deletion MC4R plasmid. The relative position of the two canonical MC4R poly(A) sites (MC4R poly(A) site 2 and 3) and another cryptic poly(A) site is indicated. The primers and the length of putative products in the RT-PCR analysis are as indicated. (B) RT-PCR analysis of MC4R polyadenylation. The position of different of poly(A) sites is indicated. Control refers to samples without plasmid transfection. *GAPDH* serves as an RNA loading control.

To analyze the unusual *JunD* poly(A) site, we employed the MC4R reporter plasmid as described in Figure 9. The JunD Mini-gene plasmids were created by cloning various fragments, which differed in length (42nt, 100nt, 150nt and 200nt, Fig 14B) and encompassed the JunD poly(A) site, into an *XhoI* site,

located in the MC4R 3' UTR upstream of the MC4R poly(A) sites (Fig 14A, B). These two alternative MC4R poly(A) sites are present 100bp and 200bp downstream of the inserted JunD fragment and can be used to assess the strength of the JunD poly(A) site (Fig 14A). The cryptic poly(A) site upstream of the *XhoI* site was reserved in JunD-MC4R plasmids, according to the cloning strategy (see Materials and Methods). The plasmids containing different lengths of JunD were subsequently transfected into HEK 293 cells. 24h after transfecting the cells, whole cell RNA was extracted using the hot phenol method (see Materials and Methods). Subsequently, cDNA was generated by Oligo dT annealing and reverse transcription. The subsequent analytic PCR reaction to assess JunD poly(A) site usage was performed by a forward primer "F" in the GFP region upstream of the MC4R 3' UTR and Oligo dT (Fig 14A). As can be seen from the RT-PCR results, for all JunD constructs, only one type of polyadenylated transcripts can be obviously detected, which corresponds to mRNAs that are cleaved and polyadenylated at the JunD AGUAAA poly(A) site (Fig 14C lane 1,2,3,4 JunD poly(A) site). Surprisingly, it appears that only 42 nucleotides including the AGUAAA hexamer are sufficient to drive highly complete 3' end processing at the JunD site. To confirm that polyadenylation in this construct occurs at the correct site, the PCR fragment was gel purified and sequenced. This analysis showed that the cleavage site matched that of endogenous *JUND* transcripts, as established by the 3' READS sequencing data (Neve, Burger et al. 2016), and maps to 14 nucleotides downstream of the AGUAAA hexamer, after a "CA" dinucleotide (Fig 14D). This confirmed that the 42 nucleotides inserted into the reporter gene represent a functional AGUAAA type poly(A) site. Considering the fact that many canonical cleavage and

polyadenylation sites are considered to require an array of auxiliary elements (Decorsiere, Cayrel et al. 2011, Tian and Graber 2012), it is surprising that only 42 nucleotides are efficient enough to direct 3' end formation at the JunD noncanonical poly(A) site. From now, the JunD WT1 construct will be referred to as JunD WT.

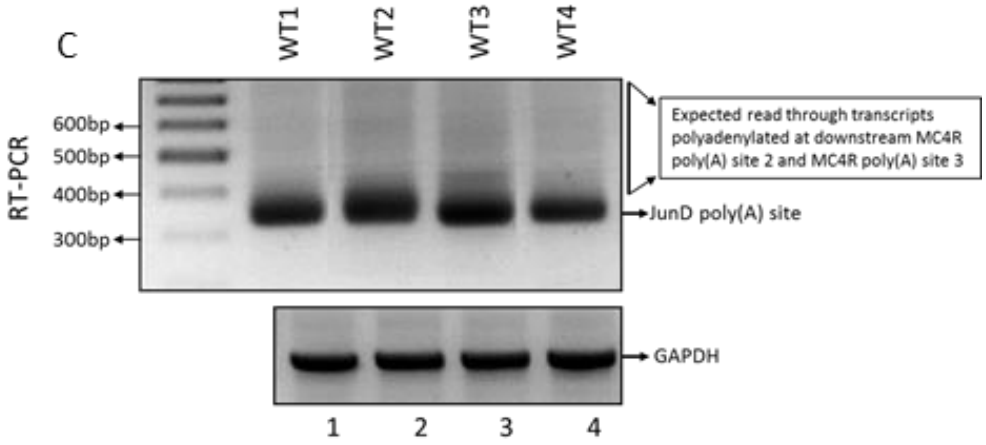
A



B

Name	Insert sequence (5' – 3')	Insert length (nt)	JunD poly(A) signal length (bp)
JunD WT1	AGTAAAGTCTCGTTACGCCAGCTCGGCTCTCCGCCTCCTTCT	42	332
JunD WT2	AGTAAAGTCTCGTTACGCCAGCTCGGCTCTCCGCCTCCTTC TTCCCCCGCCGGGGGCCTGGCGGGCTGGGCGGGGCCTGGT TCGCTTCCCCCCTACACCCC	100	332
JunD WT3	AGTAAAGTCTCGTTACGCCAGCTCGGCTCTCCGCCTCCTTC TTCCCCCGCCGGGGGCCTGGCGGGCTGGGCGGGGCCTGGT TCGCTTCCCCCCTACACCCCCGCCTTTCTCTCCTCCCTCTGT CCCTAGCTCCTCCCCTTGTTGGAGGGAG	150	332
JunD WT4	AGTAAAGTCTCGTTACGCCAGCTCGGCTCTCCGCCTCCTTC TTCCCCCGCCGGGGGCCTGGCGGGCTGGGCGGGGCCTGGT TCGCTTCCCCCCTACACCCCCGCCTTTCTCTCCTCCCTCTGT CCCTAGCTCCTCCCCTTGTTGGAGGGAGCAGAGAGCGGG TGTCTGAGAGCGGCCCGCACATCTGGCGCGCAGCGAGCG	200	332

C



D

Figure 14 The identification of JunD minimal working region

(A) Schematic depicting the JunD-MC4R constructs. Different lengths of JunD sequences were cloned into the same position in the MC4R reporter at the *XhoI* restriction site. The relative position of two downstream MC4R poly(A) sites are depicted. The length of three expected PCR products is indicated (JunD poly(A) site, MC4R poly(A) site 2 and MC4R poly(A) site 3). Note that the length of MC4R poly(A) signals is different in four JunD-MC4R constructs as indicated. (B) Sequences of different lengths of wildtype JunD cloned into the MC4R reporter. (C) RT-PCR analysis of JunD and MC4R cleavage and polyadenylation. The expected position of three poly(A) signals: JunD poly(A) site is indicated. The position of MC4R poly(A) site 2 and MC4R poly(A) site 3 varies as indicated in (A). *GAPDH* serves as an RNA loading control. (D) The cleavage site, defined by the sequencing of the gel isolated RT-PCR products, is indicated by a red arrow on the 42-nucleotide JunD WT1 sequence. The hexamer signal is indicated in blue letters.

3.4.2 The identification of *cis* elements in JunD

To further dissect the JunD minimal poly(A) sequence element and identify which sequences are essential for function, a mutagenesis based approach was applied. The mutations were designed to target potential *cis* elements and silence them (Fig 15B). Seven mutations altering the sequence of the 42-nucleotide JunD WT (previously JunD WT1) fragment were designed and cloned into *XhoI* site of the MC4R plasmid (Fig 15A). The rationale behind the seven mutations is summarised in Table 2.

Table 2 Summary of the reasons to generate JunD mutations

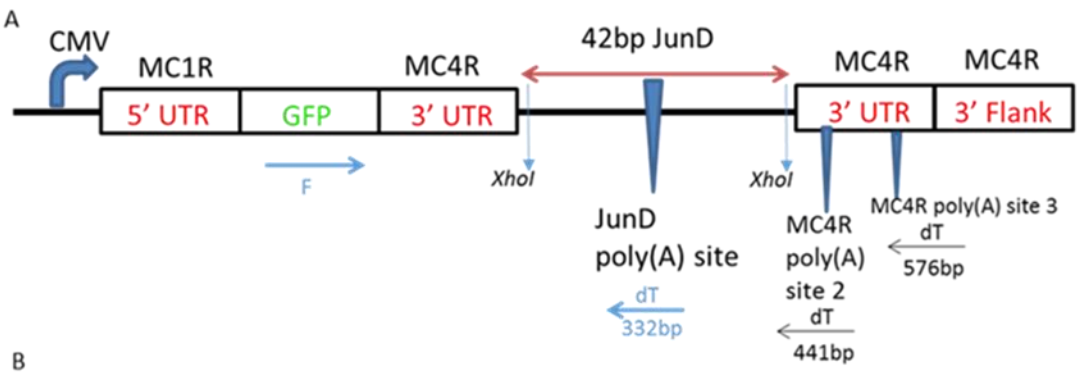
JunD-MC4R plasmid	Aim of Mutation
Mut1	Silence the noncanonical hexamer signal AGUAAA
Mut2	Destroy tandem uridines between the hexamer and the cleavage site (potential hFip1 binding site)
Mut3	Mutate all the “C”s in DSE region and destroy the putative RNA secondary structure as described in Figure 17B
Mut4	Alter all the uridines between the hexamer and the cleavage site
Mut5	Mutate the canonical cleavage site dinucleotide “CA”
Mut6	Change the noncanonical signal to a canonical one
Mut7	Silence the downstream element

After the transfection of these plasmids and RNA extraction, reverse transcribed cDNA products were, as in the previous experiment, subjected to PCR using the forward primer “F” and “dT” reverse primer (Fig 15A). For the assessment of the poly(A) strength in each mutant, the PCR products corresponding to transcripts that are polyadenylated at the JunD site can be compared to the portion of transcripts that read through the JunD site but are captured by the downstream MC4R poly(A) sites. For Mut1 and Mut7, the poly(A) signal of JunD disappears completely (Fig 15C, lane 2 and 8, JunD poly(A) site) and all resulting PCR products represent transcripts that either read through the JunD poly(A) site and are processed downstream at the MC4R sites (Fig 15C lane 2, 8, MC4R poly(A) site 2 and 3), or processed at a cryptic poly(A) site (Fig 15C lane 2 and 8, Cryptic poly(A) site). In essence, both of these mutations compromise the JunD poly(A) site significantly and show that the hexamer AGUAAA and DSE region “CUCUCCGCCUCCUUCU” are functionally important *cis* elements of the JunD poly(A) site. In Mut4, usage of

the JunD poly(A) site is significantly reduced as PCR products representing transcripts processed at the MC4R sites are prominent (Fig 15C lane 5) indicating that the four uridines adjacent to the hexamer constitute an auxiliary element for the noncanonical JunD poly(A) site. No MC4R poly(A) site usage can be observed in any of the other four mutations (Fig 15C lane 3, 4, 6 and 7), which indicates that no other *cis* elements are present in the regions targeted by the mutations. Interestingly, the mutation of the frequent cleavage site sequence “CA” as represented by Mut5 has little effect on cleavage and polyadenylation efficiency and does not alter the position of cleavage site (confirmed by sequencing). Also, changing the noncanonical hexamer to the canonical has no impact on polyadenylation. Moreover, the secondary structure predicted and highlighted in Figure 12B is not likely to influence JunD 3’ end formation.

In summary, the recognition of the noncanonical JunD poly(A) site depends on the presence of the AGUAAA, the DSE and four additional auxiliary uridines in the “GUCUCGUU” region which is located adjacent to the hexamer. We define the four auxiliary uridines as the “adjacent U-rich element” (adU-rich).

A



B

Name	Insert Sequence (5′-3′) Hexamer Cleavage site DSE	Insert length (nt)	JunD poly(A) signal length (bp)
JunD WT	**AGTAAA**GTCTCGTTACGCCAGCTCGGCTCTCCGCCTCCTTCT	42	332
JunD Mut1	**AG***C***AA***C***GTCTCGTTACGCCAGCTCGGCTCTCCGCCTCCTTCT	42	332
JunD Mut2	**AGTAAA**GTCTCG*CA*ACGCCAGCTCGGCTCTCCGCCTCCTTCT	42	332
JunD Mut3	**AGTAAA**GTCTCGTTACGCCAGCTCGG*A*TCTC*A*GC*A*TCCTT*A*T	42	332
JunD Mut4	**AGTAAA**G*CC*CCG*CC*ACGCCAGCTCGGCTCTCCGCCTCCTTCT	42	332
JunD Mut5	**AGTAAA**GTCTCGTTACGC*GG*GCTCGGCTCTCCGCCTCCTTCT	42	332
JunD Mut6	**A*A*TAAA**GTCTCGTTACGCCAGCTCGGCTCTCCGCCTCCTTCT	42	332
JunD Mut7	**AGTAAA**GTCTCGTTACGCCAGCTCGGC*A*C*A*CCGCC*A*CC*AA*C*A*	42	332

C

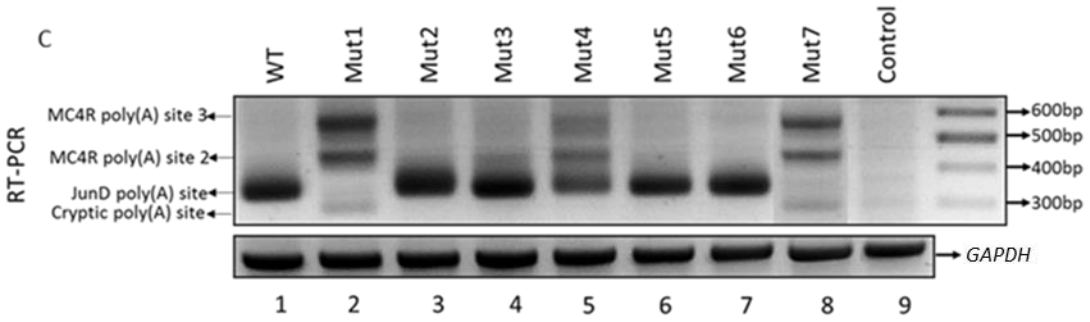


Figure 15 The identification of JunD *cis* elements

(A) Schematic depicting JunD-MC4R plasmids. Wildtype and seven mutations of JunD fragments were cloned into the same position at the *XhoI* restriction site in the MC4R plasmid. The relative position of two downstream MC4R poly(A) sites is depicted. The expected length of the three poly(A) signals is indicated. (B) Sequences of wildtype and JunD mutations cloned into the MC4R plasmid. The mutated nucleotides are indicated in red italic letters. The expected length of the JunD poly(A) signal is indicated in each plasmid. The position of the hexamer, the cleavage site and the DSE is indicated. (C) RT-PCR analysis of JunD and MC4R polyadenylation. The position of the polyadenylation signal of JunD, the first poly(A) signal of MC4R (MC4R poly(A) site 2) and the second poly(A) signal of MC4R (MC4R poly(A) site 3) is indicated. Another cryptic poly(A) signal appears when the JunD poly(A) signal is completely abolished.

WT and Mut1-7 correspond to the JunD sequences in (B). The control refers to the samples without JunD-MC4R plasmid transfection. *GAPDH* serves as an RNA loading control.

3.4.3 Recovery of the reduction of poly(A) site usage in JunD Mut4

To further address the nature of the adU-rich auxiliary element, we want to demonstrate next whether this element is linked to the noncanonical hexamer or not. To address this issue, an additional mutant reporter plasmid was constructed by altering the noncanonical hexamer into the canonical hexamer AAUAAA in the Mut4 plasmid, generating the Mut4a reporter gene (Fig 16A JunD Mut4a). The same RT-PCR analysis (Fig 14A) was applied to measure the JunD poly(A) site usage in the plasmid. JunD WT and Mut1 were selected as two controls representing the condition of full and no usage of the JunD poly(A) site. As can be seen in Figure 16B, all Mut4a transcripts are processed at the JunD poly(A) site and no PCR products representing transcripts processed at the downstream MC4R poly(A) sites can be detected (Fig 16B lane 4, MC4R poly(A) site 2 and 3). This indicates that changing the noncanonical to the canonical hexamer abolishes the dependence of the poly(A) site on the auxiliary adU-rich element.

We conclude that the four uridines function to compensate for the degeneration of the canonical hexamer to a noncanonical AGUAAA hexamer and act as an auxiliary element to ensure optimal poly(A) site usage of the JunD poly(A) site.

A

Name	Insert Sequence (5'-3')	Insert length (nt)	JunD poly(A) signal length (bp)
JunD WT	**AGTAAA**GTCTCGTTACGCCAGCTCGGCTCTCCGCCTCCTTCT	42	332
JunD Mut1	**AG*C*AA*C***GTCTCGTTACGCCAGCTCGGCTCTCCGCCTCCTTCT	42	332
JunD Mut4	**AGTAAA**G*CC*CG*CC*ACGCCAGCTCGGCTCTCCGCCTCCTTCT	42	332
JunD Mut4a	**A*A*TAAA**G*CC*CG*CC*ACGCCAGCTCGGCTCTCCGCCTCCTTCT	42	332

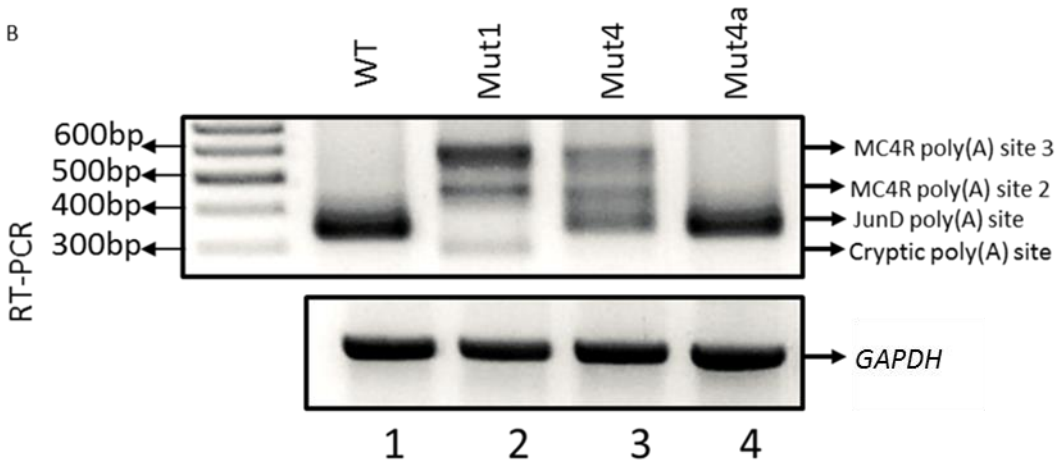


Figure 16 The canonical hexamer recovers the loss of poly(A) site usage caused by four-uridine mutation in JunD Mut4

(A) Sequences of wildtype (WT) and Mut1, 4 and 4a of JunD cloned into the MC4R plasmids. The mutated nucleotides are indicated in red italic letters. (B) The same RT-PCR approach as in Fig 18A was adopted to assess polyadenylation. In Mut4a the noncanonical AGUAAA in Mut4 is altered to the canonical hexamer AAUAAA. WT and Mut1 are two controls representing examples of the full JunD poly(A) site usage and no usage respectively. *GAPDH* serves as an RNA loading control.

To assess the effects of the mutations more quantitatively, an RNase protection approach was adopted. This technique utilizes radioactive probes with the sequence complementary to the targeting RNA to form double stranded RNA molecules that are protected from digestion by RNase A and RNase T1 (Fig 17B) (Carey, Peterson et al. 2013). With this technique, usage of different poly(A) sites can be more accurately quantified. For this analysis we used four

distinct radioactively labelled antisense riboprobes that were complementary to the sequences spanning the JunD poly(A) site in the transcripts generated by the JunD WT, Mut1, Mut4 and Mut4a constructs respectively. All four WT and mutant specific riboprobes extend beyond the second MC4R poly(A) site (MC4R poly(A) site 3), and thus can hybridise and protect transcripts that are processed at the JunD poly(A) site and those transcripts that read through the JunD poly(A) site and processed at either of the MC4R poly(A) sites. The probes further extend 30 nucleotides beyond the third poly(A) site, so are designed to also protect and identify transcripts that read through all poly(A) sites and are thus not processed at all (Fig 17A). Poly(A) site usage at each poly(A) site was quantified by establishing the ratio of the signals at these positions relative to the total signal at all positions (Fig 17C, signal of specific poly(A) site relative to signal of JunD poly(A) site plus MC4R poly(A) site 2 plus MC4R poly(A) site 3 plus read through transcripts).

Using this strategy, cells were transfected with the four plasmids and total RNA was isolated and subjected to RNase protection analysis. The RNase protection assay confirms the RT-PCR results (Fig 16B) obtained with the JunD WT plasmid showing that over 90% of all generated transcripts are cleaved at the JunD poly(A) site (Fig 17D, JunD WT, green bar). Accordingly, and again matching the results obtained with the RT-PCR approach (Fig 16B), mutation of the hexamer in JunD Mut1 reduces usage down to below 1% (Fig 17D Mut1 green bar). Removal of the auxiliary adU-rich element as in Mut4 causes an obvious drop in JunD poly(A) site usage (45%) and this is restored to almost WT levels (89%) when adU-rich mutation is combined with a canonical AAUAAA hexamer (Fig 17D Mut4, Mut4a, green bars) respectively. Also

statistically, JunD poly(A) site usage ratio in Mut4a is not significantly different from JunD WT (Table A1). Overall, the RNase protection analysis is clearly consistent with the previous RT-PCR based results (Fig 16B), confirming the initial finding that the adU-rich auxiliary element is essential for optimal usage of the JunD poly(A) site and it functions to compensate for a degenerated hexamer in JunD.

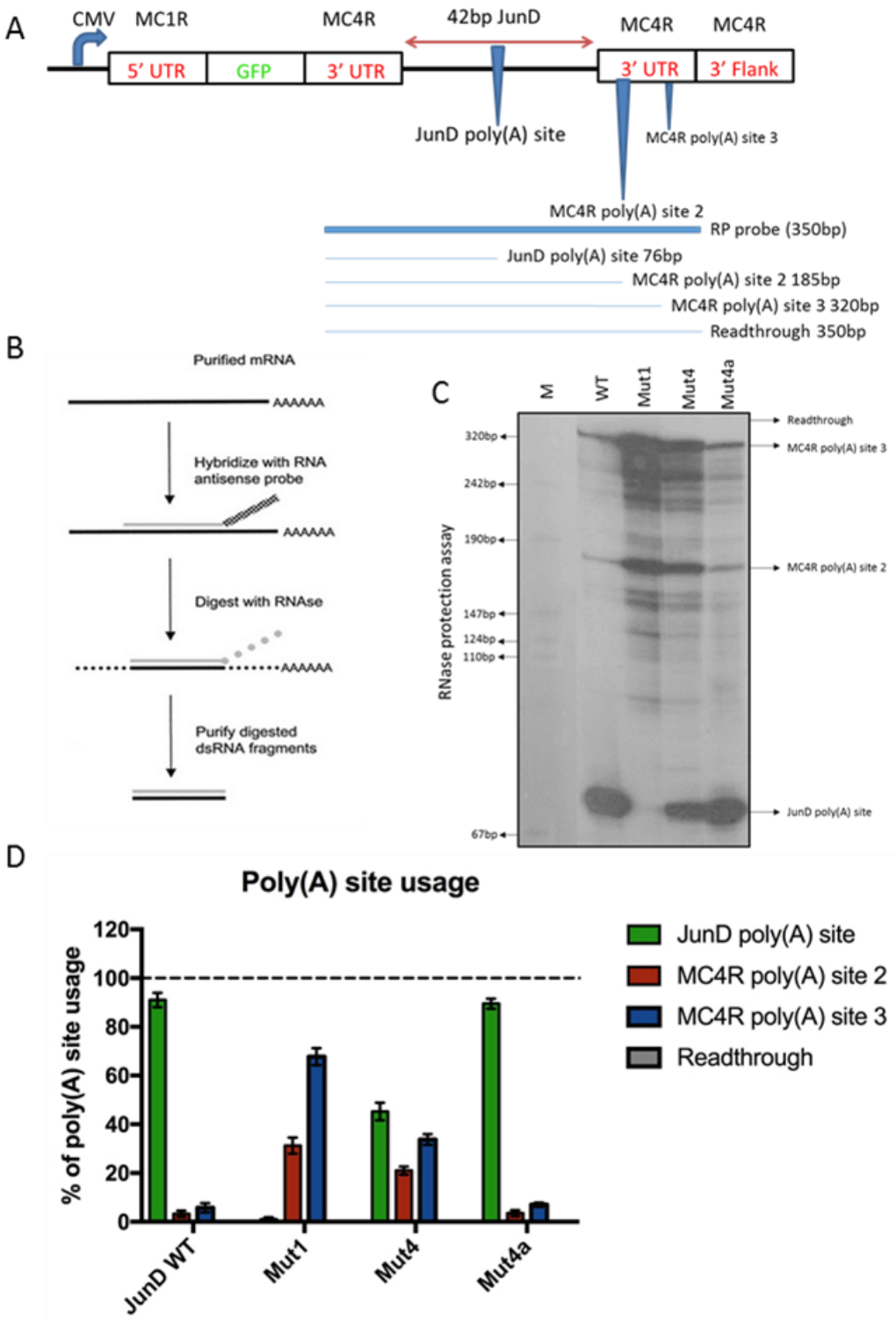


Figure 17 Quantification of the JunD poly(A) site usage

(A) Design of riboprobes used in the RNase protection assay. The sequence of the probe is complementary to the region spanning all the three poly(A) sites. A matching probe for the WT and the three mutant clones were generated to ensure full complementary of the antisense probe to the transcripts generated from each plasmid.

The lengths of the possible protected fragments are indicated. Read through transcripts result in a 350bp protected probe. (B) Schematic depicting RNase protection assay. The *in vitro* transcribed probe was labelled with α32[P]UTP. Thus the radioactive double stranded RNA can be protected and detected (Carey, Peterson et al. 2013). (C) The RNase protection assay gel exposed in phosphorimager. HEK 293 cells were transfected by the identical JunD-MC4R plasmids JunD WT, Mut1, Mut4 and Mut4a. The expected position of each poly(A) signal and read through transcripts is indicated. No read through signal was detected. (D) Quantitative result of the RNase protection assay. The absolute usage value of the poly(A) sites and read through transcripts were quantified by respective grey scale calculated by ImageJ software. The usage ratio was calculated by (Poly(A) site X)/(JunD poly(A) site + MC4R poly(A) site 2 + MC4R poly(A) site 3 + Readthrough). Green, red, blue and grey bars represent the usage ratio of JunD poly(A) site, MC4R poly(A) site 2, MC4R poly(A) site 3 and read through transcripts respectively (n=4).

3.4.4 The position of the adU-rich element relative to the AGUAAA may be critical for JunD 3' end formation

The adU-rich auxiliary element "GUCUCGUU" is located adjacent to the AGUAAA hexamer. In order to elucidate whether the close proximity between the hexamer and the adU-rich element was an important factor for the function of the 3' end processing site, an additional plasmid named MutD was constructed. Four nucleotides "ACAC" were inserted between the hexamer and the adU-rich element (Fig 18A). The plasmid was transfected into HEK 293 cells and after RNA isolation, poly(A) site usage was assessed by RT-PCR analysis as described earlier (Fig 14A). Interestingly, spacing the AGUAAA and the adU-rich element by four nucleotides had a mild negative impact on the JunD usage, as usage at the MC4R poly(A) sites were slightly increased (Fig 18B lane 2,

MutD). This suggests that the auxiliary element needs to be positioned close to the AAUAAA to act efficiently.

A

Name	Insert Sequence (5'- 3')	JunD poly(A) signal length (bp)
JunD WT	AGTAAA*GTCTCGTT*ACGCCAGCTCGGCTCTCCGCCTCCTTCT	332
JunD MutD	AGTAAA*ACACGTCTCGTT*ACGCCAGCTCGGCTCTCCGCCTCCTTCT	336

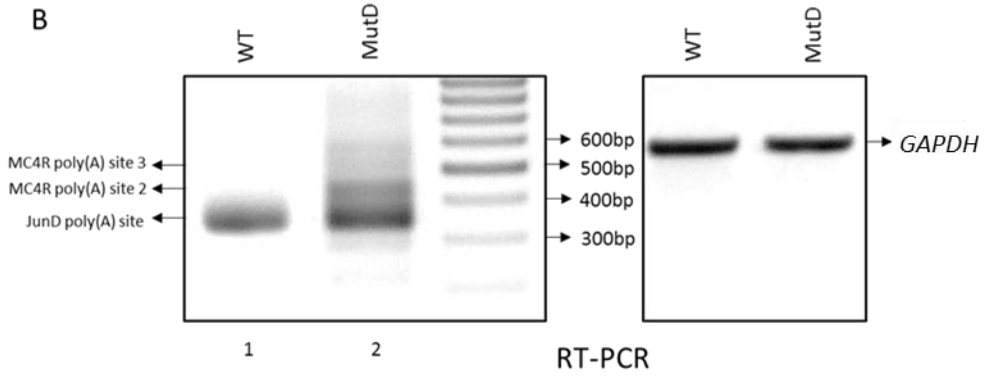


Figure 18 The close distance between AGUAAA and the adU-rich element is important for JunD cleavage and polyadenylation

(A) Sequences of wildtype and lengthened JunD mutation cloned into the MC4R plasmids. The inserted nucleotides are indicated in red italic letters. The adU-rich element is indicated in purple italic letters. The expected length of RT-PCR products representing transcripts processed at the JunD polyadenylation signal is indicated. (B) RT-PCR analysis of JunD and MC4R polyadenylation. The position of JunD poly(A) site, MC4R poly(A) site 2 and MC4R poly(A) site 3 is indicated. *GAPDH* is used as an RNA loading control.

Another gene *Zip4* was selected to further investigate the potential importance of the adU-rich element in the AGUAAA poly(A) site. This gene encodes a member of the zinc/iron-regulated transporter-like protein (ZIP) family. The encoded protein localizes to cell membranes and is required for zinc uptake in

the intestine (Dufner-Beattie, Wang et al. 2003). Its aberrant high level expression contributes to pancreatic cancer pathogenesis and progression (Li, Zhang et al. 2007). Similar to *JUND*, according to our 3' READS data, there is only a single poly(A) site with an AGUAAA hexamer present in this gene (Neve, Burger et al. 2016). Importantly, a potential adU-rich element (UCUUGUCCUU) is present downstream of the *ZIP4* AGUAAA hexamer. However, compared to *JUND*, the hexamer and the U-rich region are spaced by an additional four nucleotides "ACAC" in *ZIP4* mimicking the condition achieved in JunD MutD. To test whether the Zip4 poly(A) site functions under similar parameters as in JunD, a minimal 51-nucleotides long fragment of the Zip4 poly(A) site was cloned into the *XhoI* site in the reporter gene, generating the Zip4 WT construct. An additional construct was created in which the four nucleotides "ACAC" were deleted and as a result positioned the potential adU-rich element "UCUUGUCCUU" of the *ZIP4* gene immediately downstream of the AGUAAA, creating a similar architecture as in JunD, in the resulting Zip4 sh plasmid (Fig 19A). The same RT-PCR as described above for the JunD Mini-genes (Fig 14A) was employed to assess the Zip4 poly(A) site usage qualitatively. Unlike in the case of JunD, the minimal sequence that is required for Zip4 poly(A) to function was not defined. As the RT-PCR analysis in Figure 19B shows, when cells were transfected with the Zip4 WT construct, there is no PCR product detectable between the 300bp and 400bp mark that would represent the transcripts cleaved at the Zip4 site (Fig 19B lane1, Zip4 poly(A) site). This indicates that 51 nucleotides of the *ZIP4* 3' UTR and 3' flanking region, encompassing the AGUAAA hexamer and a potential DSE, are not sufficient to direct its 3' end formation. Surprisingly though, when the potential adU-rich region is moved

immediately adjacent to the AGUAAA, as is achieved in the Zip4 sh mutant plasmid, a PCR product representing transcripts that are cleaved at the Zip4 poly(A) site emerges (Fig 19B lane 2, Zip4 poly(A) site). To confirm the identity of this PCR product, the band was extracted and sequenced. The sequence confirmed that this product is the result of amplifying transcripts that are cleaved and polyadenylated at the Zip4 poly(A) site.

Although the Zip4 poly(A) site usage in the Zip4 WT Mini-gene is almost non-existent, deleting four nucleotides between AGUAAA hexamer and the U-rich region somehow activates the AGUAAA site. This strongly suggests an adU-rich element is present in the Zip4 poly(A) site and conforms that an adU-rich element close to the AGUAAA can act as a 3' end processing enhancer.

A

Name	Insert Sequence (5'- 3')	Zip4 poly(A) signal length (bp)
Zip4 WT	AGTAAA*GACAC**TCTTGTCCTT*GGAGCATGGCTGTGCTCTCCTTGCTGGGTG	341
Zip4 sh	AGTAAA*GTCTTGTCCTT*GGAGCATGGCTGTGCTCTCCTTGCTGGGTG	337

B

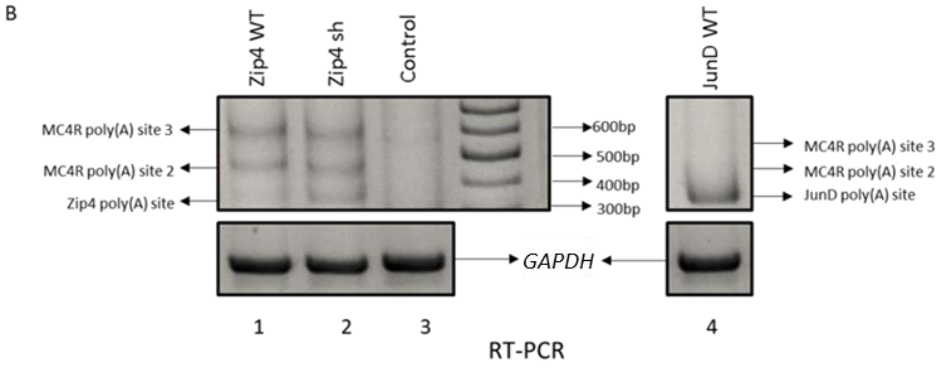


Figure 19 The distance between AGUAAA and U-rich region is important for Zip4 3' end formation

(A) Sequences of wildtype (WT) and shortened Zip4 mutation (Zip4 sh) cloned into the MC4R plasmids. The U-rich region is indicated in purple italic letters. The nucleotides (ACAC) deleted in shortened mutation are indicated in green letters in Zip4 WT. The length of the expected PCR fragments representing transcripts processed at the Zip4

poly(A) site is indicated (B) RT-PCR analysis of Zip4 and MC4R polyadenylation. JunD WT represents poly(A) site full usage control. Zip4 sh is the Zip4 shortened mutation. The control is without plasmid transfection. The position of Zip4 poly(A) site, JunD poly(A) site, MC4R poly(A) site 2 and MC4R poly(A) site 3 is indicated. *GAPDH* serves as an RNA loading control.

3.5 A U-rich motif is frequently present in single AGUAAA poly(A) site containing genes

A motif is a sequence pattern that occurs repeatedly in a group of related sequences and is often associated with some biological function. To assess whether the U-rich regions present in *JUND* and *ZIP4* poly(A) sites is a typical motif in AGUAAA containing poly(A) sites, *in silico* analysis was performed by Multiple EM for Motif Elicitation (MEME), which is a tool for discovering motifs. MEME represents motifs as position-dependent letter-probability matrices which describe the probability of each possible letter at each position in the pattern. Individual MEME motifs do not contain gaps. Patterns with variable-length gaps are split by MEME into two or more separate motifs (Bailey, Boden et al. 2009). It is considered to be a useful tool to predict the *cis* elements that are present in a group of poly(A) sites.

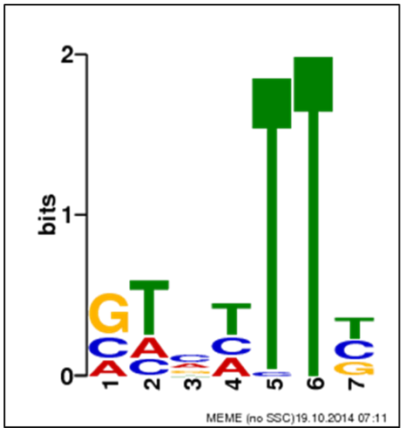
To initiate motif analysis, sequences adjacent to the AGUAAA hexamer and upstream of the cleavage site from each single poly(A) site gene (in total 173 genes, filtered from 3' READS data (Neve, Burger et al. 2016)) were extracted and applied in the MEME programme. The position of the exact cleavage site in each gene is defined by the 3' READS data (Neve, Burger et al. 2016). The parameters used to calculate motif enrichment are shown in Figure 20A. The

motif length was set between 6nt to 10nt based on the initial observation of several AGUAAA poly(A) sites including *JUND* and *ZIP4* poly(A) sites. After performing MEME, we were able to identify 96 genes (Fig 20C) with the U-rich region (Fig 20B) downstream of AGUAAA upstream of the cleavage site. Also, we set up a control group containing 173 randomly picked poly(A) sites. We made the identical analysis based on the sequence between poly(A) signals and cleavage site of these genes. We were not able to discover any U-rich motifs in the control group (Fig A1). The motif analysis reveals that the U-rich region (Fig 20B), especially the high probability of double-uridine, is present over half of the single AGUAAA poly(A) site containing genes (55.5%). This suggests the general presence of the adU-rich element amongst single AGUAAA poly(A) site.

A

Sequence file	AGUAAAurich.fasta
Distribution of motif occurrences	Zero or one per sequence
Number of different motifs	5
Minimum number of sites	50
Maximum number of sites	160
Minimum motif width	6
Maximum motif width	10
Searching	given strand only

B



C

Summary

E-value	9.1e-04
Width	7
Sites	96

Figure 20 Motif investigation amongst single AGUAAA poly(A) site containing genes

A) The parameter used to search for motifs amongst AGUAAA containing single poly(A) site genes. B) U-rich motif described by MEME amongst all single AGUAAA poly(A) site genes downstream of the hexamer upstream of the cleavage site. Higher nucleotides stand for a higher probability in that position. C) Statistical summary of the U-rich motif given by MEME.

The frequent appearance of the adU-rich element downstream of the AGUAAA hexamer in the single poly(A) site gene suggests that this element may enhance the recognition of the sub-optimal noncanonical poly(A) site. Several protein factors can be the candidate to interact with adU. WDR33 and CPSF30 are reported to directly bind the canonical hexamer signal (Chan, Huppertz et al. 2014, Schönemann, Kühn et al. 2014). Human Fip1 has been shown to bind U-rich elements up/downstream of the hexamer (Kaufmann, Martin et al. 2004). CPSF160 might also be responsible to recognise adU. We propose the model of the recognition of AGUAAA-adU-rich type poly(A) sites (Fig 21). The adU-rich element might function to enhance the interaction between CPSF and pre-mRNA, due to the close proximity between adU and the AGUAAA. And the increase of this interaction might be led by WDR33 (WDR33-adU model), CPSF30 (CPSF30-adU model), CPSF160 (CPSF160-adU model) or hFip1 (hFip1-adU model).

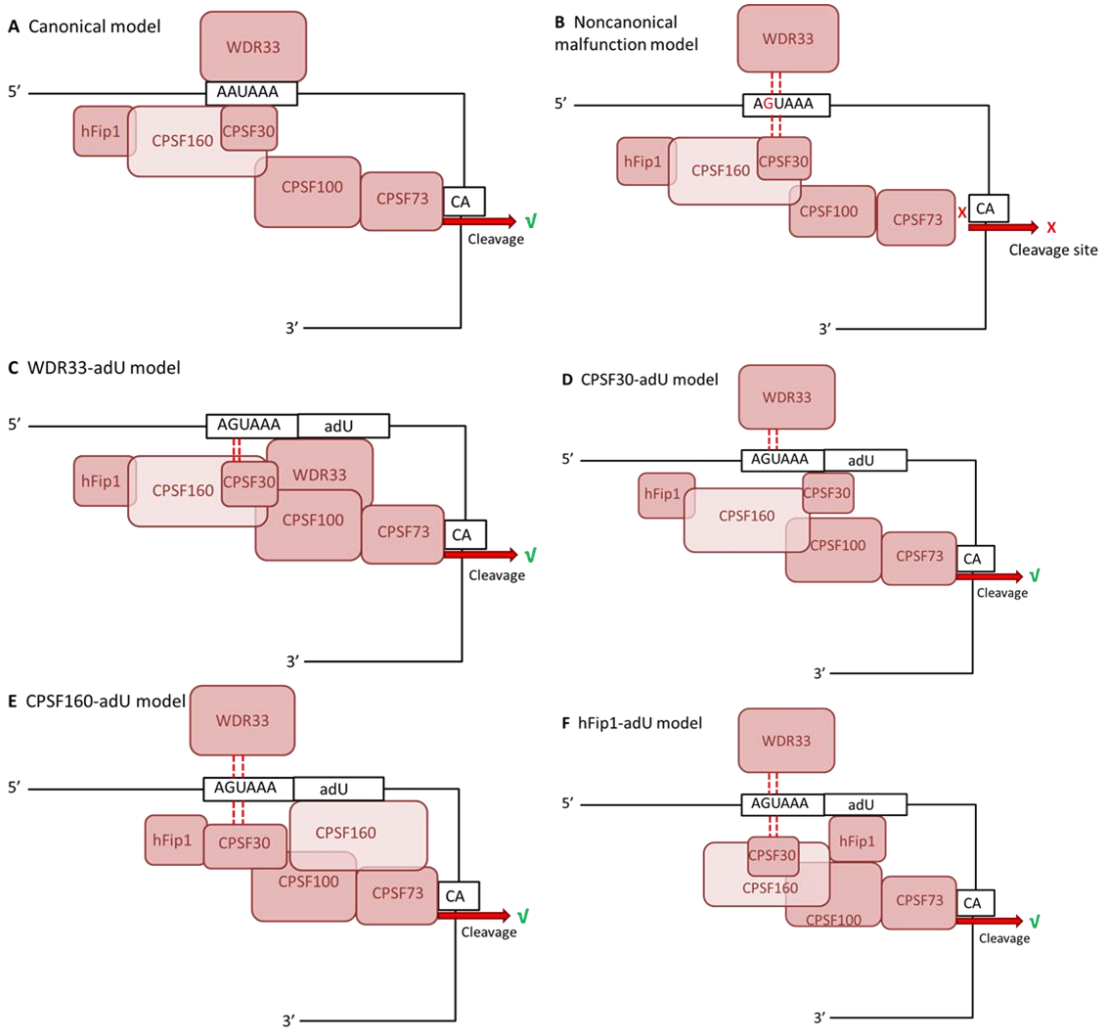


Figure 21 AGUAAA-adU-rich type poly(A) site recognition model

A) Canonical model: the recognition of canonical poly(A) site. The canonical hexamer is recognised by CPSF30 and WDR33 and further recruit other CPSF factors to perform the cleavage reaction. B) Noncanonical malfunction model: if no auxiliary element presents, the AGUAAA hexamer cannot be recognised by CPSF30 or WDR33 and cleavage reaction cannot take place. C) WDR33-adU model: the adU-rich element can be recognised by WDR33 and further recruit other CPSF factors to perform the cleavage reaction. D) CPSF30-adU model: the adU-rich element can be recognised by CPSF30 and further recruit other CPSF factors to perform the cleavage reaction. E) CPSF160-adU model: the adU-rich element can be recognised by CPSF160 and further recruit other CPSF factors to perform the cleavage reaction. F) hFip1-adU model: the adU-rich element can be recognised by hFip1 and further recruit other CPSF factors to perform the cleavage reaction. Red dashed line: weak interaction.

3.6 RNAi based approach to identify protein factors that control AGUAAA poly(A) site recognition

3.6.1 Protein candidates to knockdown

The canonical hexamer AAUAAA/AUUAAA in the pre-mRNA is specifically recognised by CPSF, a protein complex that contains CPSF160, WDR33, CPSF100, CPSF73, hFip1 and CPSF30 (Chan, Huppertz et al. 2014). Despite this, we currently do not know whether the same components are required and involved in the recognition of noncanonical poly(A) sites. It is also not clear whether the whole complex or only part of the complex are required to recognise canonical and noncanonical poly(A) sites. In order to begin to address this issue and to identify protein factors that are required to recognise the AGUAAA type noncanonical poly(A) site in JunD, we employed an RNAi based approach. For this, the following proteins of the CPSF complex were focused on: (i) hFip1: this factor has been reported to interact with U-rich regions positioned up- or downstream of the hexamer signal (Kaufmann, Martin et al. 2004) and would thus be a potential candidate for the interaction with the adU-rich element in the JunD site; (ii) CPSF30 and (iii) WDR33 were recently shown by immunoprecipitation to directly bind to canonical hexamers (Chan, Huppertz et al. 2014); (iv) CPSF160 was originally believed to be the factor that associated with the hexamer but this has now been revised and its function within the CPSF complex remains a mystery (Chan, Huppertz et al. 2014).

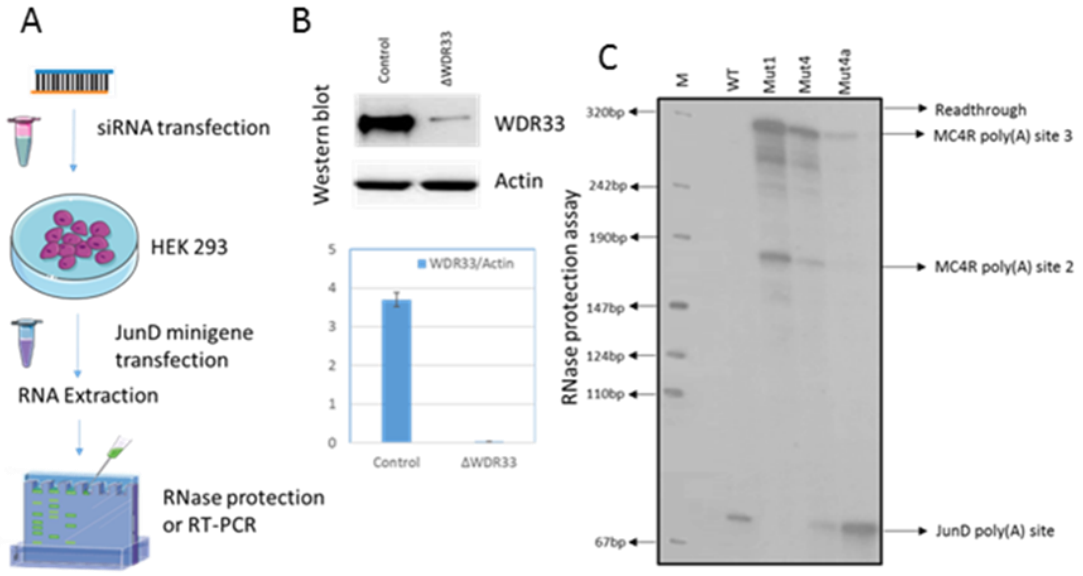
In addition to the CPSF components described above, we also investigated whether other factors play a role in the recognition of AGUAAA poly(A) sites. CFIm has been found to direct the 3’ end formation of a noncanonical poly(A)

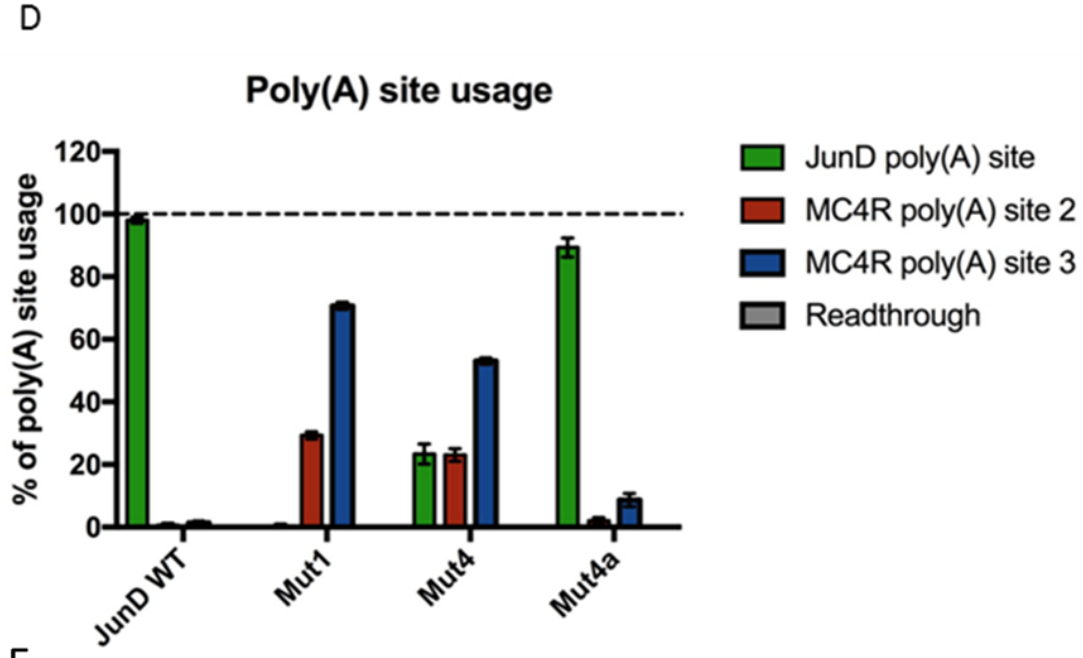
site (Venkataraman, Brown et al. 2005). In collaboration with the Norbury lab, we were able to assess the cleavage and polyadenylation in our MC4R-JunD system in CFIm68 and CFIm25 CRISPR (clustered regularly interspaced short palindromic repeats) “knockdown” cells (3.6.6 and 3.6.7).

3.6.2 The depletion of WDR33 decreased the JunD AGUAAA poly(A) site usage only in the absence of the adU-rich element

Two groups have recently shown that WDR33 directly binds to the canonical hexamer signal AAUAAA during 3’ end processing (Chan, Huppertz et al. 2014, Schönemann, Kühn et al. 2014). It is however not known whether WDR33 also recognises noncanonical sites. In order to answer this question, we applied an RNAi based approach (Fig 22A). We depleted WDR33 by transfecting HEK 293 cells with two hits of siRNAs (see Materials and Methods). The four JunD-MC4R constructs WT, Mut1, Mut4 and Mut4a were transfected as reporter genes into HEK 293 cells, 24h after the second transfection with the siRNA. The total RNA was isolated by the hot phenol method (see Materials and Methods) and at the same time approximately 8 x 10^5 of cells were lysed in Laemmli buffer for subsequent protein fraction analysis to monitor the knockdown efficiency. As verified by Western blot analysis, WDR33 was significantly depleted in the knockdown cells (Fig 22B). To assess its effects on AGUAAA and canonical poly(A) site usage, the same RNase protection approach was used as described above (Fig 17A). The usage of the poly(A) sites were quantified also as described in the previous set of experiments (Fig 17). The rationale behind this approach was that, if the depleted protein factor

is preferentially involved in recognising AGUAAA over AAUAAA poly(A) sites, the usage of the JunD poly(A) site in knockdown cells transfected by the JunD WT and JunD Mut4 plasmids is expected to decrease, compared to the usage ratio in the wildtype control (Fig 17), and vice versa. Compared to the RNase protection analysis from wildtype HEK 293 cells (Fig 22E, Control), the poly(A) site usage in WT, Mut1 and Mut4a (Fig 22D JunD WT, Mut1 and Mut4a, green bars) in WDR33 knockdown cells did not change significantly (Table A1). However, JunD poly(A) site usage in the Mut4 construct was compromised by the WDR33 depletion and reduced from 45% in wildtype results (Fig 22E, Control Mut4) to 24% as shown in Figure 21D Mut4 green bar (and statistically, the JunD poly(A) site usage in Mut4 ΔWDR33 is significantly different from the wildtype control, Table A1). This analysis suggests that WDR33 does not have a preference for either canonical poly(A) site or AGUAAA-adU-rich type poly(A) site. However, at least for JunD, WDR33 appears to be important to boost the recognition of a weak AGUAAA poly(A) site.





E

	JunD WT	Mut1	Mut4	Mut4a
Control	91%	1%	45%	89%
ΔWDR33	98%	0%	24%	89%

Figure 22 The depletion of WDR33 compromised the usage of the sub-optimal JunD poly(A) site

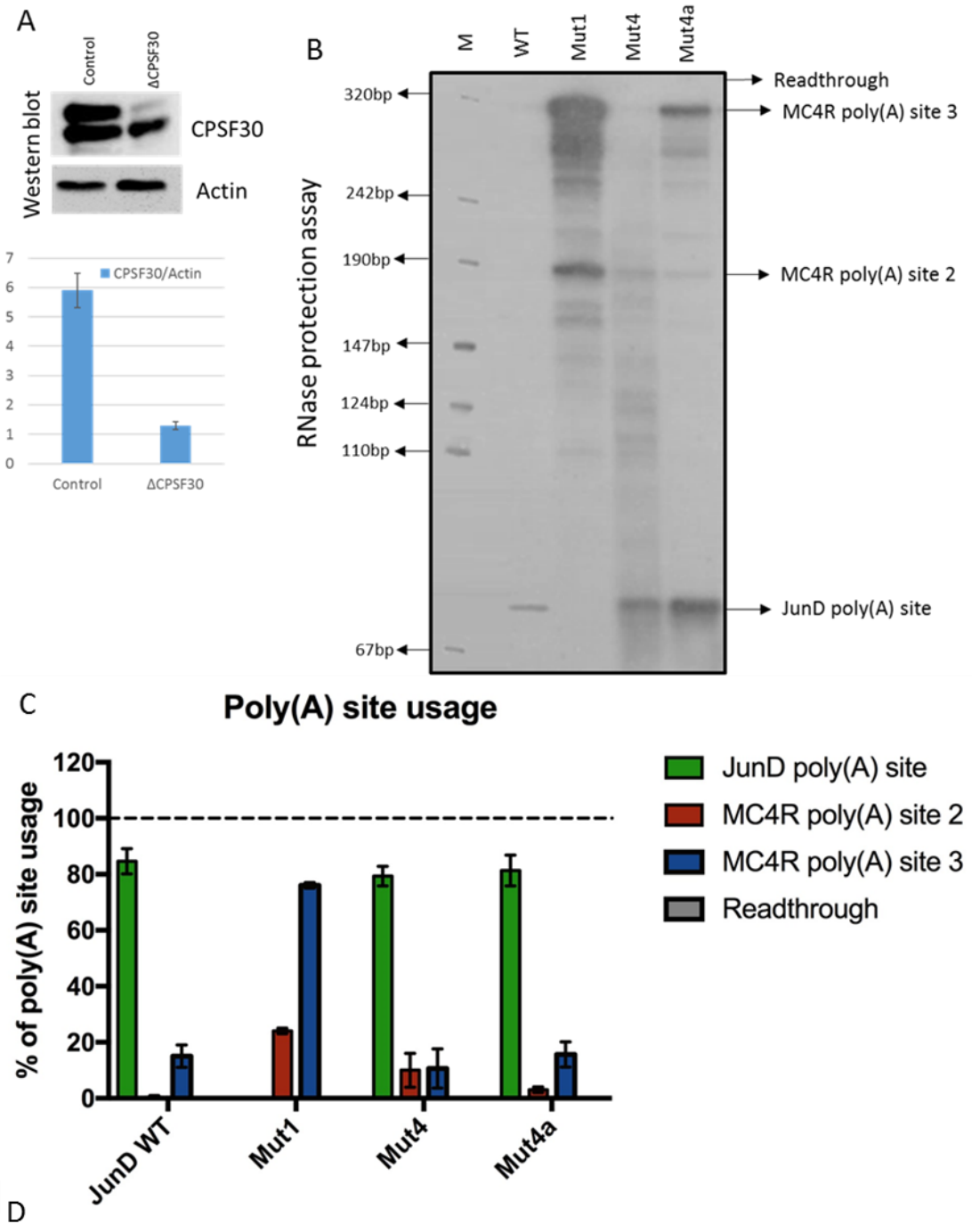
(A) Workflow of the RNase protection assay evaluating the influence of different protein factors. Different protein factors were depleted by transient siRNA transfection. Four

JunD reporters were transfected into the knockdown cell lines, 24h before RNA extraction and subsequent RNase protection assay. (B) Verification of WDR33 knockdown. The same membrane was probed with anti- WDR33 and Actin (loading control) antibody. The control is untreated HEK 293 cells. The ratio of WDR33/Actin is the ratio of respective grey scale calculated by ImageJ. (C) RNase protection assay result as exposed in the phosphorimager. WDR33 depleted HEK 293 cells were transfected with the JunD-MC4R plasmids JunD WT, Mut1, Mut4 and Mut4a. The expected position of each poly(A) signal and read through transcripts is indicated. (D) Quantitative assessment of the RNase protection assay. The usage by ratio of the poly(A) sites is quantified by the same method as described in Fig 16D. Green, red, blue and grey bars represent the usage ratio of JunD poly(A) site, MC4R poly(A) site 2, MC4R poly(A) site 3 and read through transcripts (n=3). (E) JunD poly(A) site usage in all four constructs in WDR33 knockdown cells compared to the JunD poly(A) site usage in HEK 293 control (Control).

3.6.3 The depletion of CPSF30 increased the JunD AGUAAA poly(A) site usage only in the absence of the adU-rich element

We next addressed the role of CPSF30, which is the smallest subunit in CPSF complex, in the recognition of the JunD AGUAAA poly(A) site. Again, whilst CPSF30 has been shown to directly bind to AAUAAA (Chan, Huppertz et al. 2014), it is not known whether it is involved in the recognition of noncanonical hexamers. siRNA was used to deplete HEK 293 cells of CPSF30 (Fig 22A). The usage of different poly(A) sites was monitored by RNase protection. Compared to the results of the RNase protection assay in wildtype HEK 293 cells (Fig 23D Control), the poly(A) site usage in WT, Mut1 and Mut4a did not change significantly (Table A1) upon knockdown of CPSF30 (Fig 23C WT, Mut1 and Mut4a, green bars). However, the relative usage of the JunD poly(A) site was increased from 45% in wildtype to 79% (Fig 23D) in CPSF30 knockdown in the Mut4 construct, that contained the JunD AGUAAA hexamer where the

adU-rich element was mutated. The analysis suggests that CPSF30, similar to WDR33, does not have a particular preference for either AAUAAA poly(A) site or AGUAAA-adU-rich type poly(A) sites. However, in CPSF30 depleted cells, 3' UTR shortening and switching to a weaker noncanonical poly(A) site is seen. Perhaps when CPSF30 levels are reduced, the formation of an otherwise minor CPSF sub-complex is enhanced which may be able to recognise and favour weak AGUAAA type poly(A) sites. In conclusion, our results show that the depletion of CPSF30 causes a shift from a canonical poly(A) site to a weak noncanonical poly(A) site.



D

	JunD WT	Mut1	Mut4	Mut4a
Control	91%	1%	45%	89%
ΔCPSF30	85%	0%	79%	81%

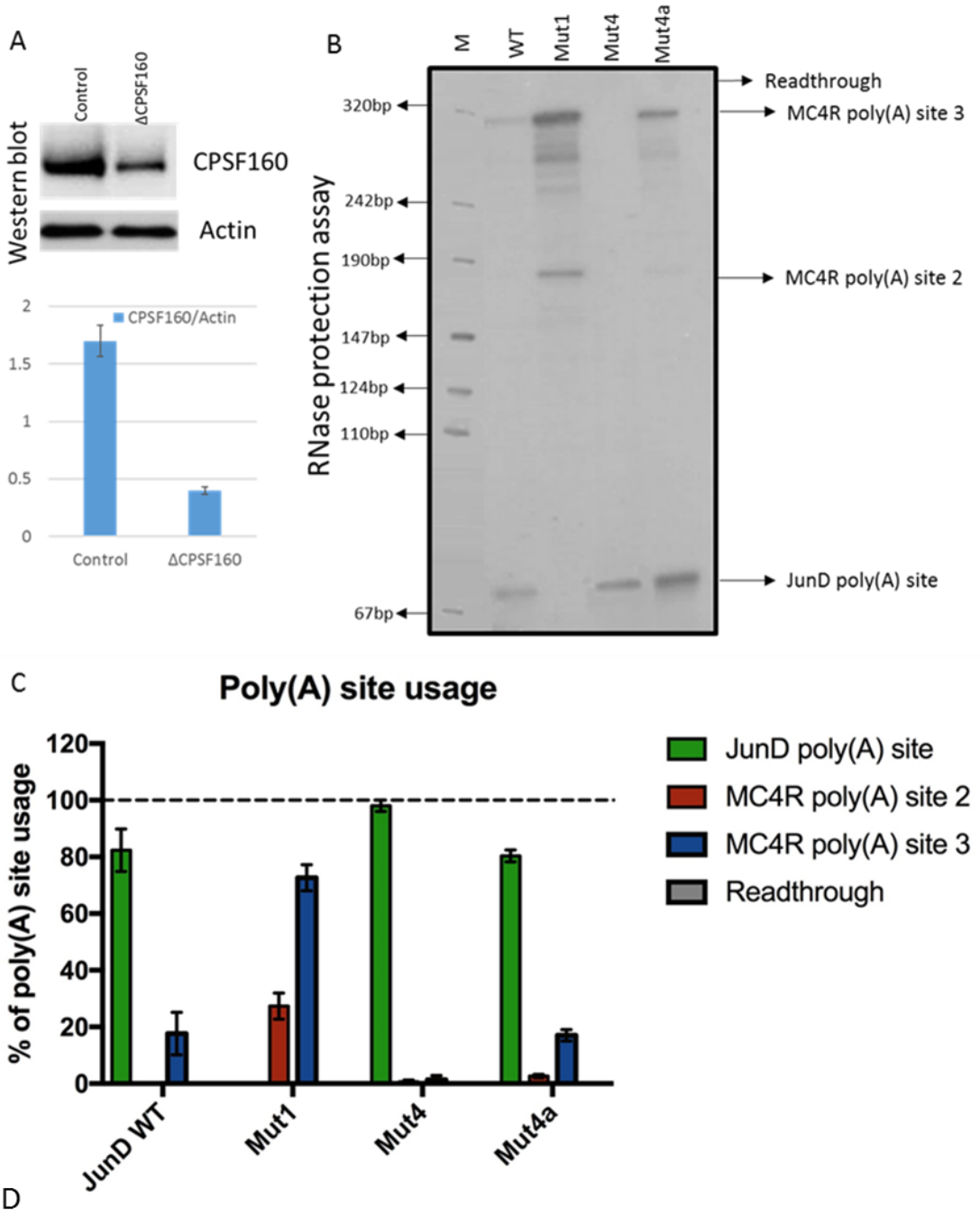
Figure 23 The depletion of CPSF30 increased the usage of the sub-optimal JunD poly(A) site

(A) Verification of CPSF30 knockdown by Western blot analysis. The same membrane was probed with anti- CPSF30 and Actin (loading control) antibody. The control refers to untreated HEK 293 cells. The ratio of CPSF30/Actin is the ratio of respective grey scale calculated by ImageJ. (B) RNase protection assay result exposed in phosphorimager. HEK 293 cells were transfected by the identical JunD-MC4R plasmids WT, Mut1, Mut4 and Mut4a. The expected position of each poly(A) signal and read through transcripts is labelled. (C) Quantitative assessment of the RNase protection assay. The poly(A) site usage by ratio of different poly(A) sites were quantified by the same method as described in Fig 16D. Green, red, blue and grey bars represent the usage ratio of JunD poly(A) site, MC4R poly(A) site 2, MC4R poly(A) site 3 and read through transcripts (n=3). (D) JunD poly(A) site usage in all four constructs in CPSF30 knockdown cells compared to the JunD poly(A) site usage in HEK 293 control (Control).

3.6.4 The depletion of CPSF160 increased the JunD AGUAAA poly(A) site usage only in the absence of the adU-rich element

CPSF160 along with WDR33 are the two largest subunits in CPSF complex (Shi, Giammartino et al. 2009). Recombinant CPSF160 was initially discovered to bind AAUAAA *in vitro* (Colgan and Manley 1997). However, its current role in hexamer recognition is not known (Chan, Huppertz et al. 2014). CPSF160 is likely to function as a scaffold and is important to recruit other factors to the complex (Murthy and Manley 1995, Chan, Huppertz et al. 2014). To test whether it is involved in the recognition of noncanonical poly(A) sites, we depleted CPSF160 in HEK 293 cells (Fig 24A). As can be seen from the RNase protection results (Fig 24C green bars), the poly(A) site usage in WT, Mut1 and Mut4a did not change significantly in CPSF160 depleted cells compared to the wildtype control (Fig 24D, Table A1). Nevertheless, similar to the depletion of CPSF30, JunD poly(A) site usage was increased to 98% from 45% in Mut4 (Fig

24D) after the depletion of CPSF160. The analysis suggests that CPSF160 does not have a preference for either AAUAAA poly(A) site or AGUAAA-adU-rich type poly(A) sites. However, it appears that if a weak noncanonical site is presented upstream of a strong canonical site, high levels of CPSF160 favour distal site usage, but when CPSF160 is depleted, switching to the weaker upstream poly(A) site is observed.



D

	JunD WT	Mut1	Mut4	Mut4a
Control	91%	1%	45%	89%
ΔCPSF160	82%	0%	98%	80%

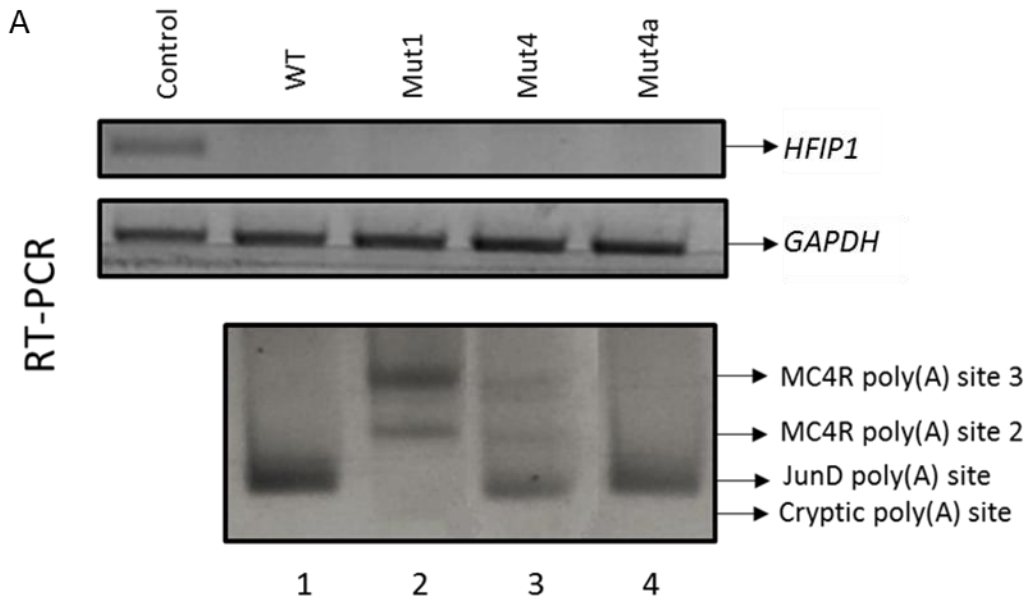
Figure 24 The depletion of CPSF160 increased the usage of the sub-optimal JunD poly(A) site

(A) Verification of CPSF160 knockdown by Western blot analysis. The same membrane was probed with anti- CPSF160 and Actin (loading control) antibody. The control represents untreated HEK 293 cells. The ratio of CPSF160/Actin is the ratio of respective grey scale calculated by ImageJ. (B) RNase protection assay result exposed in phosphorimager. HEK 293 cells depleted of CPSF160 were transfected by the identical JunD-MC4R plasmids WT, Mut1, Mut4 and Mut4a. The expected position of each poly(A) signal and read through transcript is labelled. (C) Quantitative assessment of the RNase protection assay. The usage by ratio of the poly(A) sites are quantified by the same method as described in Fig 16D. Green, red, blue and grey bars represent the usage ratio of JunD poly(A) site, MC4R poly(A) site 2, MC4R poly(A) site 3, and read through transcripts (n=3). (D) JunD poly(A) site usage in all four constructs in CPSF160 knockdown cells compared to the JunD poly(A) site usage in HEK 293 control (Control).

3.6.5 The depletion of hFip1 increased the JunD AGUAAA poly(A) site usage only in the absence of the adU-rich element

Human Fip1 (hFip1) is a subunit of CPSF complex. It has been shown to interact with U-rich elements up- or downstream of the hexamer (Kaufmann, Martin et al. 2004) and might thus be a factor that controls the AGUAAA-adU-rich type poly(A) site recognition. To test this, hFip1 was depleted in HEK 293 cells (Fig 25A, top panel) and analyzed its effect on poly(A) site usage in four JunD-MC4R plasmids by RT-PCR (Fig 25A, bottom panel). Compared to the RT-PCR result in the wildtype HEK 293 cells, no obvious change in the ration of the poly(A) site usage can be found in JunD WT, Mut1 and Mut4a. Specifically, the function of JunD poly(A) site in WT (Fig 25A, bottom panel, lane 1), which is an AGUAAA-adU-rich type poly(A) site, is not compromised by the depletion of hFip1. However, JunD poly(A) site usage was increased significantly in Mut4 (Fig 25B) after the depletion of hFip1. The analysis

suggests that hFip1 does not have a preference for either AAUAAA poly(A) site or AGUAAA-adU-rich type poly(A) sites. However, it appears that if a weak noncanonical site is presented upstream of a strong canonical site, high levels of hFip1 favour distal site usage, but when hFip1 is depleted, switching to the weaker upstream poly(A) site is observed.



B

	JunD WT	Mut1	Mut4	Mut4a
Control	98%	0%	47%	99%
ΔhFip1	100%	0%	72%	99%

Figure 25 The impact of hFip1 on JunD-MC4R 3' end formation

(A) (i) Top panel: Knockdown levels were verified by RT-PCR analysis targeting the terminal exon region of *HFip1* (*FIP1L1*) mRNA. The control refers to the untreated HEK 293 cells. (ii) Bottom panel: The identical RT-PCR method (Fig 18A) was applied to evaluate JunD-MC4R polyadenylation. HEK 293 cells depleted of hFip1 were transfected by the identical JunD-MC4R plasmids WT, Mut1, Mut4 and Mut4a. The expected position of each poly(A) signals is labelled. (B) JunD poly(A) site usage in all four constructs in hFip1 knockdown cells compared to the JunD poly(A) site usage in

HEK 293 control (Control). The absolute usage value of the poly(A) sites in RT-PCR were quantified by respective grey scale calculated by ImageJ software. The JunD poly(A) site usage ratio was calculated by (JunD poly(A) site)/(JunD poly(A) site + MC4R poly(A) site 2 + MC4R poly(A) site 3 + Cryptic poly(A) site). (n=3).

3.6.6 The depletion of CFIm68 abolished the most distal poly(A) site and generated read through transcripts

The interaction of the CFIm protein complex with the pre-mRNA is one of the earliest steps in the assembly of the 3' end processing complex (Brown and Gilmartin 2003) and facilitates the recruitment of other processing factors. This complex is composed of four polypeptides: CFIm25, CFIm68, CFIm59 and CFIm72. CFIm68 is a 68 kDa subunit. CFIm protein complex recognises UGUAN motif upstream of the hexamer mainly by its CFIm68 and CFIm25 subunits and is suggested to recruit CFIIm afterwards (Yang, Gilmartin et al. 2011). Importantly, CFIm has originally been identified as a factor that is critical for the recognition of noncanonical poly(A) sites (Venkataraman, Brown et al. 2005). Several groups demonstrated that the loss-of-function of CFIm68 and CFIm25 leads to a transcriptome-wide increase in the use of proximal polyadenylation sites during alternative cleavage and polyadenylation events in mammalian cells (Gruber, Martin et al. 2012, Martin, Gruber et al. 2012, Masamha, Xia et al. 2014). In collaboration with the Norbury Lab, we investigated the impact of CRISPR “knockdown” of CFIm68 on JunD-MC4R cleavage and polyadenylation. HEK 293 CRISPR “knockdown” cell lines, provided by Jessica Hardy (DPhil student in Chris Norbury Lab), were generated by co-transfection of Cas9 reporter and guide RNA targeting exon 3 of the CFIm68 gene. The cells in depletion of CFIm68 were selected by

puromycin. We call them “knockdown” cells because there is a residual amount of CFIm68 remaining, as seen by western blotting (Fig 26A), likely to be due to the presence of one wildtype allele (out of around 5 copies). Four JunD-MC4R constructs (WT, Mut1, Mut4 and Mut4a) were transfected into these cells and the resulting RNA was applied to RT-PCR based analysis described previously (Fig 14A). Interestingly, a general 3’ UTR shortening effect is apparent whereby in all constructs there is little, if any, usage of the most distal MC4R poly(A) site (Fig 26B, MC4R poly(A) site 3). This 3’ UTR shortening effect would be in line with previous reports. To test whether the disappearance of the most distal site is due to increased usage of the proximal sites, we subjected the RNA to quantitative assessment using the RNase protection approach.

The RNase protection analysis confirmed a dramatic drop of the most distal site, MC4R poly(A) site 3, in JunD Mut1 and Mut4 (while the MC4R poly(A) site 3 usage in JunD WT and Mut4a is already low in the wildtype, Fig 17D). Importantly, an additional longer band that represents transcripts that read through the MC4R poly(A) site 3 become visible (Fig 26C, Readthrough), and the usage ratio of the read through in WT, Mut1 and Mut4 (Fig 26D) is significantly different from the wildtype control (Table A1). If the disappearance of the most distal site is due to a switch and preferential usage of the more proximal sites, no read through should be detected. On the other hand, if depletion of CFIm68 does not cause a switch in usage from distal to proximal poly(A) site, but instead somehow inactivates distal sites, a read through transcript would be expected (Fig 26E, F). Therefore, if this is the case, then the impact of the knockdown of CFIm68 on the JunD-MC4R 3’ end formation cannot be described as 3’ UTR shortening. It is worth noting that, compared to

JunD WT, there is no read through transcript detected in Mut4a (Fig 26 F). This indicates that, the canonical hexamer mutant Mut4a can direct JunD 3' end processing efficiently independent of CFIm68 whilst the function of AGUAAA-adU-rich type poly(A) site is reliant on CFIm68.

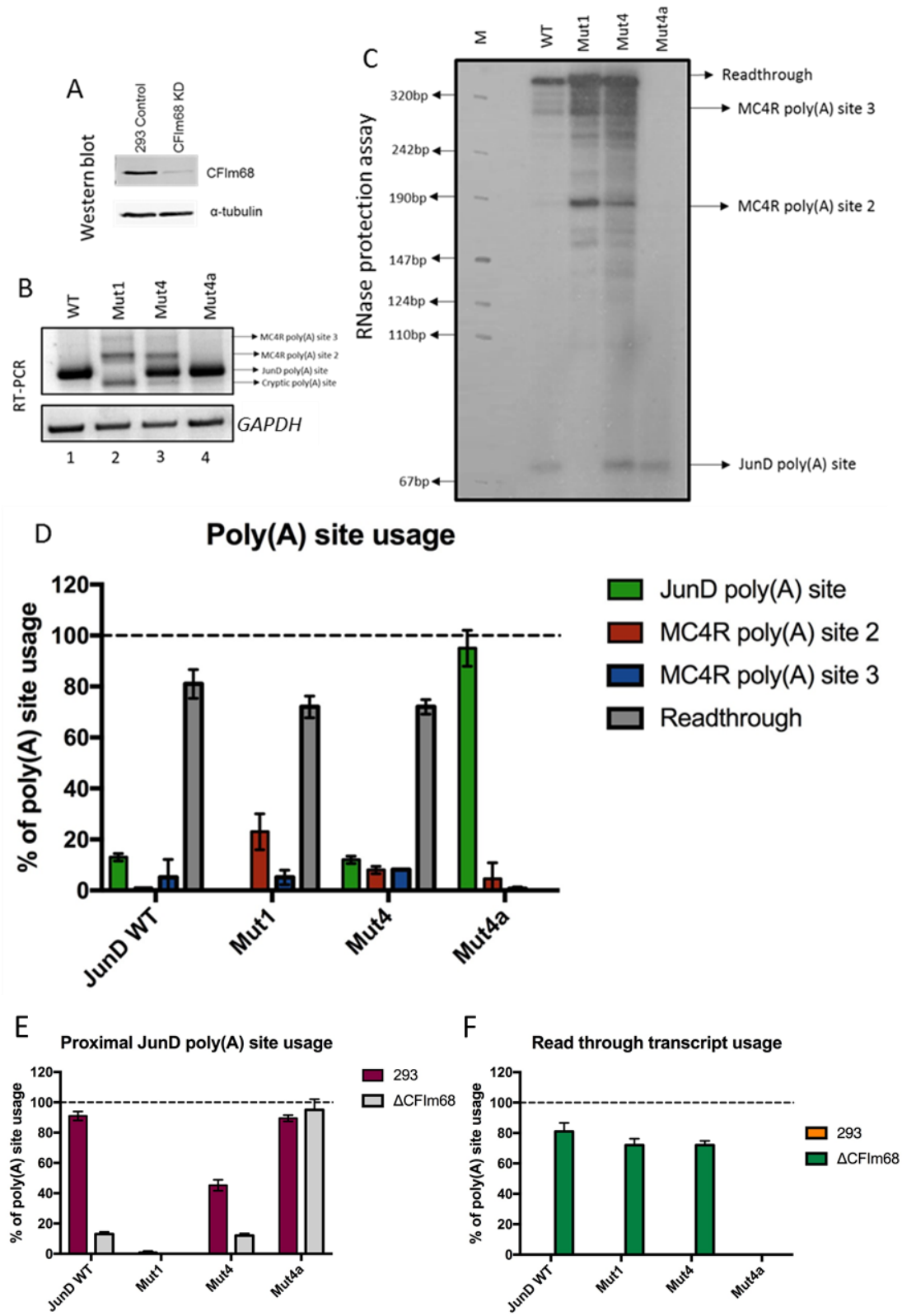


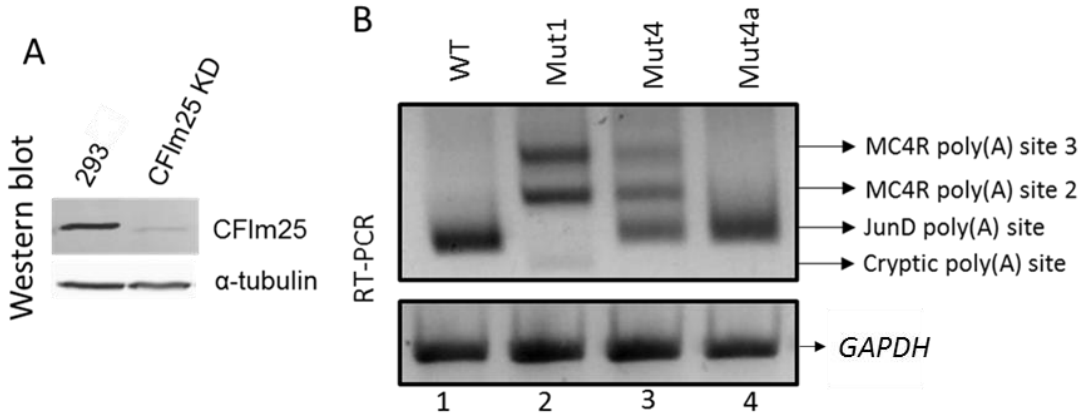
Figure 26 The depletion of CFIm68 abolished the most distal poly(A) site and generated read through transcripts

(A) Verification of CFIm68 knockdown by Western blot analysis. The same membrane was probed with anti- CFIm68 and α-tubulin (loading control) antibody. The control is

untreated HEK 293 cells. (Cells and Western blot analysis provided by Jessica Hardy). (B) RT-PCR analysis of JunD and MC4R polyadenylation in CFIm68 knockdown cells. The identical method (Fig 18A) was applied. The expected position of different poly(A) signals is as labelled. (C) RNase protection assay result exposed in phosphorimager. HEK 293 cells were transfected by four identical JunD-MC4R plasmids JunD WT, Mut1, Mut4 and Mut4a. The expected position of each poly(A) signal and read through transcripts is indicated. (D) Quantitative assessment of the RNase protection assay. The usage ratio of the poly(A) sites and read through transcripts are quantified by the same method as described in Fig 16D. Green, red, blue and grey bars represent the usage ratio of JunD poly(A) site, MC4R poly(A) site 2, MC4R poly(A) site 3 and read through transcripts (n=2). (E, F) The comparison of the JunD poly(A) site usage ratio and read through transcripts ratio between the wildtype and CFIm68 depleted data from RNase protection assay results in Fig 16D and Fig 25D. The wildtype control is indicated as 293.

3.6.7 The depletion of CFIm25 did not influence JunD-MC4R 3' end formation

The influence of CFIm25 was analyzed by the same RT-PCR approach as described above (Fig 14A). CFIm25 CRISPR knockdown cells were again provided by the Norbury lab (Fig 27A). Unlike for CFIm68, no obvious change was visualised by RT-PCR (Fig 27B), compared to the wildtype control (Fig 27C).



C

	JunD WT	Mut1	Mut4	Mut4a
Control	98%	0%	47%	99%
ΔCFIm25	100%	0%	42%	100%

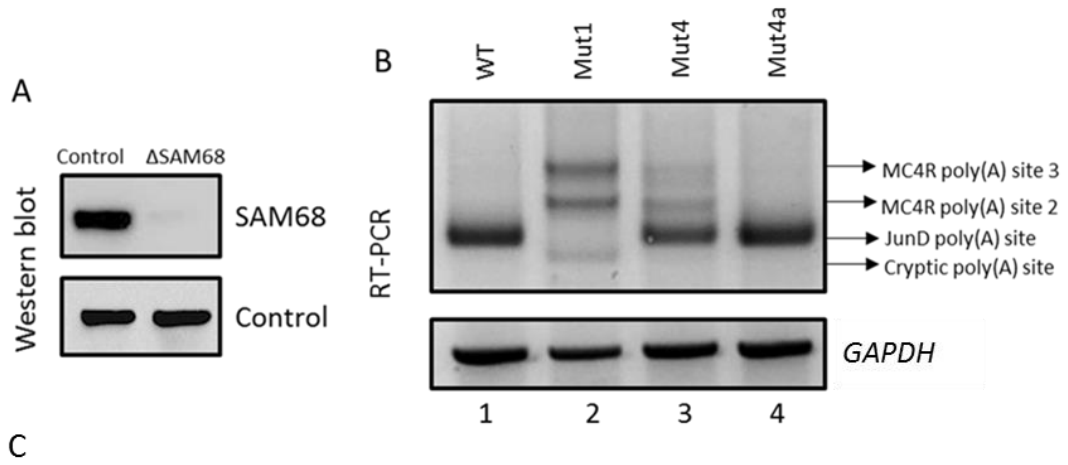
Figure 27 The impact of CFIm25 on JunD-MC4R 3' end formation

(A) Verification of CFIm25 knockdown. The same membrane was probed with anti-CFIm25 and α-tubulin (loading control) antibody. The control refers to untreated HEK 293 cells. (Cells and Western blot analysis provided by Jessica Hardy). (B) RT-PCR analysis of JunD and MC4R polyadenylation in CFIm25 depleted cells. The same method (Fig 13A) was adopted. The expected position of different poly(A) signals is as labelled. (C) JunD poly(A) site usage in all four constructs in CFIm25 knockdown cells compared to the JunD poly(A) site usage in HEK 293 control (Control). The quantitative method is described in Figure 24B. (n=3).

3.6.8 The depletion of SAM68 increased the JunD AGUAAA poly(A) site usage only in the absence of the adU-rich element

The *SAM68* gene encodes a member of the K homology domain-containing, RNA-binding, signal transduction-associated protein family (Lukong and Richard 2003). The encoded protein appears to have many functions and may be involved in a variety of cellular processes, including alternative splicing

(Paronetto, Achsel et al. 2007), cell cycle regulation (Barlat, Maurier et al. 1997), viral RNA 3' end formation (McLaren, Asai et al. 2004), tumorigenesis (Lazer, Pe'er et al. 2007), and regulation of human immunodeficiency virus (HIV) gene expression (Reddy, Xu et al. 1999, Modem, Badri et al. 2005). The protein has been found to possess the affinity to UAAA or UUAA RNA sequence (Lin, Taylor et al. 1997, Itoh, Haga et al. 2002). As its recognition site overlaps with canonical poly(A) hexamer and AGUAAA sequences, we hypothesised that SAM68 might be a factor that could participate in mammalian 3' end processing. The global role of this protein will be addressed in the next chapter. As a pilot experiment, the impact of low levels of SAM68 on JunD-MC4R polyadenylation was assessed by the same strategy described in Figure 22A. SAM68 was transiently depleted in HEK 293 cells using siRNAs. After the verification of the knockdown effect of SAM68 by Western blot (Fig 28A), the poly(A) site usage in the different JunD reporter genes was evaluated using our RT-PCR approach as described (Fig 28B). Compared to the RT-PCR result in wildtype HEK 293 cells (Fig 28C), no obvious change of the poly(A) site usage by ratio was detected for JunD WT, Mut1 and Mut4a upon the depletion of SAM68. However, the JunD poly(A) site usage in Mut4 is significantly increased after the knockdown of SAM68. From this, we suggest that SAM68 participates in JunD-MC4R cleavage and polyadenylation and prefers canonical poly(A) sites to the sub-optimal noncanonical poly(A) sites. Nevertheless, since both the AGUAAA and AUUAAA poly(A) sites contain the UAAA SAM68 binding sequence, the influence of SAM68 on Mut4 cannot be explained by its UAAA affinity.



C

	JunD WT	Mut1	Mut4	Mut4a
Control	98%	0%	47%	99%
ΔSAM68	100%	1%	68%	100%

Figure 28 The impact of SAM68 on JunD-MC4R 3' end formation

(A) Verification of SAM68 knockdown by Western blot analysis. The same membrane was probed with anti- SAM68 and Actin (loading control) antibody. The control refers to untreated HEK 293 cells. (B) The identical RT-PCR method (Fig 13A) was performed to evaluate JunD and MC4R polyadenylation. The expected position of different poly(A) signals is as labelled. (C) JunD poly(A) site usage in all four constructs in SAM68 knockdown cells compared to the JunD poly(A) site usage in HEK 293 control (Control). The quantitative method is described in Figure 24B. (n=3).

3.7 Discussion

We started the investigation of noncanonical AGUAAA poly(A) site recognition by analysing the RNA secondary structure environments, surrounding the hexamer signal, by employing the Oxfold algorithm (Anderson, Haas et al. 2013). From this *in silico* study, we were able to extract potential functionally relevant stem loop structures by using a sophisticated "forecasting algorithm" for AGUAAA containing poly(A) sites (Fig 10B). However, the close inspection

of one such example, JunD, showed that even if there is a high probability of a secondary structure (Fig 12B), disruption of it had little, if any, influence in poly(A) site usage (Fig 15C, lane 4). This indicates that, trying to address the involvement of structure in poly(A) site recognition at a global scale may produce limited meaningful results.

To study AGUAAA poly(A) site recognition further, the sequences necessary for 3' end processing in JunD were dissected by combining the Mini-gene approach and the mutagenesis analysis. Fewer than 50 nucleotides of JunD proved sufficient to direct cleavage and polyadenylation, which suggests noncanonical poly(A) sites, such as AGUAAA site, can be fully functional as a very short fragment (Fig 14C, lane 1). This is surprising as *JUND* is an intronless noncanonical poly(A) site gene which generally has been shown to rely on an array of auxiliary elements for 3' end processing (Huang and Carmichael 1997, Conrad and Steitz 2005, Dalziel, Nunes et al. 2006). This might explain why a seemingly equivalent region including the AGUAAA hexamer and a potential DSE from the *ZIP4* gene was not sufficient to direct processing at this minimal noncanonical site (Fig 19B, lane 1). However, what makes the Zip4 result intriguing is the fact that whilst the wildtype architecture of the poly(A) site is inactive within the 51-nucleotide context, it can be activated if a potential adU-rich element, defined in JunD mutagenesis analysis, is positioned immediately adjacent to the AGUAAA hexamer (Fig 19B, lane 2). The adU-rich element together with the AGUAAA hexamer and the DSE have been shown to be essential for the optimal JunD poly(A) site usage (Fig 15C). Thus, it is suggested that, the recognition of the AGUAAA hexamer can be boosted, if a U-rich region is present immediately adjacent to the hexamer. In

the example of MC4R major poly(A) site (deleted in the plasmid described in Fig 9) (Nunes, Li et al. 2010), the deletion of AUUAAA or the upstream A-rich sequence does not affect 3' end formation. The function of this poly(A) site is abolished only when both of them are mutated. Strikingly, the sequence similar to the adU-rich element (Fig 9, "GUCCUU") is present adjacent to the A-rich sequence in *MC4R* which deepens our understanding of this highly unusual poly(A) site. This suggests that the A-rich-adU-rich region and the AUUAAA are two potential poly(A) signals in MC4R poly(A) site. Cleavage and polyadenylation appear to be equally and efficiently directed by both of these sequence elements. This implies that a noncanonical poly(A) signal can functionally substitute for the canonical hexamer with the aid of adjacent U-rich element and direct cleavage and polyadenylation.

From the individual AGUAAA poly(A) site study, we identified the necessary *cis* elements that are sufficient and essential to drive 3' end formation in the *JUND* gene. As expected, these elements include the AGUAAA hexamer and a short DSE. However, we have shown that a novel auxiliary adU-rich element that is immediately adjacent to AGUAAA is required for full function of this site. The experiments involving Zip4 further suggest that the positioning of a U-rich region immediately adjacent to a noncanonical hexamer can activate an otherwise dormant poly(A) site. A similar U-rich region, found adjacent to the A-rich region in *MC4R*, might contribute to the recognition of 3' end processing machinery. This further illustrates the complexity of poly(A) site recognition and the difficulty to assign a particular strength to a poly(A) site based on sequence composition alone. This limitation makes the interpretation of alternative cleavage and polyadenylation profiles more difficult as noncanonical sites are

generally considered to be weak. However, the example of JunD shows this is not necessarily the case.

In the motif analysis, the MEME programme describes the presence of a U-rich motif between the hexamer and the cleavage point amongst single AGUAAA poly(A) site genes (Fig 20B). The core part of the motif is the double-uridine. The motif can be found downstream of *JUND* and *ZIP4* AGUAAA hexamer. In the control group which contains randomly picked 173 poly(A) sites, we failed to identify any similar U-rich motif (Fig A1). This study indicates that the sub-optimal hexamer AGUAAA is often associated with the auxiliary adU-rich element. Thus, the adU-rich element may be a common enhancer to ensure efficient recruitment of the 3' end processing machinery at many AGUAAA poly(A) sites. Interestingly, the sequence of the adU-rich region found in *MC4R* poly(A) site (Nunes, Li et al. 2010) meets the motif sequence requirement described by MEME (except the seventh nucleotide which appears to be versatile). This implies that the adU-rich element might not only assist the recognition of AGUAAA poly(A) site but also other poly(A) site with degenerated hexamers. The global importance of the U-rich region dissected in *JUND* and *ZIP4* pre-mRNA is suggested.

To address AGUAAA poly(A) site recognition from the perspective of trans-factors, we investigated the effects caused by the depletion of different protein factors involved in 3' end processing on different types of poly(A) sites in our system. This is the first study to assess the role of WDR33, CPSF30, CPSF160, hFip1, CFIm68, CFIm25 and SAM68 on noncanonical AGUAAA type poly(A) site recognition.

WDR33, CPSF30, CPSF160 and hFip1 are subunits of CPSF complex (Chan, Huppertz et al. 2014), which is an essential protein complex for both the cleavage and polyadenylation reaction. These protein factors were depleted transiently in HEK 293 cells by siRNA transfection (Fig 22A). hFip1 has been shown to interact with U-rich enhancers surrounding the poly(A) site (Kaufmann, Martin et al. 2004). However, since the efficient poly(A) site usage of JunD WT, an AGUAAA-adU type poly(A) site, is independent of hFip1 (Fig 25A, bottom panel, lane 1), it is unlikely that the adU-rich element in JunD poly(A) site is specifically recognised by hFip1. WDR33 and CPSF30 have been shown to interact with the canonical poly(A) signal (Chan, Huppertz et al. 2014, Schönemann, Kühn et al. 2014). CPSF160 coordinates the interaction of CPSF complex with other poly(A) protein complexes (Murthy and Manley 1995). In the quantitative RNase protection analysis, AAUAAA and AGUAAA-adU-rich type poly(A) sites are indicated to be recognised equally well by these three protein factors as well as hFip1 in RT-PCR (considering the fact that the usage ratio of JunD poly(A) site in WT and Mut4a is not significantly different from each other in the four protein factors depleted results, Table A1). Interestingly, WDR33 favours the noncanonical poly(A) sites with sub-optimal *cis* element composition over the AAUAAA or AGUAAA-adU-rich type poly(A) sites (Fig 22E). In contrast, CPSF30, CPSF160 and hFip1 have the opposite impact regarding the preference on types of poly(A) sites (Fig 23D, Fig 24D and Fig 25B). Since the different dependencies on the CPSF subunits raise the intriguing possibility that different sub-complexes may be formed that cater specifically for the cleavage or the polyadenylation steps respectively (Schönemann, Kühn et al. 2014), the preference of these three CPSF subunits

on different types of the poly(A) sites might also be caused by the formation of different CPSF sub-complexes when the respective protein levels decrease.

We also addressed a potential role of CFIm subunits in JunD-MC4R 3' cleavage and polyadenylation. The RNase protection analyzes the 3' end formation in JunD-MC4R reporter genes when transfected in CRISPR depleted CFIm68 cells. However, what was previously interpreted as activation of the proximal poly(A) sites over distal sites (Martin, Gruber et al. 2012), our analysis provides an alternative interpretation why a steady state towards the proximal site is observed upon CFIm68 knockdown. In our example, compared to the wildtype control (Fig 17D), we do not see an increase in proximal site usage (Fig 26E) but rather a decrease of distal site usage evidenced by the increased read through (Fig 26F) at this site. This read through increases the relative abundance of transcripts cleaved at the proximal site because few transcripts are cleaved at the distal site. This is a significant finding as it contravenes the current belief that CFIm depletion causes 3' UTR shortening by switching poly(A) site usage from distal to proximal sites. We show that this, at least in some cases, might be a misinterpretation. Since the oligo dT priming approach to establish the library fails to capture the read through transcripts, in collaboration with the Norbury Lab, we are currently assessing whether increased read through level is a global phenomenon caused by low levels of CFIm68. Interestingly, 3' end formation in JunD Mut4a is not affected by the low levels of CFIm68 as no read through transcript is detectable. It has been shown that CFIm can function as a primary determinant of poly(A) site recognition in the absence of the canonical hexamer (Venkataraman, Brown et al. 2005). This might explain that the UGUAN motif in 3' UTR of MC4R upstream

of the JunD fragment contributes to the 3’ end formation of JunD AGUAAA poly(A) site. However, the 3’ end formation of the canonical poly(A) site JunD Mut4a is CFIm68 independent.

Interestingly, although depletion of CFIm25 has widely been reported to cause 3’ UTR shortening (Kubo, Wada et al. 2006, Kim, Yamamoto et al. 2010, Masamha, Xia et al. 2014), here CFIm25 depletion, in contrast to CFIm68, has no obvious effect on JunD-MC4R 3’ end formation as shown in RT-PCR analysis (Fig 27C). This suggests distinct impacts of both the two subunits from CFIm complex on some poly(A) sites. Alternatively, the level of CFIm25 depletion in the CRISPR cells is not severe enough to cause an effect.

SAM68 is a protein factor that binds A/U rich sequences (Lin, Taylor et al. 1997, Itoh, Haga et al. 2002). Our knockdown and RT-PCR experiments show that, similar to CPSF30, CPSF160 and hFip1, depletion of SAM68 only increases the usage of the sub-optimal JunD poly(A) site but does not have preference between canonical sites and the AGUAAA-adU JunD poly(A) site (Fig 28C). This suggests that SAM68 participates in the 3’ end processing, at least in the MC4R-JunD Mut4 construct.

Chapter 4

The global role of WDR33, CPSF30 and SAM68 on canonical and noncanonical poly(A) site usage

Chapter 4

4.1 Introduction

We have analysed the different roles of a number of protein trans-factors for JunD-MC4R 3' end formation in a Mini-gene context in Chapter 3. To explore a potential global influence of WDR33, CPSF30 and SAM68 on the canonical and noncanonical poly(A) site recognition, we employed a whole transcriptome 3' end sequencing based approach in combination with RNAi (Fig 29A). Nuclear RNA was isolated after double-transfection with the siRNAs that target the trans-factor mRNA. The libraries were generated by Oligo dT based synthesis of cDNA followed by PCR amplification.

4.2 The global role of hexamer binding protein WDR33 on canonical and noncanonical poly(A) site usage

Results from two different laboratories have shown that WDR33, a protein component of CPSF complex, directly binds to the canonical hexamer AAUAAA (Chan, Huppertz et al. 2014, Schönemann, Kühn et al. 2014). It is now believed to be the core RNA interacting part of the CPSF complex. In the individual gene study described above, we have shown that depletion of WDR33 decreases the usage of the noncanonical AGUAAA poly(A) site when the *cis* elements are sub-optimal, such as JunD Mut4 (Fig 22E). To address the global role of

WDR33 on the different types of poly(A) sites, we performed RNA sequencing on RNA isolated from cells that have depleted WDR33 levels.

WDR33 transient knockdown was generated by double transfection of siRNA (see Materials and Methods). The control cells were transfected by siRNA negative control in the same approach. RNA and protein were isolated at the same time. Before library preparation, as indicated above, we verified the fractionation quality of RNA by detecting nuclear-retained noncoding RNA *MALAT1* (Tripathi, Ellis et al. 2010) in RT-PCR analysis (Fig 29C, RT-PCR) and visualising the precursor ribosomal RNA on a denaturing RNA gel (Fig 29C, RNA denaturing gel). WDR33 knockdown efficiency was verified by Western blot analysis (Fig 29B). Following the initial quality control step, libraries were prepared using the Lexogen QuantSeq kit. Library quality and concentration was assessed by Agilent Bioanalyzer (Fig 29D). The libraries were sequenced on the Ion Proton sequencer. The sequencing data was processed through the algorithm developed by Bin Tian, our collaborator at Rutgers Cancer Institute of New Jersey (Hoque, Ji et al. 2012). A third quality control step was applied after sequencing by confirming the RNA-seq result of one particular gene *CDCA8* by RT-PCR. According to the RNA-seq data, the proximal poly(A) site usage of *CDCA8* was increased in WDR33 depletion cells. The RT-PCR result was consistent with the sequencing data (Fig 29E).

The global influence of WDR33 on APA was measured and produced by the algorithm developed by Tian Lab. Knockdown of WDR33 shows a modest effect on either UTR-APA (4% significant APA change events compared to the negative control) or CR-APA (2% significant APA change events compared to the negative control) (Fig 33A).

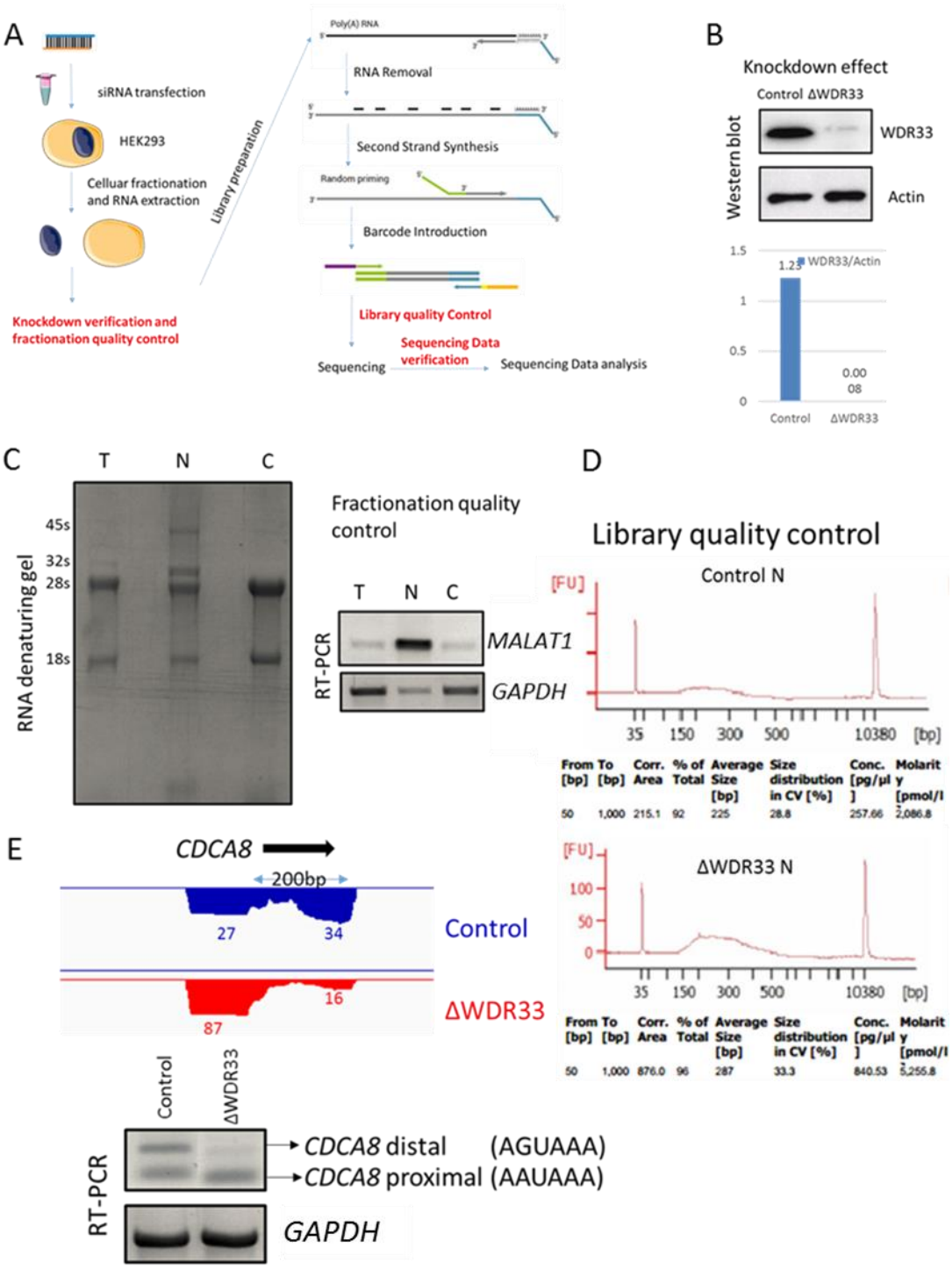

A
siRNA transfection
HEK293
Celluar fractionation and RNA extraction
Knockdown verification and fractionation quality control
Library preparation
Poly(A) RNA
RNA Removal
Second Strand Synthesis
Random priming
Barcode Introduction
Library quality Control
Sequencing Data verification
Sequencing
Sequencing Data analysis
B
Knockdown effect
Control ΔWDR33
Western blot
WDR33
Actin
WDR33/Actin
1.23
0.00 08
Control
ΔWDR33
C
T N C
RNA denaturing gel
45s
32s
28s
18s
Fractionation quality control
RT-PCR
MALAT1
GAPDH
D
Library quality control
Control N
[FU]
35 150 300 500 10380 [bp]
From [bp] To [bp] Corr. Area % of Total Average Size [bp] Size distribution in CV [%] Conc. [pg/µl] Molarity [pmol/l]
50 1,000 215.1 92 225 28.8 257.66 2,086.8
ΔWDR33 N
100
50
0
876.0 96 287 33.3 840.53 6,255.8
E
CDCA8
200bp
27
34
16
87
Control
ΔWDR33
RT-PCR
CDCA8 distal (AGUAAA)
CDCA8 proximal (AAUAAA)
GAPDH

Figure 29 WDR33 knockdown and sequencing strategy and three quality control steps

(A) Workflow of the RNA sequencing evaluating the global influence of different protein factors. Different protein factors were depleted by transient siRNA transfection. Nuclear fraction of RNA was isolated. First quality control of the samples for knockdown verification and fractionation quality was done before library preparation. Library was generated by Lexogen QuantSeq kit by RT-PCR principle. The second library quality control was assessed by Agilent Bioanalyzer. The third quality control to verify the sequencing data was assessed by RT-PCR. Three quality control steps are labelled in red letters in the workflow. (B) Verification of WDR33 knockdown by Western blot analysis. The same membrane was probed with anti- WDR33 and Actin (loading control) antibody. The control is HEK 293 cells transfected with siRNA negative control. The ratio of WDR33/Actin was the ratio of grey scale of the respective blots calculated by ImageJ. (C) Fractionation quality control using two approaches: RNA gel detecting precursor ribosomal RNA and RT-PCR detecting *MALAT1* in nuclear fraction. Only WDR33 knockdown samples were shown. (D) Library quality control done by Agilent Bioanalyzer. (E) Verification of the sequencing data by detecting *CDCA8* in RT-PCR. The sequencing data was visualised in IGV software. Two poly(A) sites were detected in a single PCR reaction. The type of the poly(A) site is indicated.

The effect of WDR33 on APA (Fig 33A) produced by the algorithm stated above does not specifically calculate the extent of different types of poly(A) sites affected by the protein factor. For this, two gene categories were created based on the raw sequencing data to investigate the influence of WDR33 on canonical and noncanonical poly(A) sites. The first “AGTAAA” gene category contains genes undergoing UTR-APA with tandem poly(A) sites, where one poly(A) site is directed by AGUAAA and another is directed by AAUAAA/AUUAAA. Globally, there are 158 genes in this category (Fig 30A). The second expanding gene category, “All noncanonical”, contains genes undergoing UTR-APA with tandem poly(A) sites, where one poly(A) site is directed by the canonical

hexamer, and another is directed by all the noncanonical poly(A) signals. There are 1204 genes in this category (Fig 30A). In both gene categories, a bias towards the proximal poly(A) site position for the noncanonical poly(A) site has been found (65% genes with proximal AGUAAA poly(A) sites in the “AGTAAA” gene category, 72% genes with proximal noncanonical poly(A) sites in the “All noncanonical” gene category, Fig 30B), as expected, according to the reported data (Wang, Dowell et al. 2013). The noncanonical poly(A) site usage ratio was calculated in individual genes first and then the distribution of the usage ratio was compared between the negative control and knockdown samples in each gene category (Fig 30A). The AGUAAA poly(A) site usage ratio was decreased in 77 out of 158 genes, less than 50%, in depletion of WDR33. Comparatively, the AGUAAA poly(A) site usage increased in 81 genes (Fig 33C). In addition, in the “All noncanonical” gene category, which contains 1204 genes, 609 genes showed decreased noncanonical poly(A) site usage and 588 genes were with increased noncanonical poly(A) site usage in WDR33 knockdown cells (Fig 33C). However, the distribution of global noncanonical poly(A) site usage in WDR33 knockdown is not significantly different from the negative control in both two gene categories. This indicates that WDR33 does not preferentially increase the usage of either the AGUAAA/noncanonical poly(A) site or the canonical poly(A) site in tandem UTR-APA genes. In the RNase protection assay, the depletion of WDR33 resulted in a decreased usage of the JunD AGUAAA poly(A) site which lacks an adU-rich element. We used MEME to further explore whether globally, the downregulated noncanonical poly(A) site caused by WDR33 knockdown specifically contains a similar U-rich region in JunD. For this, genes in the “AGTAAA” gene category were further divided into

two sub-groups, a group of genes with upregulated AGUAAA poly(A) site usage as the control and another group of genes with downregulated AGUAAA poly(A) site usage after the depletion of WDR33. The sequence including AGUAAA hexamer and 20 nucleotides downstream in each gene was extracted and analysed by MEME for each group. No significant motifs other than the hexamer were identified in either group. This means that genes with downregulated AGUAAA poly(A) site usage do not necessarily contain a U-rich enhancer, as identified before. Also, we performed the same analysis based on the usage of the canonical poly(A) site, and again, no significant motifs apart from the hexamer were identified.

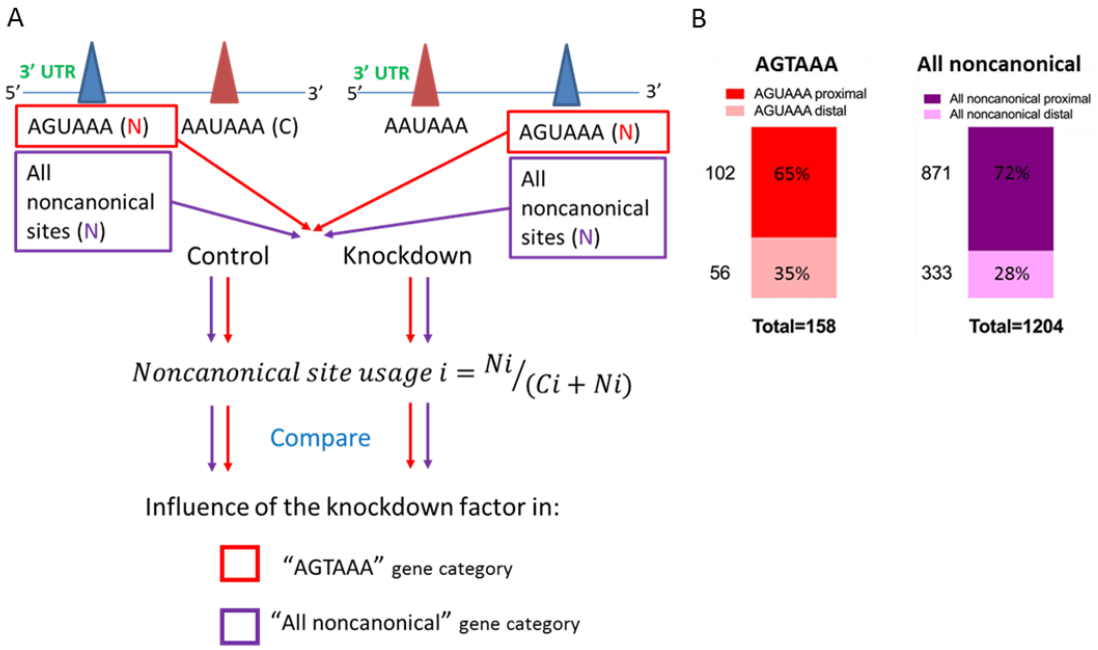


Figure 30 Schematic workflow to analyze the influence of the protein factors on canonical and noncanonical poly(A) site selection

(A) Two gene categories "AGTAAA" (in red box) and "All noncanonical" (in purple box) were generated and analyzed (in red arrows or purple arrows) separately. Noncanonical poly(A) site usage was calculated in the formula indicated. The distribution of the usage ratio in each gene category was compared between the negative control and protein knockdown samples. Ni and Ci refer to the reads number

per million reads of the respective poly(A) signal. (B) Numbers and percentage of the genes with AGUAAA/ noncanonical poly(A) site at the proximal position in two gene categories.

4.3 The global role of hexamer binding protein CPSF30 on canonical and noncanonical poly(A) site usage

Two recent studies have shown that CPSF30, the smallest protein subunit of CPSF complex, directly binds to the canonical hexamer AAUAAA and this subunit is essential for 3' end formation (Chan, Huppertz et al. 2014, Schönemann, Kühn et al. 2014). In the individual gene study, we have shown that the depletion of CPSF30 increases the usage of noncanonical poly(A) site when the condition of the *cis* elements is sub-optimal (Fig 23C). To address the global role of CPSF30 on the different type of poly(A) sites, we used the identical RNA sequencing strategy analysing the influence of WDR33 (Fig 29A). After three quality control steps (Fig 31), we started to analyze the sequencing data in the workflow stated in Figure 30A. Firstly and similar to WDR33, the depletion of CPSF30 has modest effect on APA (4% significant UTR-APA change events compared to the negative control and 2% significant CR-APA change events compared to the negative control, Fig 33A). Also, CPSF30 is not suggested to preferentially increase the usage of either the AGUAAA/noncanonical poly(A) site or the canonical poly(A) site, in tandem UTR-APA genes. This is supported by the fact that, the change of the distribution of noncanonical poly(A) site usage after the depletion of CPSF30, is not significantly different compared to the negative control. In "AGTAAA" gene category, the AGUAAA poly(A) site usage was decreased in 73 out of 158

genes in depletion of CPSF30. Comparatively, the AGUAAA poly(A) site usage increased in 83 genes (Fig 33C). In the “All noncanonical” gene category, 570 genes were with decreased noncanonical poly(A) site usage and 620 genes showed increased noncanonical poly(A) site usage in CPSF30 depleted cells (Fig 33C). We then performed the identical MEME motif analysis as described based on the two sub-groups, categorised according to the alteration of the AGUAAA poly(A) site usage after CPSF30 knockdown. No significant motifs apart from the hexamer signal were identified within in either sub-groups. This reveals that, unlike the upregulated sub-optimal poly(A) site in JunD Mut4, the AGUAAA poly(A) sites, upregulated after the depletion of CPSF30, do not necessarily contain a U-rich region. Also, we performed the same analysis based on the usage of the canonical poly(A) site, and again, no significant motifs apart from the hexamer were identified.

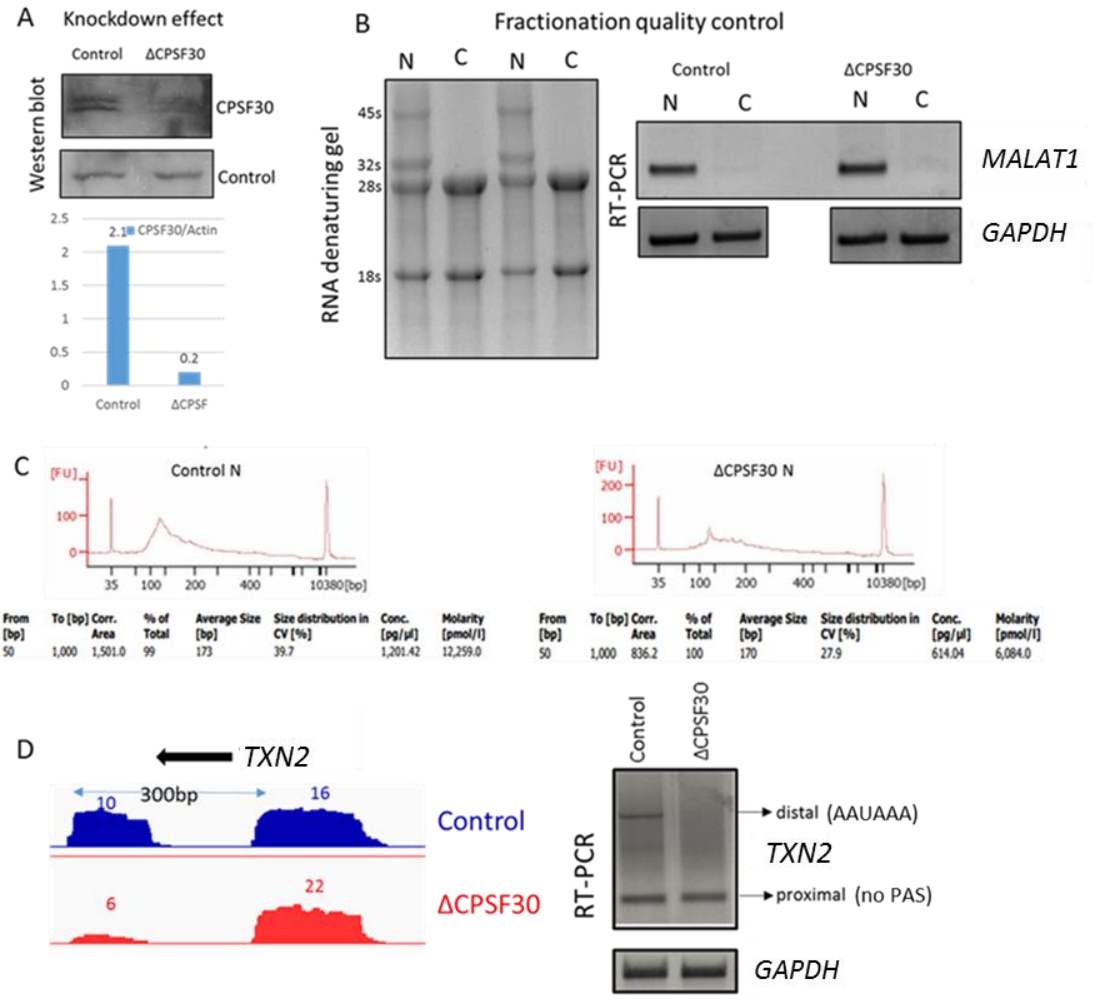


Figure 31 Three quality control steps in CPSF30 knockdown and sequencing

(A) Verification of CPSF30 knockdown. The same membrane was probed with anti-CPSF30 and Actin (loading control) antibody. The control was HEK 293 cells transfected with siRNA negative control. The ratio of CPSF30/Actin was the ratio of the grey scale of the respective blots calculated by ImageJ. (B) Fractionation quality control using two approaches as described. (C) Library quality control assessed by Agilent Bioanalyzer. (D) Verification of the sequencing data by detecting *TXN2* in RT-PCR. The sequencing data was visualised in IGV software. Two poly(A) sites were detected in a single PCR reaction. The type of the poly(A) site is indicated.

4.4 The global role of SAM68 on canonical and noncanonical poly(A) site usage

SAM68 has been reported to have affinity to the RNA sequence UAAA (Itoh, Haga et al. 2002) and has been shown to participate in alternative splicing and viral 3' end formation (Lukong and Richard 2003). Similar to CPSF30, depletion of SAM68 increases the usage of sub-optimal JunD poly(A) site. And this result suggests its role in 3' end formation. Thus, we employed the identical sequencing approach (Fig 29A) with three quality control steps (Fig 32) to investigate its role on canonical and noncanonical poly(A) site selection. The sequencing data analysis workflow is stated in Figure 30A. Firstly, from a general APA aspect, according to the distribution of significant APA changes produced by the algorithm, SAM68 has little effect on APA (2% UTR-APA changes and 2% CR-APA affected, Fig 33A). Interestingly, a clear bias towards 3' UTR shortening events amongst significant poly(A) site switching cases has been observed.

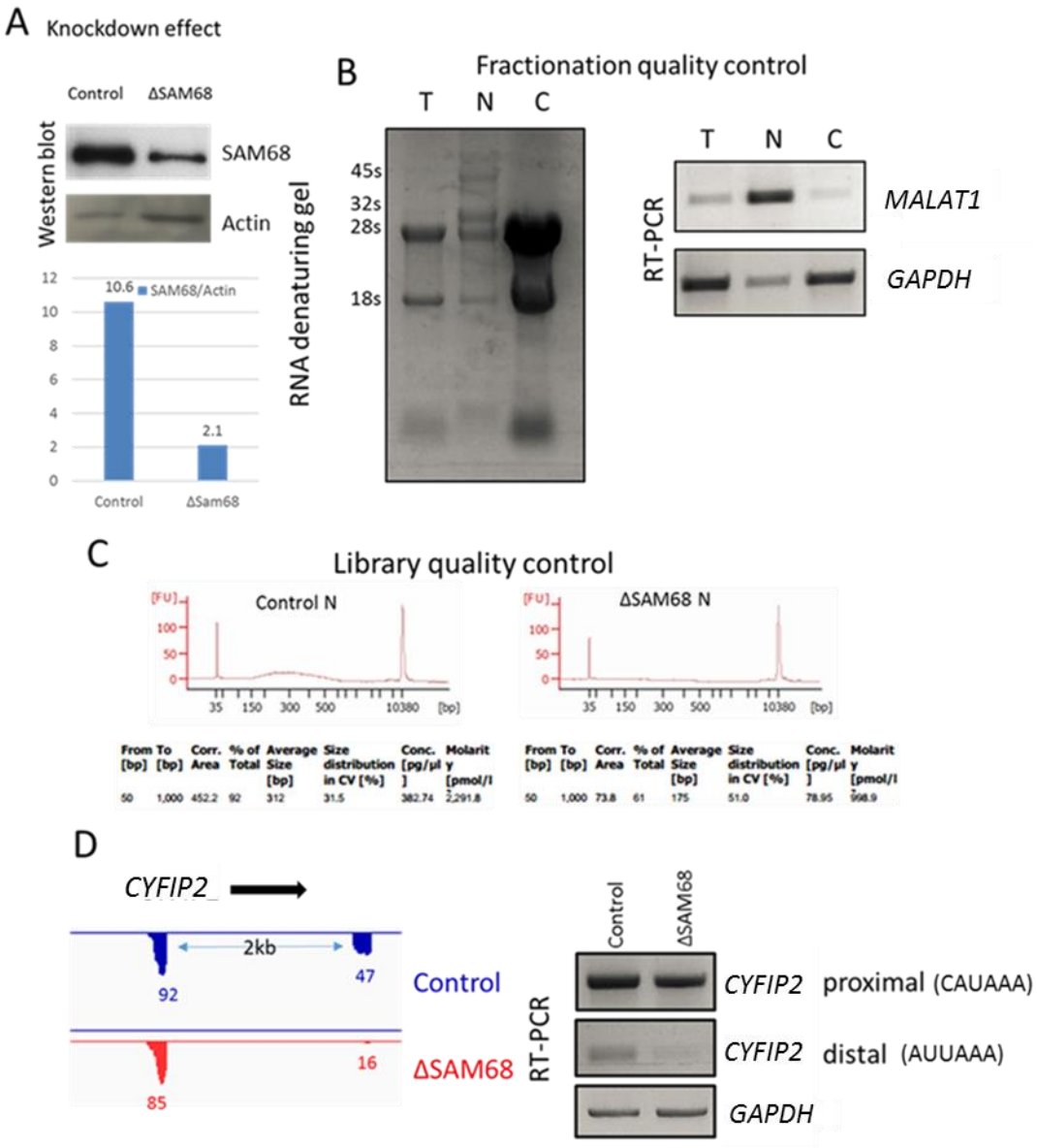


Figure 32 Three quality control steps in SAM68 knockdown and sequencing

(A) Verification of SAM68 knockdown. The same membrane was probed with anti-SAM68 and Actin (loading control) antibody. Control was HEK 293 cells transfected with siRNA negative control. The ratio of SAM68/Actin was the ratio of the grey scale of the respective blots calculated by ImageJ. (B) Fractionation quality control using two approaches as described. Only SAM68 knockdown samples were shown. (C) Library quality control generated by Agilent Bioanalyzer. (D) Verification of the sequencing data by detecting *CYFIP2* in RT-PCR. The sequencing data was visualised in IGV software. Two poly(A) sites were detected from two separate PCR reactions because of the 2kb distance between two poly(A) sites. The type of the poly(A) site is indicated.

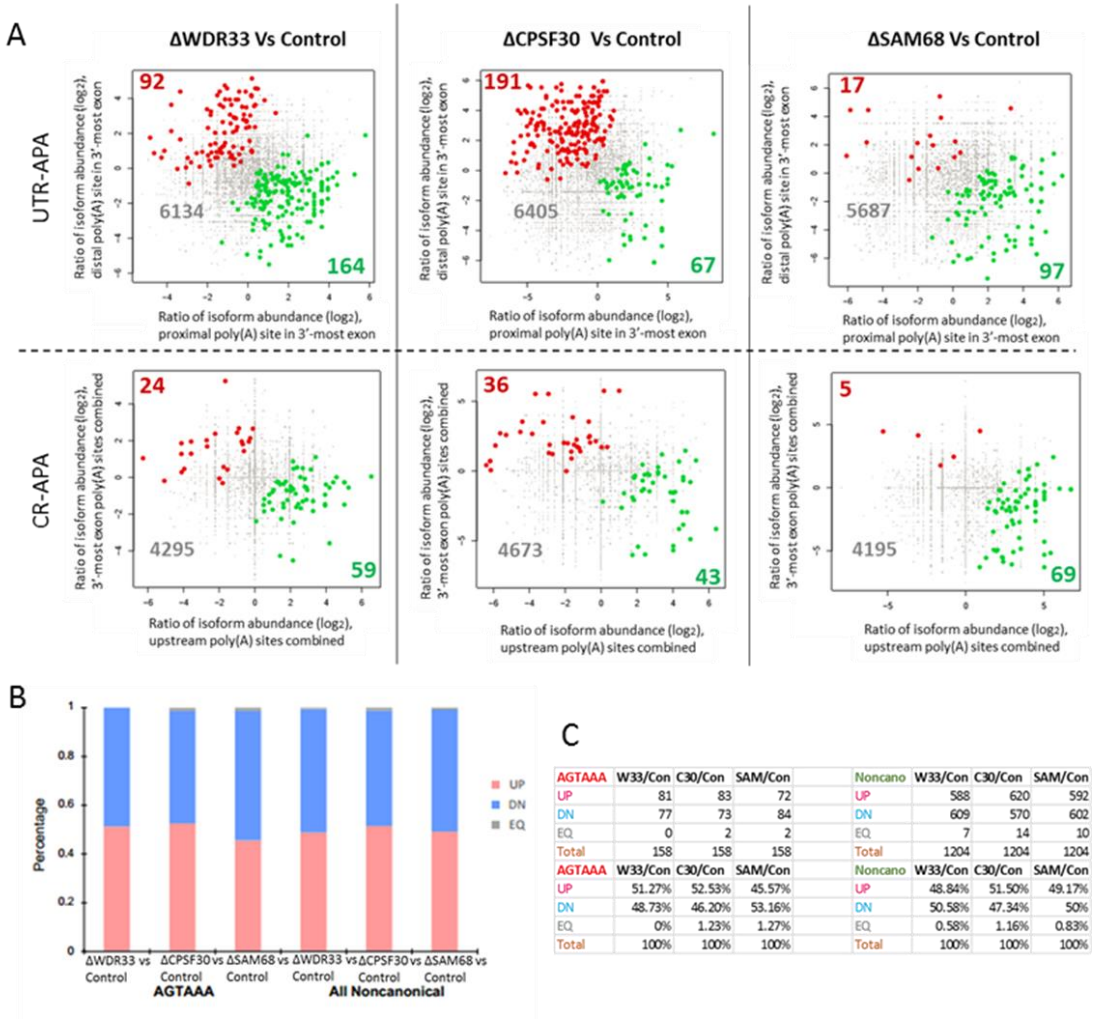
Also, SAM68 is not suggested to preferentially increase the usage of either AGUAAA/noncanonical poly(A) site or canonical poly(A) site in tandem UTR-APA genes, since the change of the distribution of noncanonical poly(A) site usage after the depletion of SAM68 is not significantly different, compared to the negative control, in either of the two gene categories. In the “AGTAAA” gene category, the AGUAAA poly(A) site usage was decreased in 84 out of 158 genes, more than 50%, in depletion of SAM68. Meanwhile, the AGUAAA poly(A) site usage increased in 72 genes (Fig 33C). In the “All noncanonical” gene category, 602 genes showed decreased noncanonical poly(A) site usage and 592 genes showed increased noncanonical poly(A) site usage in SAM68 knockdown cells (Fig 33C).

4.5 Summary – Influence of WDR33, CPSF30 and SAM68

The distribution of significant global APA changes produced by the algorithm developed by the Tian Lab and the distribution of up- or downregulation of noncanonical poly(A) site usage in two gene categories caused by WDR33, CPSF30 and SAM68 are summarised in Figure 33. Globally, all the three protein factors have a modest effect on APA. However, the significant APA changes caused by SAM68 depletion bias towards 3’ UTR shortening have been found. Additionally, the distribution of the noncanonical poly(A) site usage in knockdown samples is statistically not significantly different from the distribution in the negative control. This suggests that WDR33, CPSF30 or

SAM68 do not have preference for either the canonical or the noncanonical poly(A) sites in tandem poly(A) sites genes.

In addition, in the motif analysis, which aims to identify *cis* elements specifically associated with up- or downregulated AGUAAA poly(A) site usage caused by WDR33 and CPSF30 knockdown, no significant motif, apart from the hexamer signal, was found. This suggests that the adU-rich element we identified in JunD is not likely to globally promote the recognition of AGUAAA poly(A) site during tandem UTR-APA.



AGTAAA	W33/Con	C30/Con	SAM/Con
UP	81	83	72
DN	77	73	84
EQ	0	2	2
Total	158	158	158

Noncano	W33/Con	C30/Con	SAM/Con
UP	588	620	592
DN	609	570	602
EQ	7	14	10
Total	1204	1204	1204

AGTAAA	W33/Con	C30/Con	SAM/Con
UP	51.27%	52.53%	45.57%
DN	48.73%	46.20%	53.16%
EQ	0%	1.23%	1.27%
Total	100%	100%	100%

Noncano	W33/Con	C30/Con	SAM/Con
UP	48.84%	51.50%	49.17%
DN	50.58%	47.34%	50%
EQ	0.58%	1.16%	0.83%
Total	100%	100%	100%

Figure 33 Globally WDR33, CPSF30 and SAM68 do not significantly influence the canonical and noncanonical poly(A) site selection

(A) Upper panel: Genes with UTR-APA isoforms that differ significantly in relative representation of their two most highly expressed isoforms in cells with depleted protein factors compared to the control. Genes are highlighted when the shorter isoforms (green) or longer isoforms (red) have a significantly higher relative representation in the knockdown samples compared to the negative control. Generated by the algorithm developed by Tian Lab. Bottom panel: Genes with CR-APA, which significantly differ in their representation of upstream region (intronic and exonic) APA isoforms relative to 3' UTR isoforms between the protein factor knockdown samples and the control. Genes are highlighted when the upstream APA isoforms (green) or 3' UTR isoforms (red) have a significantly higher relative representation in the knockdown samples compared to the negative control. Significantly regulated genes are those in which $P \leqslant 0.01$ (Fisher's exact test) and the abundance change is >5%. (B) Distribution of up- and downregulation of noncanonical

poly(A) site usage caused by the depletion of protein factors in two gene categories. “AGTAAA” are genes with tandem poly(A) sites in 3’ UTR, one site directed by the AGUAAA signal and another directed by the AAUAAA/AUUAAA signal. There are 158 genes in this category globally. “All noncanonical” are genes with tandem poly(A) sites in 3’ UTR, one site directed by the AAUAAA/AUUAAA and another directed by any other poly(A) signals. There are 1204 genes in this category globally. A noncanonical poly(A) site is defined as upregulated if the protein knockdown causes its usage ratio higher than the ratio in the negative control, and vice versa. (C) This table shows the numbers and percentage of genes up- or downregulated caused by depletion of different protein factors. W33: WDR33, C30: CPSF30, SAM: SAM68, Con: Control.

As can be seen in Table 3, the average AGUAAA poly(A) site usage in ratio is lower than the canonical poly(A) site usage in “AGTAAA” gene category in the control group. After the knockdown of any of the three proteins, the AGUAAA poly(A) site usage is almost the same as the control. This again suggests that WDR33, CPSF30 or SAM68 do not have preference for either the canonical or AGUAAA poly(A) sites in tandem poly(A) sites genes.

Table 3 Average of the AGUAAA/ canonical poly(A) site usage ratio in “AGTAAA” gene category

	Control ratio	**ΔWDR33 ratio**	**ΔCPSF30 ratio**	**ΔSAM68 ratio**
AGUAAA poly(A) site usage	43.8%	43.1%	43.8%	42.4%
Canonical poly(A) site usage	56.2%	56.9%	56.2%	57.6%

Finally, a Venn diagram was constructed to identify the genes which showed altered noncanonical poly(A) site usage with depletion of any of the three protein factors including WDR33, CPSF30 and SAM68 (Fig 34). Around half of the genes that showed altered noncanonical poly(A) site usage with depletion of either WDR33, CPSF30 or SAM68 overlapped with each other. This was

seen for both increased and decreased noncanonical poly(A) site usage in both of the gene categories. This suggests that many APA events might be influenced by the different CPSF sub-complexes, caused by the low levels of WDR33 or CPSF30, in a similar way. For SAM68, the overlapping results might be another indication of its participation during mammalian 3' end processing. However, no link between this protein and CPSF has been found to date.

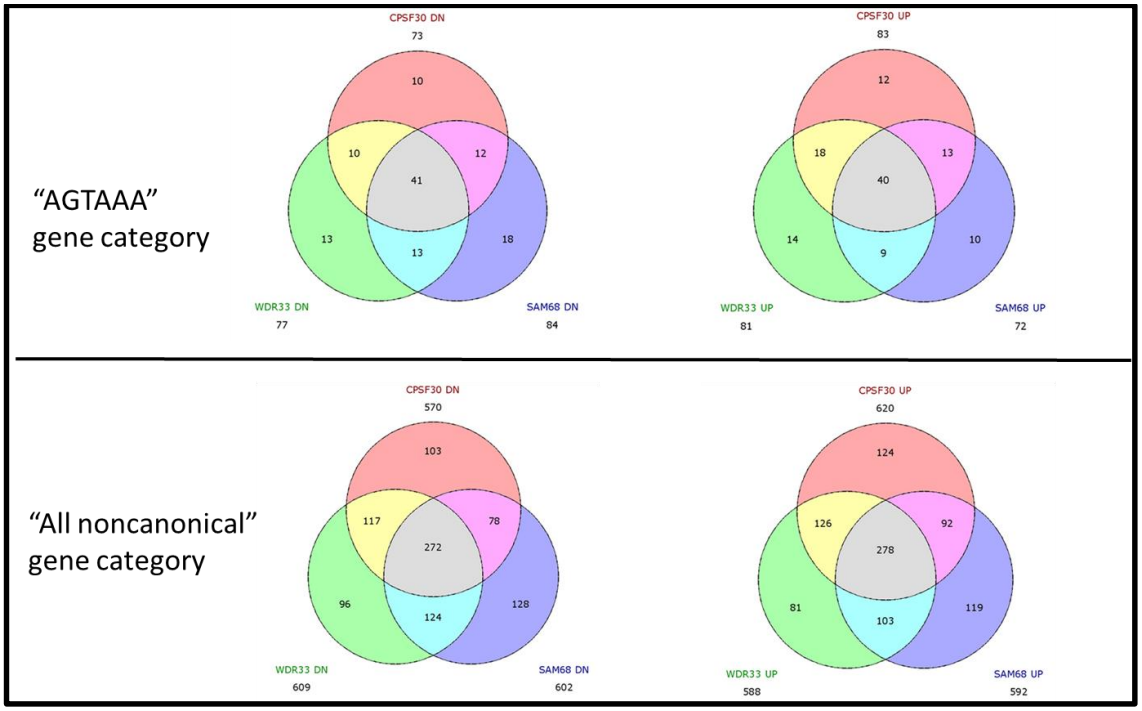


Figure 34 Venn diagram depicting the overlapping influence caused by WDR33, CPSF30 and SAM68

Numbers of the genes with overlapping up- or downregulated noncanonical poly(A) site usage in "AGTAAA" and "All noncanonical" gene categories caused by WDR33, CPSD30 and SAM68.

4.6 Discussion

Two hexamer signal binding protein factors WDR33 and CPSF30 (Chan, Huppertz et al. 2014, Schönemann, Kühn et al. 2014), together with a potential

hexamer interacting candidate SAM68, were depleted and their role on canonical and noncanonical poly(A) site selection was investigated in two gene categories, which enabled us to compare the usage between a canonical and a noncanonical poly(A) sites, during UTR-APA. In contrast to the influence of WDR33 and CPSF30 on JunD-MC4R 3' end formation (Chapter 3), the change of the distribution of noncanonical poly(A) site usage is not significantly different after the protein factor knockdown. This suggests that these two protein factors do not have obvious influence on canonical and noncanonical poly(A) site selection. Also, the modest effect of both of these two protein factors on APA has been indicated, and this is consistent with a global analysis based on mouse cell lines, which reveals that, the low levels of WDR33 and CPSF30 cause relatively few genes to switch the usage of poly(A) sites (Li, You et al. 2015). Together with our data, in a global view, WDR33 and CPSF30 are likely to equally recognise different types of poly(A) sites and they might compensate for each other during poly(A) site recognition.

SAM68 has been shown to interact with the UAAA sequence (Taylor and Shalloway 1994, Lin, Taylor et al. 1997, Itoh, Haga et al. 2002). We have shown that depletion of SAM68 increases the usage of sub-optimal AGUAAA poly(A) site (Fig 28C). It is not known whether this protein participates in mammalian 3' end processing and further influences canonical and noncanonical poly(A) sites. After the knockdown and sequencing experiment, we suggest that the global influence of SAM68 on canonical and noncanonical poly(A) site selection in either gene category is not obvious. Upon SAM68 knockdown, however, individual genes are undergoing APA events (Fig A2) and show a general trend

towards 3’ UTR shortening. This indicates that SAM68 might influence 3’ end formation in a small sub-group of genes.

In addition, we investigated whether there were motifs existing downstream of the AGUAAA signal, specifically in up- or downregulated sub-groups caused by WDR33 or CPSF30 knockdown. In the MEME analysis, no significant motif other than the hexamer signal was found in any sub-groups. This suggests that the adU-rich element is not likely to globally promote the recognition of AGUAAA poly(A) sites during tandem UTR-APA. Considering the global presence of the U-rich region in single AGUAAA poly(A) site genes (Fig 20), the adU-rich element might only facilitate 3’ end formation in that gene group.

Chapter 5

The Investigation of the sequence requirement of the ARAF poly(A) site as a representative of cleavage and polyadenylation at a “no PAS” site

Chapter 5

5.1 Introduction

In the previous chapters, we used JunD as an individual example to investigate the noncanonical AGUAAA poly(A) site. Also, recognition of the so called noncanonical “no PAS” poly(A) sites, which lack any of the canonical or noncanonical hexamers, are poorly understood. Our group have found that additional auxiliary elements can promote the recognition of these poly(A) sites (Dalziel, Nunes et al. 2006, Nunes, Li et al. 2010).

In this chapter, we investigate how these highly unusual noncanonical poly(A) sites, which lack the defined hexamer signals, can be recognised. The analysis of the individual gene example *ARAF* is presented which is a representative of genes with “no PAS” poly(A) sites.

5.2 *ARAF* is a candidate to investigate 3’ end processing at a “no PAS” site

ARAF is a member of the RAF protein family (Baljuls, Schmitz et al. 2008). *ARAF* codes for cytoplasmic serine/threonine kinases that are regulated by binding RAS (McCubrey, Steelman et al. 2007). In addition, *ARAF* is infrequently mutated in human cancer (Baljuls, Schmitz et al. 2008).

Our RNA 3’ READS sequencing maps the *ARAF* poly(A) site to a position that is in agreement with other RNA-seq data published earlier by Merck (Derti, Garrett-Engele et al. 2012), which reveals that ARAF cleavage and

polyadenylation is directed downstream of a sequence element that has no resemblance with any prior described poly(A) hexamers (Fig 35). The sequence element “AAGAAGCACAGAA” which is found 10nt upstream of the major cleavage site is a potential candidate to serve as the upstream poly(A) signal for ARAF 3’ end processing. This putative poly(A) signal does not belong to any types of hexamer signal so that the ARAF poly(A) site falls into the so called “no PAS” site category representing around 9% of all poly(A) sites (Fig 5). The fact that 3’ end formation of the ARAF pre-mRNA does not require a canonical or noncanonical hexamer signal makes it an intriguing example to study how such un usual site is recognised.

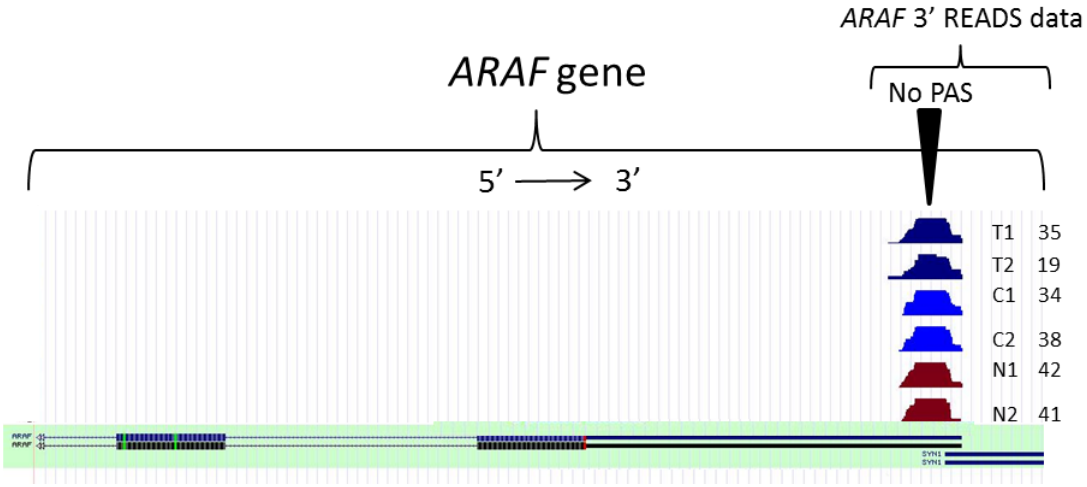


Figure 35 RNA 3’ READS sequencing data of ARAF in HEK 293 cells

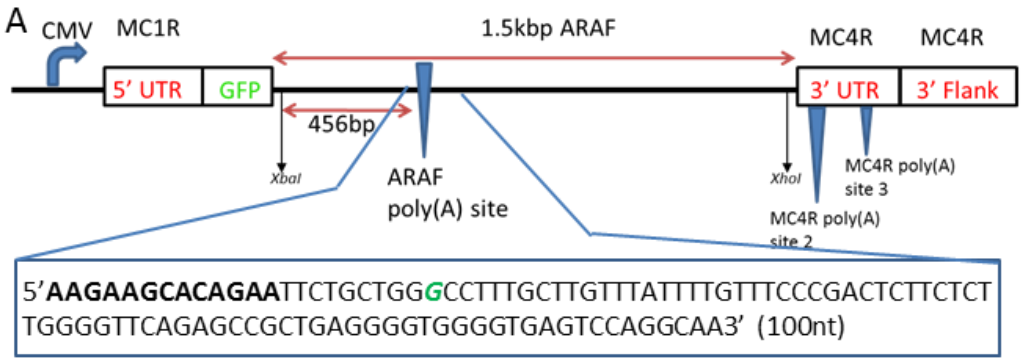
RNA 3’ READS sequencing data (Neve, Burger et al. 2016) of ARAF gene in HEK 293 cell line of whole cell (T1, T2), cytoplasmic (C1, C2) and nuclear (N1, N2) fractions. The exact cleavage site varies from 3 nucleotides to 21 nucleotides downstream of AAGAAGCACAGAA, the putative sequence element serves as a poly(A) signal like role in ARAF pre-mRNA. The dominant cleavage site is 10-nucleotide after the “AAGAAGCACAGAA” sequence. The data is visualised in UCSC. Reads per million in all the six samples are shown.

5.3 Characterisation of the ARAF poly(A) signal

The initial series of experiments analyzing *ARAF* followed the same strategy to investigate the *JUND* poly(A) site. We aimed to establish a Mini-gene based system, that used the ARAF poly(A) site for 3' end formation, to define the minimal *cis* elements that are required for this "no PAS" poly(A) site.

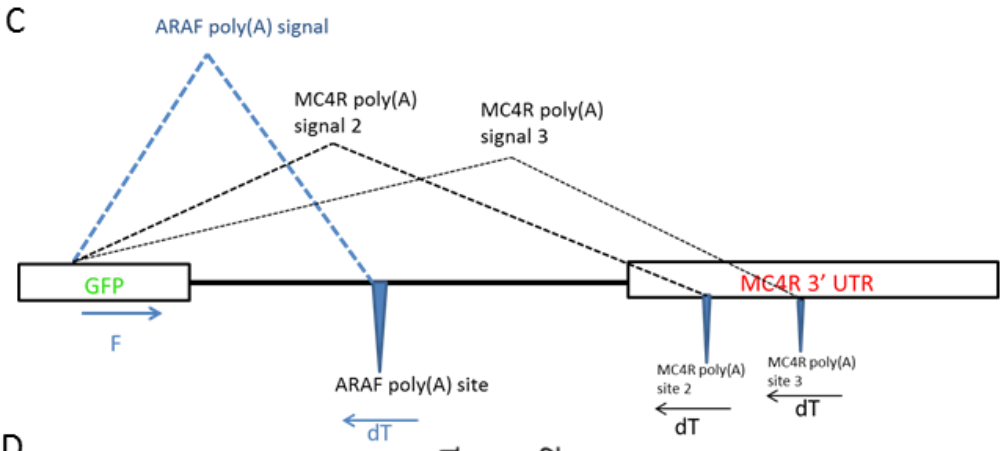
To do this, we first cloned 1.5kb wildtype ARAF (Fig 36B A1WT) into the above described MC4R plasmid (Fig 9) upstream of two functional MC4R poly(A) sites by *XbaI* and *XhoI* digestion (Fig 36A). The cryptic poly(A) site was removed after the restriction enzyme digestion of the plasmid described in Figure 9. This way, a poly(A) competition vector is created and the two functional poly(A) sites of the MC4R that are present approximately 100bp and 200bp downstream of the inserted ARAF fragment can be used to measure the ratio of transcripts that read through the upstream positioned ARAF poly(A) site. From this "parent" plasmid, additional constructs (Fig 36B A1Mut1 and A1Mut2) were created which contained mutations that aimed to disrupt the putative poly(A) signal, "AAGAAGCACAGAA". The plasmids containing wildtype and mutations in "AAGAAGCACAGAA" were subsequently transfected into HEK 293 cells. After 24 hours, whole cell RNA was isolated by the hot phenol method (see Materials and Methods). The RT-PCR method was the same as described in Figure 13A, but no potential "Cryptic poly(A) site" can be detected (Fig 36C). There is no A-rich sequence downstream of the ARAF cleavage site (Fig 36A blue box). This means that the internal priming issue is not present in the ARAF mRNA. As can be seen in Figure 36D ARAF poly(A) site, lane 1 A1WT, a strong band corresponding to the size of the ARAF poly(A) signal, was detected. This band was gel purified and sequenced. It was then confirmed to be the

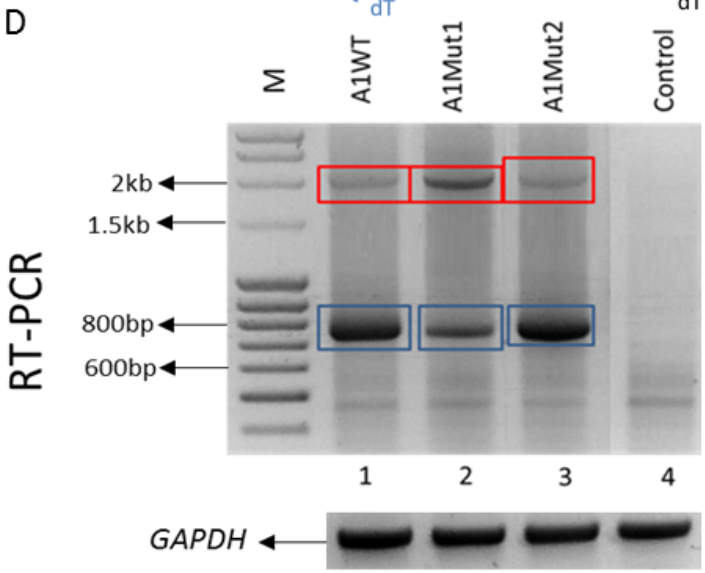
polyadenylation site of ARAF despite the lack of any previously known hexamer signal. The exact cleavage site has been characterised to be 10 nucleotides downstream of the putative “AAGAAGCACAGAA” after “GG” sequence. However, the second alternative MC4R poly(A) site (Fig 36D lane 1 MC4R poly(A) site 3) was detected and confirmed by sequencing which indicates that polyadenylation in the 1.5kb of wildtype ARAF fragment is leaky to some degree. Since A1Mut1 caused obviously decreased usage of ARAF poly(A) site (Fig 36D lane 2 ARAF poly(A) site and MC4R poly(A) site 3, 36E), it is suggested that the two mutated adenosines are important for its 3’ end processing. A1Mut2 did not disrupt ARAF poly(A) site usage significantly (Fig 36D lane 3 MC4R poly(A) site 3, 36E). This indicates that the mutated “A” is not likely to be part of the poly(A) signal. Thus, “AAGAAG” may be part of a wider recognition sequence that is required to direct cleavage and polyadenylation in the *ARAF* gene and no clear hexamer can be identified.



B

Name	Insert Sequence (5' – 3')
A1WT	440bp upstream + AAGAAGCACAGAA + 1037bp downstream
A1Mut1	440bp upstream + ACGACGCACAGAA + 1037bp downstream
A1Mut2	440bp upstream + AAGAAGCACAGAC + 1037bp downstream





E

ARAF poly(A) site usage in ratio

	%
A1WT	88
A1Mut1	43
A1Mut2	87

Figure 36 Mutagenesis analysis on the potential poly(A) signal in *ARAF*

(A) Schematic depicting ARAF fragments inserted into the *XbaI* and *XhoI* site of MC4R reporter constructs to assess ARAF poly(A) site usage. The relative position of the ARAF poly(A) site, 10 nucleotides after "AAGAAGCACAGAA", and two downstream

MC4R poly(A) sites (MC4R poly(A) site 2 and 3), is indicated. The 100-nucleotide sequence from the potential PAS (in bold) is shown. Green italic “G” is the major cleavage site. (B) Putative poly(A) signal sequences of ARAF wildtype (A1WT) and two mutants (A1Mut1 and A1Mut2) cloned into the MC4R plasmids. The mutated nucleotides are indicated in red italic letters. (C) Diagram depicting the constructs, primers and putative products in the RT-PCR analysis. The forward primer (F) in the GFP region and reverse primer (dT) used in RT-PCR analysis are indicated. Expected RT-PCR products represent the ARAF poly(A) signal, MC4R poly(A) signal 2 and MC4R poly(A) signal 3 are indicated. (D) RT-PCR analysis of ARAF and MC4R polyadenylation. The ARAF poly(A) site is indicated in blue boxes. The MC4R poly(A) site 3 is indicated in red frames. No MC4R poly(A) site 2 can be detected. *GAPDH* serves as an RNA loading control. The control lane is the RT-PCR result on untransfected HEK 293 cells. (E) ARAF poly(A) site usage in all three constructs. The absolute usage value of the poly(A) sites in RT-PCR were quantified by respective grey scale calculated by ImageJ software. The ARAF poly(A) site usage ratio was calculated by (ARAF poly(A) site)/(ARAF poly(A) site + MC4R poly(A) site 2 + MC4R poly(A) site 3). (n=3).

From the aspect of RNA secondary structure, ARAF is further suggested to be a highly unusual poly(A) site. According to the principle of RNA base-pairing, the six nucleotides within AGUAAA, AAUAAA or AUUAAA hexamer cannot interact with each other, perhaps to facilitate the accessibility to the CPSF complex. Early studies of the role of RNA secondary structure in the recognition of the AAUAAA hexamer reveal that this element is recognised in a single-stranded form (Sittler, Gallinaro et al. 1995). In contrast, a helical structure can be formed within the putative poly(A) signal of ARAF, when the secondary structure of the sequence surrounding the ARAF poly(A) site was predicted on Oxfold (Fig 37). This is surprising as none of the sequences constituting the canonical and AGUAAA hexamers (Fig 10B) can fold into structures.

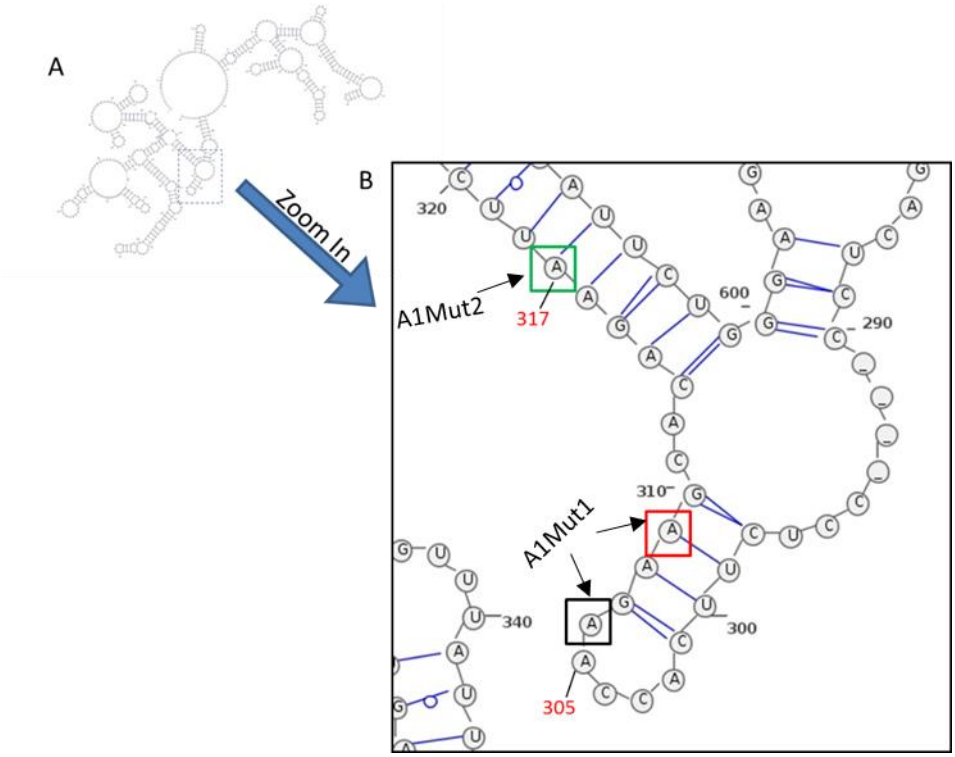


Figure 37 The RNA secondary structure prediction of *ARAF* pre-mRNA

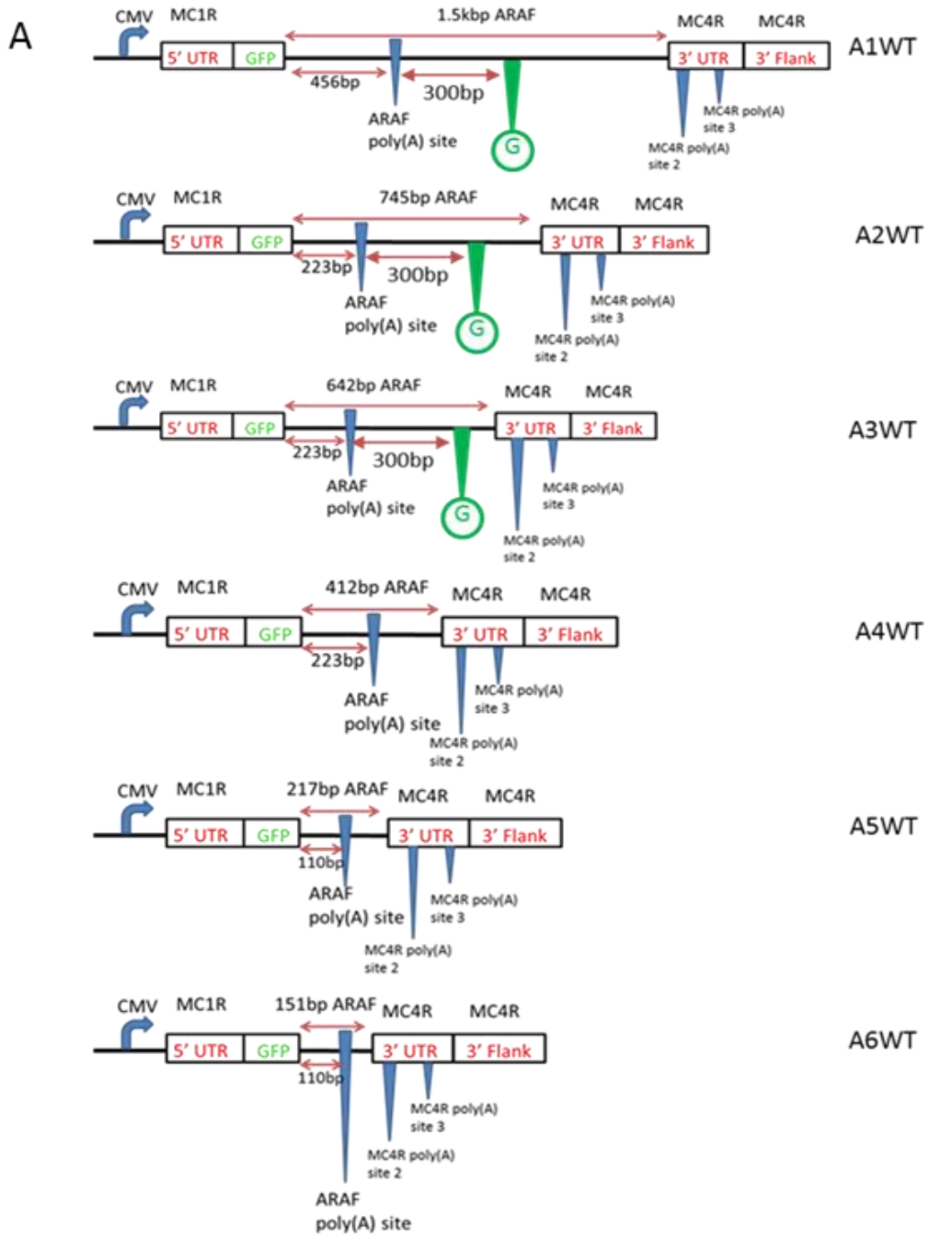
A) Visualisation of the *ARAF* RNA secondary structure calculated by Oxfold, drawn by the VARNA package (Darty, Denise et al. 2009). Five 600-nucleotides RNA sequences of *ARAF* from human, rhesus, mouse, dog and pig were used. The "AAGAAGCACAGAA" region was located in the middle of all the sequences. B) Partial RNA secondary structure zoomed in from the dashed area in (A). The "AAGAAGCACAGAA" sequence located between 305th nucleotide and 317th nucleotide as outlined in the structure. Three adenosines mutated in the mutagenesis study are squared by black, red and green boxes. The type of the mutation is indicated as A1Mut1 and A1Mut2.

5.4 The investigation of auxiliary elements in *ARAF*

After the examination of the ARAF 3' flanking region, a G-rich region (52 nucleotides long) was identified approximately 300-nucleotide downstream of the ARAF poly(A) site (Fig 38B). This G-rich region is very similar in position

and nucleotide composition to the G-rich sequence forming a G-Quadruplex structure (Hazel†, Huppert‡ et al. 2004, Risitano and Fox 2004, Huppert and Balasubramanian 2005, Petraccone, Erra et al. 2005, Todd, Johnston et al. 2005, Dalziel, Nunes et al. 2006), a characteristic RNA secondary structure that acts as a distant poly(A) enhancer (Decorsiere, Cayrel et al. 2011). The sequence of this G-rich region was extracted and applied in Quadruplex forming G-Rich Sequences Mapper (QGRS) (Kikin, D'Antonio et al. 2006), a software that searches for the sequences possible to form G-Quadruplexes and calculates the possibility displayed as a G-score (higher G-score means higher possibility to form a G-Quadruplex). The G-score is calculated as 59 (Fig 38C), which is higher than the score of the valid G-Quadruplex, which scored 49 with the G-rich sequence “GGGTGGGTGGCAGAGCTGGGCTGGG” (Kikin, D'Antonio et al. 2006). Thus, the analysis from QGRS suggests that the given G-rich sequence in ARAF is highly probable to form a G-Quadruplex.

In order to test whether the putative G-Quadruplex carries any important function for ARAF 3’ end formation, five additional ARAF-MC4R constructs with different lengths of ARAF 3’ UTR and 3’ flanking region were cloned into the same location as the ARAF insert in the A1WT (Fig 38A). A1WT (1.5kbp ARAF insert), A2WT (745bp ARAF insert) and A3WT (642bp ARAF insert) encompass the putative G-Quadruplex, which is located 300-nucleotides downstream of the ARAF poly(A) site (Fig 38B). The constructs A4WT (412bp ARAF insert), A5WT (217bp ARAF insert) and A6WT (151bp ARAF insert) all represent examples that lack this region.



B CTGGGGAGGGGGTCCCTGGGGGAACCGAGGGGTCCTGGAAGTGGGAGGGGCT

C QGRS sequences found (overlaps not included)

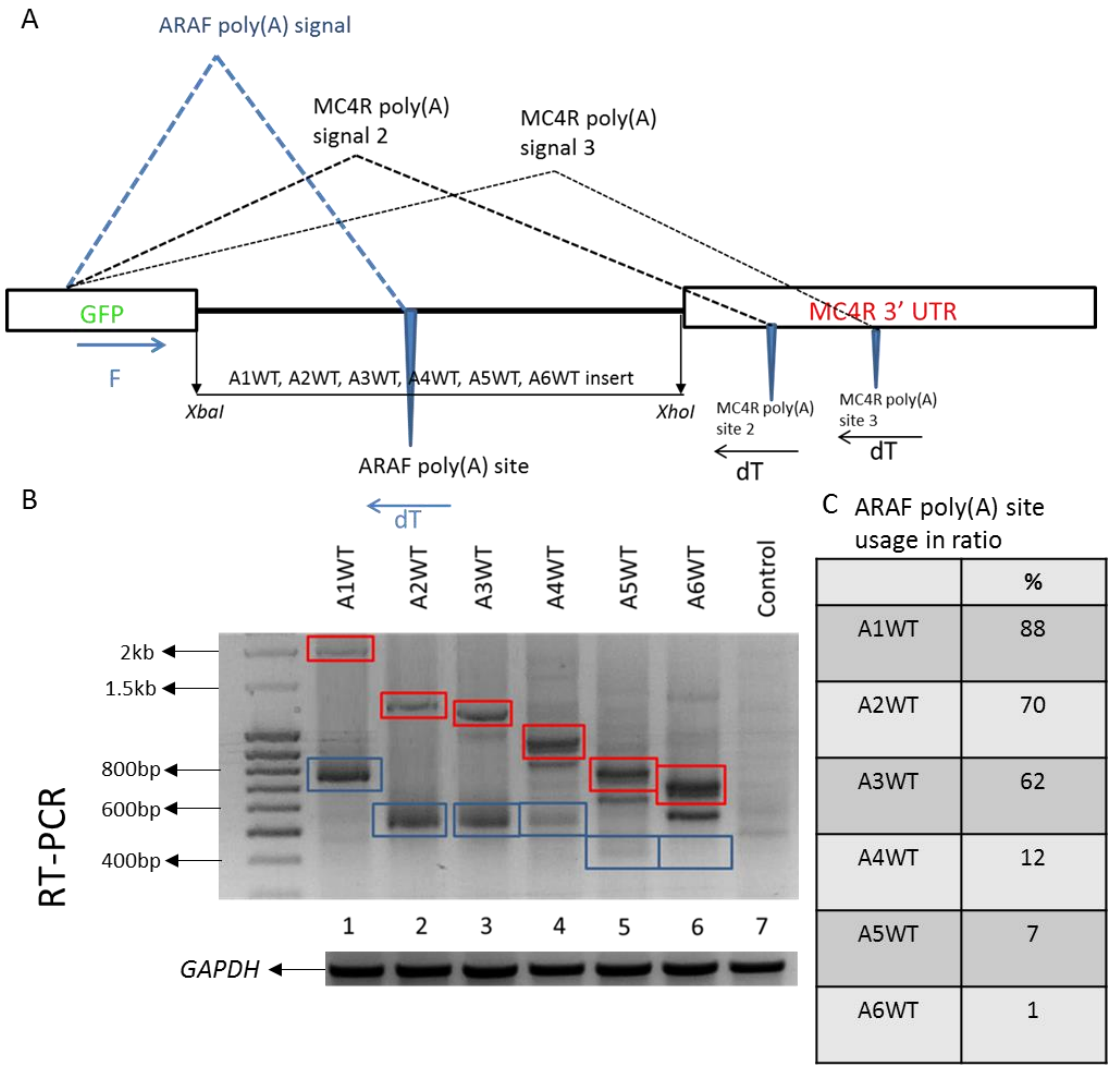
Position	Length	QGRS	G-Score
1	30	GGGGAGGGGGTCCCTGGGGGAACCGAGGGG	59
35	14	GGAAGTGGGAGGGG	17

Figure 38 Maps of ARAF-MC4R constructs with different lengths of ARAF fragments (A) Six different lengths of *ARAF* insert in MC4R constructs used to verify the importance of the G-rich region 300nt downstream of the putative *ARAF* poly(A) signal.

1.5k nucleotides (A1WT), 745 nucleotides (A2WT), 642 nucleotides (A3WT), 412 nucleotides (A4WT), 217 nucleotides (A5WT) and 151 nucleotides (A6WT) of ARAF fragments were cloned into the *XbaI* and *XhoI* site in MC4R plasmids. ARAF poly(A) site represents the cleavage site 10 nucleotides after "AAGAAGCACAGAA". There are two alternative MC4R poly(A) sites downstream of ARAF insert. The potential G-Quadruplex motif is indicated as "G" in green circle which presents in A1WT, A2WT and A3WT. (B) The sequence of G-rich region 300-nucleotide downstream of the *ARAF* poly(A) site presents in construct A1WT, A2WT and A3WT. (C) G-Quadruplex prediction result of G-rich sequence (52 nucleotides long) in QGRS programme. Two potential G-Quadruplexes were predicted. The upper one is with very high G-score 59.

These six constructs were transfected into HEK 293 cells and poly(A) site activity was measured by RT-PCR as described in Fig 39A. To assess ARAF poly(A) site usage, the poly(A) site usage of "ARAF poly(A) site", "MC4R poly(A) site 2" and "MC4R poly(A) site 3" among all the six constructs were compared. As can be seen in Figure 39B lanes 1-3 ARAF poly(A) site, ARAF 3' end is efficiently processed in the construct A1WT, A2WT and A3WT. Surprisingly, ARAF poly(A) site usage is reduced in A4WT, A5WT and A6WT (Fig 39B lanes 4-6, MC4R poly(A) site 2 and MC4R poly(A) site 3, Fig 39C). Interestingly, the three constructs that result in reduced ARAF poly(A) site usage all lack the putative G-Quadruplex element which is the possible binding site for hnRNP H, an RNA binding protein that is able to strengthen otherwise weak poly(A) sites (Decorsiere, Cayrel et al. 2011). However, since A3WT with the G-Quadruplex sequence is 230bp longer than the A4WT (without the G-Quadruplex) and the length of the G-Quadruplex is only 30nt, there might be other unknown auxiliary elements present in A3WT but not in A4WT contributing to the 3' end formation of ARAF. Thus, we can only conclude from this that ARAF poly(A) site represents an example where the degeneration of a core cleavage and polyadenylation element is compensated for by the presence of long distance

enhancer elements, possibly the G-Quadruplex, in the 3' flanking region. This is very much reminiscent with an example previously described in our lab (Dalziel, Nunes et al. 2006).



	%
A1WT	88
A2WT	70
A3WT	62
A4WT	12
A5WT	7
A6WT	1

Figure 39 RT-PCR Analysis of ARAF (with six different lengths of fragments) 3' end formation in HEK 293 cells

A) Diagram depicting the construct, primers and output in the RT-PCR analysis. The forward primer in the GFP region and reverse primer oligo dT were used in RT-PCR analysis. The ARAF RT-PCR products "ARAF poly(A) signal", MC4R RT-PCR products "MC4R poly(A) signal 2" and "MC4R poly(A) signal 3" as indicated represent the possible outputs that can be detected by the RT-PCR analysis. B) RT-PCR analysis of total RNA isolated from HEK293 cells transfected with A1WT, A2WT, A3WT, A4WT, A5WT and A6WT (lanes 1 to 6) and from untreated HEK 293 cells (lane 7). The ARAF poly(A) site is indicated in blue boxes. MC4R poly(A) site 3 is indicated in red boxes. MC4R poly(A) site 2 cannot be detected in A1WT, A2WT and A3WT and

are the bands below red boxes in A4WT, A5WT and A6WT. Control represents samples without ARAF-MC4R plasmid transfection. *GAPDH* serves as an RNA loading control. C) ARAF poly(A) site usage in all six constructs. The quantitative method is described in Figure 36E. (n=3).

5.5 Discussion

In this study, we used a reporter gene and mutagenesis approach to investigate the sequences that are important for *ARAF* 3' end processing. No clear hexamer signal can be identified in the *ARAF* gene. A distant G-rich sequence in the 3' flanking region is possible to enhance the function of the ARAF poly(A) site with degenerated hexamer signal. This indicates that "no PAS" poly(A) sites may be reliant on the enhancer sequences. This is somewhat reminiscent to the situation in the *MC1R* gene where, a canonical hexamer is present, but the DSE is degenerated and compensated for by a distant G-rich enhancer (Dalziel, Nunes et al. 2006). The analysis on the ARAF poly(A) site shows that both the hexamer (Nunes, Li et al. 2010) and DSE (Dalziel, Nunes et al. 2006) can degenerate in the presence of strong enhancer elements. This further illustrates that for many poly(A) sites, the comparison of merely the core sequences cannot be used to make reasonable assumptions about their strength.

Chapter 6

Discussion

Chapter 6

6.1 Discussion

The *cis*-element requirement of ARAF and JunD poly(A) sites

Alternative cleavage and polyadenylation occurs when a single gene has multiple poly(A) sites. Like alternative splicing, APA results in multiple distinct RNA transcripts being generated from a single gene. Global analysis has found that approximately 70% of all human genes undergo APA (Derti, Garrett-Engele et al. 2012). The relative positions of these poly(A) sites in a gene will determine the coding and regulatory elements that are available in the different transcript isoforms and ultimately influence the transcripts isoform stability (Al-Ahmadi, Al-Ghamdi et al. 2009, Mayr and Bartel 2009, Dickson, Kruse et al. 2013), translational efficiency (Mayr and Bartel 2009, Spies, Burge et al. 2013, Whyteside, Turner et al. 2014) and protein identity (Alt, Bothwell et al. 1980, Ni and Kuperwasser 2016). APA has been found to be altered by many factors such as the relative level of specific protein factors and different cell types. Despite all this, the architecture of a single poly(A) site lay the foundation of the individual APA event. Without the thorough understanding of how a single poly(A) site is recognised, APA cannot be well explained.

In canonical cleavage and polyadenylation, the reaction is initiated by the interaction of CPSF complex with the canonical hexamer signal AAUAAA or AUUAAA normally 10-20 nucleotides upstream of the cleavage site and the interaction of CstF complex with the less defined DSE region downstream of

the cleavage site (Proudfoot 2011). With the CPSF and CstF complex anchored, around the poly(A) site, additional factors participate in the assembly of a complicated functional 3' end processing machinery. The cleavage occurs between the hexamer and the DSE and the cleavage site sequence arguably has some degree of preference for "CA" or "UA" dinucleotides in humans (Li and Du 2013). The 30% of the poly(A) sites that are not directed by AAUAAA/AUUAAA canonical hexamers (Neve, Burger et al. 2016) are defined as noncanonical poly(A) sites. These noncanonical poly(A) sites have been shown to be linked to alternative cleavage and polyadenylation (Wang, Dowell et al. 2013). Since the mutation from canonical hexamer to noncanonical hexamer generally results in a significant inhibition of 3' end processing (Wickens and Stephenson 1984, Sheets, Ogg et al. 1990), it is believed that the mechanism of noncanonical cleavage and polyadenylation is distinct. In addition, the mechanism is poorly understood because of limited examples studied.

We investigated the recognition of noncanonical poly(A) sites from an individual example *ARAF*, which lacks any described hexamer sequence. The only probable sequence element to serve as the role of hexamer is within the sequence "AAGAAGCACAGAA", 10 nucleotides upstream of the main cleavage site. Approximately 1.5kb of *ARAF* 3' UTR and 3' flanking region is not sufficient to direct its 3' end formation, suggesting some distant elements are needed for this highly unusual poly(A) site (Fig 36D). The sequence requirement of *ARAF* poly(A) site was investigated based on mutagenesis analysis. Two adenosines (2^{nd} and 5^{th} nucleotides in the putative signal) were found to be important for ARAF cleavage and polyadenylation (Fig 36D). From

the RNA secondary structure study of *ARAF* poly(A) site, a helical structure including the putative PAS is predicted by Oxfold (Fig 37). However, Early studies reveal that the canonical hexamer is recognised in a single-stranded form (Sittler, Gallinaro et al. 1995). Together with the evidence from mutagenesis analysis, we have shown that the accurate hexamer-like signal is severely degenerated and cannot be described since neither of the mutations can completely abolish its 3' end formation. A G-rich region approximately 300 nucleotides downstream of the cleavage site was highly likely to form a G-Quadruplex structure based on the QGRS G-Quadruplex prediction algorithm. The presence of this structure has been confirmed in p53 pre-mRNA and this G-Quadruplex has been shown to interact with hnRNP H/F to maintain 3' end formation during DNA damage (Decorsiere, Cayrel et al. 2011). The poly(A) site usage of ARAF decreased in ARAF-MC4R plasmids that lack the 230nt region containing the G-Quadruplex sequence (Fig 39B). Thus, it is indicated that G-Quadruplex or other unknown elements serve as auxiliary elements of *ARAF* to facilitate its 3' end processing. This is very similar to the architecture of the poly(A) sites found in the human papillomavirus genome, where the G-rich region has been found to regulate the expression of early and late genes (Oberg, Fay et al. 2005).

The study of the individual case *ARAF* is in agreement with our concept based on the study on *MC1R* (Dalziel, Nunes et al. 2006) which reveals that weak poly(A) sites may frequently require enhancer sequences located deep in the flanking regions to ensure their cleavage and polyadenylation. In the MC1R poly(A) site, the DSE region is degenerated and the distant G-rich element, which is 440nt downstream of the poly(A) site, is required for optimal 3' end

processing. Conversely, in *ARAF*, the hexamer poly(A) signal is retrogressive and some auxiliary element, possibly the G-Quadruplex, is suggested to promote the recognition of the poly(A) site. This indicates that a distant G-rich region can compensate for the degeneration of either the hexamer signal or DSE. This further illustrates that for many poly(A) sites, the sole comparison of core sequences cannot be used to make assumptions about their strength.

This thesis also described the sequence requirement of the AGUAAA poly(A) site JunD. The 3' end processing was directed by a surprisingly short fragment, suggesting no distant auxiliary element is needed (Fig 14C). Mutagenesis analysis revealed that 3' end formation of JunD was highly tolerant to nucleotide alteration (Fig 15C). These ineffective mutations alter the RNA secondary structure of wildtype JunD. Moreover, a significant helical structure in *JUND* predicted by Oxfold was turned out not to affect JunD cleavage and polyadenylation (Fig 12B and Fig 15C). Therefore, RNA secondary structure is not suggested as an important factor during JunD 3' end processing, even though AGUAAA poly(A) site associated structures have globally been predicted by Oxfold (Fig 10).

Only a short 42-nucleotides fragment containing the AGUAAA hexamer signal and a downstream element turned out to be necessary for JunD cleavage and polyadenylation. A novel adU-rich element immediately adjacent to the 3' of hexamer was identified to be critical for its optimal 3' end formation (Fig 15C). This adU-rich element was not required when a canonical hexamer AAUAAA was present (Fig 17D). Lengthening the distance between the adU-rich element and the hexamer decreased JunD poly(A) site usage (Fig 18B). On the other hand, shortening the distance between hexamer and U-rich region in another

AGUAAA poly(A) site Zip4 increased its poly(A) site usage (Fig 19B). These distance alteration mutations are not likely to change JunD and Zip4 RNA secondary structures in the same way in terms of the accessibility of the hexamer signal, the U-rich region and the DSE. Therefore, the distance itself appears to be an important factor for poly(A) site usage in JunD and Zip4. Furthermore, a U-rich motif has been identified (Fig 20) by MEME amongst many single AGUAAA poly(A) site genes in the “adjacent region” (between hexamer signal and cleavage site). Interestingly, a similar U-rich region matching the consensus described by MEME has been found adjacent to the A-rich region in *MC4R* poly(A) site, and this A-rich region has been reported to function equally well compared to a canonical hexamer signal (Nunes, Li et al. 2010). From the individual analysis in *JUND* and *ZIP4*, it is suggested that a similar U-rich region in MC4R can function as an adU-rich element to promote the recognition of A-rich region independent of the presence of hexamer signal. Compared to the DSE in *MC4R*, which well matches the consensus proposed (Tian and Graber 2012), the DSE in *JUND* cannot be described as an optimal element due to the lack of “GU” dinucleotide. In yeast and plants, the core poly(A) signals are an A-rich region instead of the hexamer signal and a U-rich region surrounding the cleavage site (Tian and Graber 2012). This is very similar to the architecture of the JunD and MC4R poly(A) site which suggests that there is a group of yeast/plant-like conserved poly(A) sites in human genes.

The hexamer signals in *ARAF* and *JUND* are both degenerated, but, the mechanism of recognising these two poly(A) sites are distinctive. Some distant auxiliary element, possibly G-rich enhancer, is indicated to be critical for the recognition of ARAF. In contrast, in JunD, only the adU-rich element has been

identified as enhancer. This again emphasises that the whole architecture of a poly(A) site rather than just the core sequence composition determines the strength of the poly(A) site.

Different impact of protein factors on JunD 3’ end formation

An essential step to assemble the 3’ end machinery and initiate 3’ end processing is the recruitment of the CPSF (Colgan and Manley 1997). This complex directly interacts with the hexamer via two subunits WDR33 and CPSF30 (Chan, Huppertz et al. 2014, Schönemann, Kühn et al. 2014). Whether the two subunits bind to pre-mRNAs concurrently or in a time and/or poly(A) type dependent manner remains unknown. The role of the CPSF160 subunit, previously strongly suggested to mediate the interaction of CPSF complex with pre-mRNA, needs to be further elucidated during the protein-RNA interaction. Nevertheless, CPSF160 is a critical component of the cleavage and polyadenylation machinery and establishes important protein-protein interactions (Murthy and Manley 1995). An RNAi based protein depletion strategy was employed to investigate the preference of the protein factors on different types of poly(A) sites. Interestingly, the knockdown of protein factors WDR33, CPSF30 and CPSF160 do not significantly change the noncanonical poly(A) site usage in the context of JunD WT and JunD Mut4a (Fig 22E, 23D, 24D). This suggests that both AGUAAA-adU-rich and canonical signals are efficiently recognised by any sub-complex of CPSF that may be formed in response to the low levels of a particular protein subunit. However, when no adU-rich element is present downstream of the AGUAAA which weakens the

site considerably, a preference between the canonical and the sub-optimal noncanonical poly(A) sites becomes apparent. WDR33 preferentially recognises sub-optimal noncanonical poly(A) sites when in competition with a canonical site. This is in contrast to CPSF30 and CPSF160, which preferentially recognise canonical poly(A) sites when they are in competition with a weak site (Fig 22E, 23D, 24D). These results suggest that the recognition of strong and weak poly(A) sites is dependent on the availability of CPSF factors that may form distinct CPSF sub-complexes (Schönemann, Kühn et al. 2014) in cells. The main function of hFip1 is to interact with poly(A) polymerase and bind to the U-rich regions that are located either up- or downstream of the hexamer on pre-mRNAs (Kaufmann, Martin et al. 2004). We initially thought that the adU-rich element discovered in JunD may interact with hFip1. However, the depletion of hFip1 does not obviously compromise 3' end processing in JunD WT, which is an AGUAAA-adU-rich type poly(A) site (Fig 25B). Therefore, our results suggest that the adU-rich element is not recognised by hFip1.

To date, the CFIm complex is arguably the best characterised factor that regulates alternative cleavage and polyadenylation at the point of cleavage. This is demonstrated by numerous examples where the depletion of two subunits CFIm68 and CFIm25 results in the preferred usage of the proximal poly(A) sites (Kubo, Wada et al. 2006, Kim, Yamamoto et al. 2010, Gruber, Martin et al. 2012, Masamha, Xia et al. 2014, Li, You et al. 2015). In contrast with the previously analyzed and described global influence of CFIm68 on alternative cleavage and polyadenylation, knockdown of CFIm68 does not lead to a shift from the distal site to the proximal site during JunD-MC4R 3' end formation in our Mini-gene system. Most importantly, in the JunD-MC4R poly(A)

site competition experiments, no increased usage was seen at the proximal site but instead usage of the distal site decreased which is evidenced by the appearance of the read through transcripts (Fig 26E, F). This is hugely significant because the shortening effect upon CFIm knockdown is widely attributed to an activation of the proximal poly(A) site. However, at this stage, we do not know whether the increase in read through at the distal site upon CFIm68 knockdown is a global phenomenon, or only an exceptional case in our JunD-MC4R system. Interestingly, read through transcripts can only be detected in the plasmids with AGUAAA poly(A) signal but disappear in Mut4a which has the AAUAAA poly(A) signal (Fig 26D). The UGUAN element, which is bound by CFIm68, has been shown to promote the recognition of the poly(A) site without a canonical signal (Venkataraman, Brown et al. 2005). This may suggest that the UGUAN motif found in the MC4R 3' UTR upstream of the JunD poly(A) site contributes to JunD 3' end formation. However, several lines of evidence contradict this interpretation. Firstly, no UGUAN motif can be found upstream of the endogenous *JUND* sequence; secondly, the UGUAN motif has been proved not to be required for a similar intronless gene *MC4R*; thirdly, the depletion of CFIm25 does not obviously affect JunD-MC4R 3' end formation (Fig 30B), as the recognition of the UGUAN motif is required cooperatively by both CFIm68 and CFIm25 subunits (Yang, Coseno et al. 2011). CFIm68 somehow compromises the proximal noncanonical poly(A) site and inhibits the canonical distal site in the JunD-MC4R system, and the mechanism under the impact of CFIm68 on this particular case is elusive. From the impact of CFIm68 on MC4R-JunD constructs, we have proposed an alternative hypothesis of the role of CFIm68 on regulation of alternative cleavage and polyadenylation.

From RT-PCR analysis, we do not find any evidence to show that CFIm25 influences JunD-MC4R 3' end formation (Fig 27C). This indicates that the impact of CFIm25 on APA is not identical with CFIm68 in some cases. Or, perhaps the knockdown does not cause a severe enough depletion of the CFIm25 and influence JunD-MC4R 3' end formation. To further elucidate the role of CFIm68 and CFIm25 on APA, the level of read through transcripts after depletion of either of these two factors should be investigated using global sequencing analysis.

Recent studies have revealed that both WDR33 and CPSF30 directly bind to AAUAAA and are required for CPSF complex to interact with RNA (Chan, Huppertz et al. 2014, Schönemann, Kühn et al. 2014). Interestingly, no increased levels of read through transcripts can be detected in our RNase protection assay when either factor is depleted (Fig 22D and Fig 23C). Whether the low levels of the two factors result in different CPSF sub-complexes, which can compensate for each other, needs to be clarified in the future. Alternatively, a UGUAN motif in MC4R 3' UTR could be recognised by CFIm and further promote the recruitment of CPSF as revealed in a noncanonical poly(A) site study (Venkataraman, Brown et al. 2005).

The global influence of WDR33, CPSF30 and SAM68

The poly(A) site type dependent recognition of WDR33 and CPSF30 has been suggested previously. In order to globally investigate the role of these two protein factors in the context of the recognition of canonical and noncanonical poly(A) sites, an RNAi based knockdown and sequencing strategy was applied.

Two gene categories, (i) the “AGTAAA” gene category contains genes with tandem UTR-APA poly(A) sites, one directed by the AGUAAA hexamer and another directed by the canonical hexamer; (ii) the “All noncanonical” gene category contains genes with tandem UTR-APA poly(A) sites, one directed by the canonical hexamer and another directed by any other poly(A) signal, were generated as the ideal environment to address this issue. Statistically, the distribution of noncanonical poly(A) site usage in the two protein knockdown samples is not significantly different from the distribution in negative control. This suggests that WDR33 and CPSF30 have no significant global influence on canonical and noncanonical poly(A) site selection. Motif analysis was employed to investigate whether the AGUAAA poly(A) sites with downregulated usage ratio caused by WDR33, or upregulated usage ratio caused by CPSF30, specifically encompassed a potential adU-rich element. Interestingly, no U-rich motif can be found downstream of the AGUAAA hexamer in any sub-groups as described. Comparatively, as mentioned, the potential adU-rich motif has been identified amongst single AGUAAA poly(A) site genes. Globally, in the “AGTAAA” gene category, the average AGUAAA poly(A) site usage is obviously lower than the canonical poly(A) site usage (Table 3, row AGUAAA vs row Canonical), which reveals that AGUAAA poly(A) sites without the adU-rich element are generally weaker than canonical poly(A) sites. Thus, the adU-rich element might be a safeguard element to aid AGUAAA to efficiently direct the usage of the only poly(A) site in the gene.

Although the global impact of CPSF30 and WDR33 on APA and canonical and noncanonical poly(A) site selection is not significant, individual cases of poly(A) site switching occur after the depletion of CPSF30 and WDR33 (Fig A3). This

suggests that CPSF30 and WDR33 favour certain kinds of poly(A) sites. Nevertheless, these sites cannot be described as AGUAAA lacking adU-rich elements or by the type of poly(A) signals. Interestingly, in some tandem poly(A) sites containing genes, the poly(A) site usage ratio of the AGUAAA poly(A) site, lacking an adU-rich element, is higher than the canonical poly(A) site usage ratio (Fig A4). This indicates that not only the adU-rich element can promote AGUAAA poly(A) site recognition, which suggests that the architecture of the AGUAAA poly(A) site is far more complicated than we expected. Without the understanding of the architecture of the noncanonical poly(A) sites, using the individual poly(A) site analysis described in Chapter 3, it would not be practical to generate an algorithm for accurately defining the strength of each individual poly(A) site, and further attempt to understand how global alternative cleavage and polyadenylation occurs.

The protein candidate SAM68 has been shown to be involved in viral 3' end formation and can bind to the RNA sequences UAAA and UUAA (Lin, Taylor et al. 1997, Itoh, Haga et al. 2002). In chapter 3, we have shown that this protein can influence the 3' end formation in an individual example construct MC4R-JunD-Mut4. However, from the knockdown and RNA-seq analysis, the global influence of SAM68 on canonical and noncanonical poly(A) site selection is not obvious. Interestingly, though globally the depletion of SAM68 results in few cases of APA changes, defined by the algorithm, some significant poly(A) site switching events are observed. And among these APA events, 3' UTR shortening is an obvious trend (Fig 33A). This suggests that SAM68 might influence 3' UTR processing in a small group of genes. Interestingly, the poly(A) sites with downregulation of poly(A) site usage do not necessarily lack the

UAAA containing hexamer signal (Fig A2), which indicates that SAM68 might not influence the mammalian 3’ end formation through interacting with the UAAA containing poly(A) signal and further inhibiting the poly(A) site usage. Thus, the mechanism under the 3’ end shortening caused by SAM68 in a small group of genes remains unclear.

6.2 Conclusion

Considering the results presented and discussed above, this project concludes that different mechanisms of the recognition of noncanonical poly(A) sites exist. A distance dependent adU-rich element adjacent to the hexamer is suggested to aid the recognition of AGUAAA signal in single poly(A) site genes and is required for the optimal JunD poly(A) site usage. Some auxiliary element, possibly the remote G-Quadruplex, is suggested to promote *ARAF* 3' end formation. The poly(A) sites are not recognised by protein factors in a simple core poly(A) signal type dependent manner. The strength of a poly(A) site is determined by the architecture rather than the core elements. The architecture of AGUAAA noncanonical poly(A) sites is more complicated than we expected and cannot be simply described by a hexamer, a DSE and an adU-rich element. To fully understand the mechanism of APA, further individual analysis based on different types of poly(A) sites needs to be carried out to elucidate different poly(A) site architectures.

Perhaps, in the future, it will be naïve to categorise poly(A) sites into canonical and noncanonical types.

6.3 Outlook

Firstly, for the individual poly(A) site study, the sequence requirement of Zip4 needs to be thoroughly investigated. This will not only elucidate the recognition of another AGUAAA poly(A) site but also interconnect the splicing process with noncanonical polyadenylation.

Secondly, the importance of adU-rich element on MC4R 3' end formation needs to be studied. The canonical hexamer together with the adU-rich element will be deleted to evaluate the 3' processing efficiency of the first poly(A) site of MC4R. From this experiment, we will know whether the adU-rich element can also enhance the usage of other types of noncanonical poly(A) sites.

Thirdly, the G-rich sequence possibly forming G-Quadruplex needs to be accurately deleted to assess its importance on ARAF cleavage and polyadenylation.

Finally, the level of read through transcripts need to be investigated globally upon the depletion of CFIm68. This sequencing experiment will enable us to know whether the knockdown of CFIm68 causes global poly(A) site shortening or increase of the read through transcripts downstream of the distal poly(A) sites.

Chapter 7

Appendix

Chapter 7

7.1 Supplementary figures

A

Sequence file	173randompa.fasta
Distribution of motif occurrences	Zero or one per sequence
Number of different motifs	5
Minimum number of sites	50
Maximum number of sites	160
Minimum motif width	6
Maximum motif width	10
Searching	given stand only

B

Figure A1 Motif investigation amongst randomly selected poly(A) sites

A) The parameter used to search for motifs among 173 randomly selected genes. B) Motif anlysis result calculated by MEME. Sequences downstream of the poly((A) signal upstream of the cleavage site in all 173 randomly selected poly(A) sites were applied in MEME. Higher nucleotides stand for a higher probability in that position.

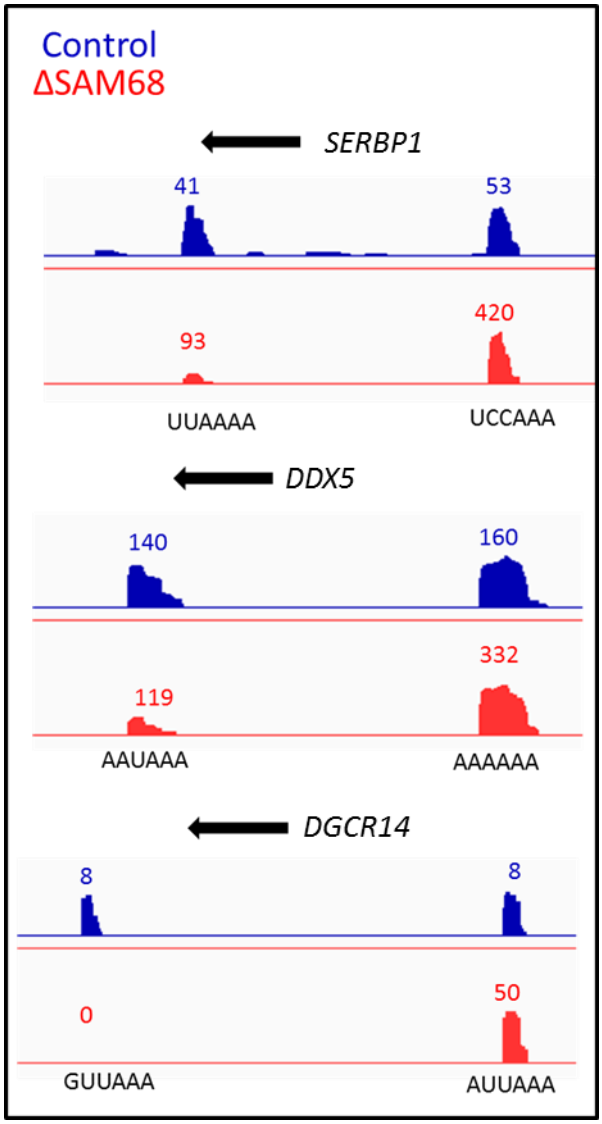


Figure A2 Individual cases of poly(A) site switch caused by the depletion of SAM68.

Visualisation of the sequencing data caused by depletion of SAM68 in IGV software. Gene name, hexamer of the poly(A) site and rpm is as indicated.

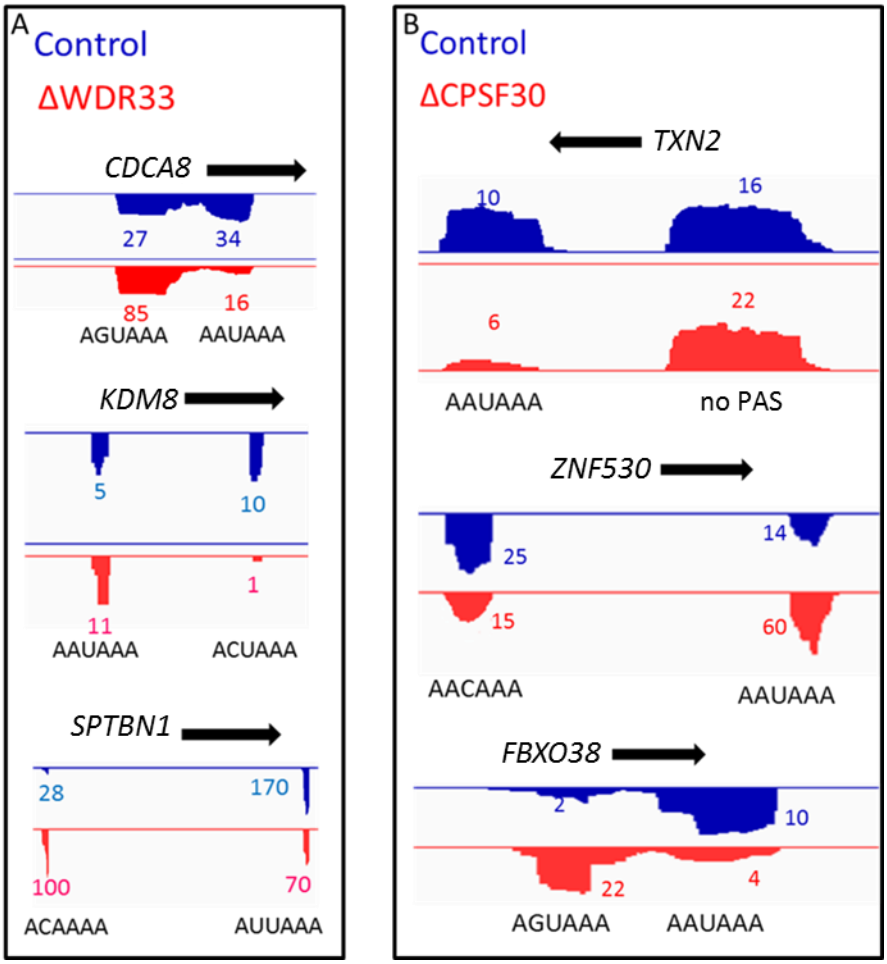


Figure A3 Individual cases of poly(A) site switch caused by the depletion of WDR33 and CPSF30.

Visualisation of the sequencing data caused by depletion of (A) WDR33 and (B) CPSF30 in IGV software. Gene name, hexamer of the poly(A) site and rpm is as indicated.

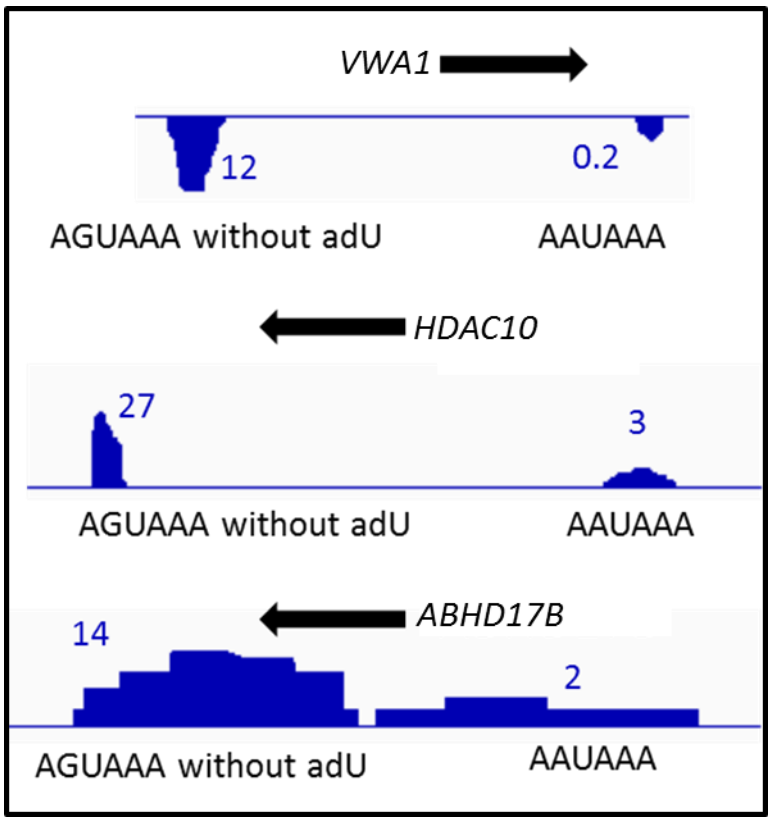


Figure A4 Individual cases of AGUAAA poly(A) sites lack adU-rich regions stronger than AAUAAA poly(A) sites.

Visualisation of the sequencing data in the negative control in IGV software. Gene name, type of poly(A) site and rpm is as indicated.

7.2 Supplementary tables

Table A1 Two-tailed paired t-test compares the JunD poly(A) site usage between the wildtype (293) and knockdown data in RNase protection assay and RT PCR. Significant level: ns: not significant, *: P<0.05; **: P<0.01

RP	Test pair			
Plasmid	Type		P value	Significant level
JunD WT	293	ΔWDR33	0.072827	ns
	293	ΔCPSF30	0.22626	ns
	293	ΔCPSF160	0.065449	ns
Mut1	293	ΔWDR33	0.183503	ns
	293	ΔCPSF30	0.225403	ns
	293	ΔCPSF160	0.225403	ns
Mut4	293	ΔWDR33	0.009684	**
	293	ΔCPSF30	0.013072	*
	293	ΔCPSF160	0.003492	**
Mut4a	293	ΔWDR33	0.678366	ns
	293	ΔCPSF30	0.076619	ns
	293	ΔCPSF160	0.6116	ns

RT-PCR	Test pair			
Plasmid	Type		P value	Significant level
JunD WT	293	ΔhFip1	0.18895	ns
	293	ΔCFIm25	0.23769	ns
	293	ΔSAM68	0.13256	ns
Mut1	293	ΔhFip1	0.223098	ns
	293	ΔCFIm25	0.222276	ns
	293	ΔSAM68	0.225117	ns
Mut4	293	ΔhFip1	0.007657	**
	293	ΔCFIm25	0.131498	ns
	293	ΔSAM68	0.00294	**
Mut4a	293	ΔhFip1	0.58798	ns
	293	ΔCFIm25	0.237698	ns
	293	ΔSAM68	0.328796	ns

7.3 List of plasmids constructed for the project

Plasmid name	Usage	Insert	Backbone
pUC_MC4Rwtd1	Mini-gene transfection		pUC18
pUC_JunDwt1MC4R	Mini-gene transfection	*JUND*	pUC18
pUC_JunDwt2MC4R	Mini-gene transfection	*JUND*	pUC18
pUC_JunDwt3MC4R	Mini-gene transfection	*JUND*	pUC18
pUC_JunDwt4MC4R	Mini-gene transfection	*JUND*	pUC18
pUC_JunDM1MC4R	Mini-gene transfection	*JUND*	pUC18
pUC_JunDM2MC4R	Mini-gene transfection	*JUND*	pUC18
pUC_JunDM3MC4R	Mini-gene transfection	*JUND*	pUC18
pUC_JunDM4MC4R	Mini-gene transfection	*JUND*	pUC18
pUC_JunDM5MC4R	Mini-gene transfection	*JUND*	pUC18
pUC_JunDM6MC4R	Mini-gene transfection	*JUND*	pUC18
pUC_JunDM7MC4R	Mini-gene transfection	*JUND*	pUC18
pUC_JunDM4aMC4R	Mini-gene transfection	*JUND*	pUC18
pUC_JunDMDMC4R	Mini-gene transfection	*JUND*	pUC18
pUC_Zip4wtMC4R	Mini-gene transfection	*Zip4*	pUC18
pUC_Zip4shMC4R	Mini-gene transfection	*Zip4*	pUC18
pUC_A1WTMC4R	Mini-gene transfection	*ARAF*	pUC18
pUC_A1M1MC4R	Mini-gene transfection	*ARAF*	pUC18
pUC_A1M2MC4R	Mini-gene transfection	*ARAF*	pUC18
pUC_A2WTMC4R	Mini-gene transfection	*ARAF*	pUC18
pUC_A3WTMC4R	Mini-gene transfection	*ARAF*	pUC18
pUC_A4WTMC4R	Mini-gene transfection	*ARAF*	pUC18
pUC_A5WTMC4R	Mini-gene transfection	*ARAF*	pUC18
pUC_A6WTMC4R	Mini-gene transfection	*ARAF*	pUC18
pG4_Jwt	RNase protection	*JUND, MC4R*	pGEM4
pG4_JM1	RNase protection	*JUND, MC4R*	pGEM4
pG4_JM4	RNase protection	*JUND, MC4R*	pGEM4
pG4_JM4a	RNase protection	*JUND, MC4R*	pGEM4

7.4 Oligonucleotides used in the project

MC4R cloning primers	
Primer labels	Sequence (5' - 3')
MC4R F1	CCGCGACTCTAGATTAAATGGGG
MC4R R1	ATGCTCGAGGTAAATCCACAGT
MC4R F2	TACCTCGAGCATAAGCATTGGAC
MC4R R2	CAGTGCCAAGCTTGCATGCCA
JunD cloning primers	
Primer labels	Sequence (5' - 3')
JunD WT1 F	TCGAAGTAAAGTCTCGTTACGCCAGCTCGGCTCTCCGCCTCCTTCTT
JunD WT1 R	TCGAAAGAAGGAGGCGGAGAGCCGAGCTGGCGTAACGAGACTTTACT
JunD WT F2	ATGCTCGAGAGTAAAGTCTCGTTACGCCAGC
JunD WT R2	GATCTCGAGGGGGTGTAGGGGGGAAGCGAA
JunD WT R3	GTACTCGAGCTCCCTCCAACAAGGGGAGGA
JunD WT R4	CGACTCGAGCGCTCGCTGCGCGCCAGATGT
JunD Mut1 F	TCGAAGCAACGTCTCGTTACGCCAGCTCGGCTCTCCGCCTCCTTCTT
JunD Mut1 R	TCGAAAGAAGGAGGCGGAGAGCCGAGCTGGCGTAACGAGACGTTGCT
JunD Mut2 F	TCGAAGTAAAGTCTCGCAACGCCAGCTCGGCTCTCCGCCTCCTTCTT
JunD Mut2 R	TCGAAAGAAGGAGGCGGAGAGCCGAGCTGGCGTTGCGAGACTTTACT
JunD Mut3 F	TCGAAGTAAAGTCTCGTTACGCCAGCTCGGATCTCAGCATCCTTCAT
JunD Mut3 R	TCGAATGAAGGATGCTGAGATCCGAGCTGGCGTAACGAGACTTTACT
JunD Mut4 F	TCGAAGTAAAGCCCCGCCACGCCAGCTCGGCTCTCCGCCTCCTTCCT
JunD Mut4 R	TCGAAGGAAGGAGGCGGAGAGCCGAGCTGGCGTGGCGGGGCTTTACT
JunD Mut5 F	TCGAAGTAAAGTCTCGTTACGCGGGCTCGGCTCTCCGCCTCCTTCCT
JunD Mut5 R	TCGAAGGAAGGAGGCGGAGAGCCGAGCCCGCGTAACGAGACTTTACT
JunD Mut6 F	TCGAAATAAAGTCTCGTTACGCCAGCTCGGCTCTCCGCCTCCTTCCT
JunD Mut6 R	TCGAAGGAAGGAGGCGGAGAGCCGAGCTGGCGTAACGAGACTTTATT
JunD Mut7 F	TCGAAGTAAAGTCTCGTTACGCCAGCTCGGCACACCGCCACCAACCA
JunD Mut7 R	TCGATGGTTGGTGGCGGTGTGCCGAGCTGGCGTAACGAGACTTTACT
JunD Mut4a F	TCGAAATAAAGCCCCGCCACGCCAGCTCGGCTCTCCGCCTCCTTCCT
JunD Mut4a R	TCGAAGGAAGGAGGCGGAGAGCCGAGCTGGCGTGGCGGGGCTTTATT
JunD MutD F	TCGAAGTAAAACACGTCTCGTTACGCCAGCTCGGCTCTCCGCCTCCTTCTT
JunD MutD R	TCGAAAGAAGGAGGCGGAGAGCCGAGCTGGCGTAACGAGACGTGTTTTACT
Zip4 cloning primers	
Primer labels	Sequence (5' - 3')
Zip4 WT F	TCGAAGTAAAGACACTCTTGTCCTTGGAGCATGGCTGTGCTCTCCTTGCTGGGTG
Zip4 WT R	TCGACACCCAGCAAGGAGAGCACAGCCATGCTCCAAGGACAAGAGTGTCTTTACT
Zip4 sh F	TCGAAGTAAAGTCTTGTCCTTGGAGCATGGCTGTGCTCTCCTTGCTGGGTG
Zip4 sh R	TCGACACCCAGCAAGGAGAGCACAGCCATGCTCCAAGGACAAGACTTTACT

ARAF cloning primers	
Primer labels	Sequence (5' - 3')
ARAF WT F1	CGACTCTAGAGCCCCGCCCAAGCCACCAGGGA
ARAF WT R1	ACGCTCGAGCTGAGACCATCCGCAGCCTGAG
ARAF WT R2	GGTGAAGAAAGAGAGGAGTCCTTG
ARAF F Mut1	ACGACGCACAGAATTCTGCTGGG
ARAF F Mut2	AAGAAGCACAGACTTCTGCTGGG
ARAF WT F2	ATGTCTAGATTTGGGGATACTTCTAAATTTTGGGAGCT
ARAF WT R3	ACGCTCGAGTCCAAATCCTCAAGAAACCCCTAC
ARAF WT F3	ATGTCTAGAATTTTGGGGTTCCCAGCACC
ARAF WT R4	ACGCTCGAGGGATTTCCCTTCACCAGCTCC
ARAF WT R5	ACGCTCGAGCTCCTTTTGTGCCGTGCTTG
ARAF WT R6	ACGCTCGAGATTTTGGGGTTCCCAGCACC
ARAF WT R7	ACGCTCGAGCCCAAGAGAAGAGTCGGGAAACA
RT-PCR primers	
Primer labels	Sequence (5' - 3')
MC4R F1	TCACTCTCGGCATGGACGAG
UAP	GCCCACGCGTCGACTAGTACT
AP	GCCCACGCGTCGACTAGTACTTTTTTTTTTTTTTT
GAPDH F1	CTGAGAACGGGAAGCTTGTC
GAPDH R1	CGTCCACCACTGACACGTTG
Fip1 F	GAACGATACAGATACAGGGAATATGC
Fip1 R	TCGGTGCTCTCCTGTTCAG
Malat RT	GACCCTACTGAAGAGCATTGGAGA
Malat F	AGAGAATTCCCTGCAAATTGTTAACAGAA
Malat R	TCAGCTTCCGCTAAGATGCTAGCTT
CDCAf	GGAGCAGTGTTTTATGTTTATCCCTG
TXNf	AGTGCTTCCGAGCTGGCGGACTGCCCAGGGG
CYFp	ACCCTTATCTGGAGGAGGAAGAG
CYFd	TTTCCCTGGCTTGCCAGTG
RP primers	
Primer labels	Sequence (5' - 3')
RPJDf_EcoRI	ATCGAATTCTGGCCATAAAATATGAATCTATGTT
RPJDr_Hind3	TGCAAGCTTCCAGGAGCTTCATTTCTACTTCTGC
pGEM4seqR	GCAGCTGGCTTATCGAAATTAATAC
pGEM4seqF	CCTTATGTATCATACACATACGATTTAG

Chapter 8

Bibliography

Chapter 8

Adelman, K. and J. T. Lis (2012). "Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans." Nature Reviews Genetics **13**(10): 720-731.
Agarwal, S. K., S. C. Guru, C. Heppner, M. R. Erdos, R. M. Collins, S. Y. Park, S. Saggar, S. C. Chandrasekharappa, F. S. Collins, A. M. Spiegel, S. J. Marx and A. L. Burns (1999). "Menin Interacts with the AP1 Transcription Factor JunD and Represses JunD-Activated Transcription." Cell **96**(1): 143-152.
Ahmed, Y. F., G. M. Gilmartin, S. M. Hanly, J. R. Nevins and W. C. Greene (1991). "The HTLV-I rex response element mediates a novel form of mRNA polyadenylation." Cell **64**(4): 727-737.
Al-Ahmadi, W., M. Al-Ghamdi, L. Al-Haj, M. Al-Saif and K. S. Khabar (2009). "Alternative polyadenylation variants of the RNA binding protein, HuR: abundance, role of AU-rich elements and auto-Regulation." Nucleic Acids Res **37**(11): 3612-3624.
Alt, F. W., A. L. Bothwell, M. Knapp, E. Siden, E. Mather, M. Koshland and D. Baltimore (1980). "Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends." Cell **20**(2): 293-301.
Anderson, J. W. J., P. A. Haas, L.-A. Mathieson, V. Volynkin, R. Lyngsø, P. Tataru and J. Hein (2013). "Oxfold: kinetic folding of RNA using stochastic context-free grammars and evolutionary information." Bioinformatics **29**(6): 704-710.
Ara, T., F. Lopez, W. Ritchie, P. Benech and D. Gautheret (2006). "Conservation of alternative polyadenylation patterns in mammalian genes." BMC Genomics **7**(1): 189.
Archambault, J. and J. D. Friesen (1993). "Genetics of eukaryotic RNA polymerases I, II, and III." Microbiological reviews **57**(3): 703-724.
Ashe, M. P., A. Furger and N. J. Proudfoot (2000). "Stem-loop 1 of the U1 snRNP plays a critical role in the suppression of HIV-1 polyadenylation." RNA (New York, N.Y.) **6**(2): 170-177.
Ashe, M. P., P. Griffin, W. James and N. J. Proudfoot (1995). "Poly(A) site selection in the HIV-1 provirus: inhibition of promoter-proximal polyadenylation by the downstream major splice donor site." Genes & Development **9**(23): 3008-3025.
Ast, G. (2004). "How did alternative splicing evolve?" Nature Reviews Genetics **5**(10): 773-782.
Avery, O. T., C. M. MacLeod and M. McCarty (1944). "STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III." The Journal of Experimental Medicine **79**(2): 137-158.
Bai, Y., T. C. Auperin, C.-Y. Chou, G.-G. Chang, J. L. Manley and L. Tong (2007). "Crystal Structure of Murine CstF-77: Dimeric Association and Implications for Polyadenylation of mRNA Precursors." Molecular Cell **25**(6): 863-875.
Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li and W. S. Noble (2009). "MEME Suite: tools for motif discovery and searching." Nucleic Acids Research **37**(suppl 2).
Baljuls, A., W. Schmitz, T. Mueller, R. P. Zahedi, A. Sickmann, M. Hekman and U. R. Rapp (2008). "Positive Regulation of A-RAF by Phosphorylation of Isoform-specific Hinge Segment and Identification of Novel Phosphorylation Sites." Journal of Biological Chemistry **283**(40): 27239-27254.
Barlat, I., F. Maurier, M. Duchesne, E. Guitard, B. Tocque and F. Schweighoffer (1997). "A Role for Sam68 in Cell Cycle Progression Antagonized by a Spliced Variant within the KH Domain." Journal of Biological Chemistry **272**(6): 3129-3132.

Bennett, C., M. Brunkow, F. Ramsdell, K. O'Briant, Q. Zhu, R. Fuleihan, A. Shigeoka, H. Ochs and P. Chance (2001). "A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA→ AAUGAA) leads to the IPEX syndrome." Immunogenetics **53**(6): 435-439.
Berget, S. M. (1995). "Exon Recognition in Vertebrate Splicing." Journal of Biological Chemistry **270**(6): 2411-2414.
Bienroth, S., W. Keller and E. Wahle (1993). "Assembly of a processive messenger RNA polyadenylation complex." The EMBO journal **12**(2): 585-594.
Bienroth, S., E. Wahle, C. Suter-Crazzolara and W. Keller (1991). "Purification of the cleavage and polyadenylation factor involved in the 3'-processing of messenger RNA precursors." The Journal of biological chemistry **266**(29): 19768-19776.
Black, D. L. (2003). "MECHANISMS OF ALTERNATIVE PRE-MESSENGER RNA SPLICING." Biochemistry **72**(1): 291-336.
Brackenridge, S., H. L. Ashe, M. Giacca and N. J. Proudfoot (1997). "Transcription and polyadenylation in a short human intergenic region." Nucleic Acids Research **25**(12): 2326-2335.
Brown, K. M. and G. M. Gilmartin (2003). "A Mechanism for the Regulation of Pre-mRNA 3′ Processing by Human Cleavage Factor Im." Molecular Cell **12**(6): 1467-1476.
Calvo, O. and J. L. Manley (2001). "Evolutionarily Conserved Interaction between CstF-64 and PC4 Links Transcription, Polyadenylation, and Termination." Molecular Cell **7**(5): 1013-1023.
Cañadillas, J. and G. Varani (2003). "Recognition of GU‐rich polyadenylation regulatory elements by human CstF‐64 protein." The EMBO Journal **22**(11): 2821-2830.
Carey, M. F., C. L. Peterson and S. T. Smale (2013). "The RNase Protection Assay." Cold Spring Harbor Protocols **2013**(3).
Carmody, S. R. and S. R. Wente (2009). "mRNA nuclear export at a glance." Journal of Cell Science **122**(12): 1933-1937.
Carswell, S. and J. C. Alwine (1989). "Efficiency of utilization of the simian virus 40 late polyadenylation site: effects of upstream sequences." Molecular and cellular biology **9**(10): 4248-4258.
Cartegni, L., S. L. Chew and A. R. Krainer (2002). "Listening to silence and understanding nonsense: exonic mutations that affect splicing." Nature Reviews Genetics **3**(4): 285-298.
Castello, A., B. Fischer, K. Eichelbaum, R. Horos, B. M. Beckmann, C. Strein, N. E. Davey, D. T. Humphreys, T. Preiss, L. M. Steinmetz, J. Krijgsveld and M. W. Hentze (2012). "Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins." Cell **149**(6): 1393-1406.
Chan, S., E. A. Choi and Y. Shi (2011). "Pre‐mRNA 3′‐end processing complex assembly and function." Wiley Interdisciplinary Reviews: RNA **2**(3): 321-335.
Chan, S. L., I. Huppertz, C. Yao, L. Weng, J. J. Moresco, J. R. Yates, J. Ule, J. L. Manley and Y. Shi (2014). "CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3′ processing." Genes & Development **28**(21): 2370-2380.
Chang, H., J. Lim, M. Ha and V. N. Kim (2014). "TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications." Mol Cell **53**(6): 1044-1052.
Chao, L. C., A. Jamil, S. J. Kim, L. Huang and H. G. Martinson (1999). "Assembly of the Cleavage and Polyadenylation Apparatus Requires About 10 Seconds In Vivo and Is Faster for Strong than for Weak Poly(A) Sites." Molecular and Cellular Biology **19**(8): 5588-5600.
Chen, F., C. C. MacDonald and J. Wilusz (1995). "Cleavage site determinants in the mammalian polyadenylation signal." Nucleic acids research **23**(14): 2614-2620.
Chen, M. and J. L. Manley (2009). "Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches." Nature Reviews Molecular Cell Biology **10**(11): 741-754.
Chiba, K., J. Yamamoto, Y. Yamaguchi and H. Handa (2010). "Promoter-proximal pausing and its release: Molecular mechanisms and physiological functions." Experimental Cell Research **316**(17): 2723-2730.

Cho, E.-J., T. Takagi, C. R. Moore and S. Buratowski (1997). "mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain." Genes & Development **11**(24): 3319-3326.
Church, G. M. (2006). "Genomes for All." Scientific American **294**(1): 46-54.
Colgan, D., K. G. K. Murthy, C. Prives and J. L. Manley (1996). "Cell-cycle related regulation of poly(A) polymerase by phosphorylation." Nature **384**(6606): 282-285.
Colgan, D. F. and J. L. Manley (1997). "Mechanism and regulation of mRNA polyadenylation." Genes & Development **11**(21): 2755-2766.
Connelly, S. and J. L. Manley (1988). "A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II." Genes & development **2**(4): 440-452.
Conrad, N. K. and J. A. Steitz (2005). "A Kaposi's sarcoma virus RNA element that increases the nuclear abundance of intronless transcripts." The EMBO Journal **24**(10): 1831-1841.
Cook, P. R. (1999). "The Organization of Replication and Transcription." Science **284**(5421): 1790-1795.
Cooke, C. and J. C. Alwine (1996). "The cap and the 3' splice site similarly affect polyadenylation efficiency." Molecular and Cellular Biology **16**(6): 2579-2584.
Crick, F. (1970). "Central Dogma of Molecular Biology." Nature **227**(5258): 561-563.
Dalziel, M., N. Nunes and A. Furger (2006). "Two G-rich regulatory elements located adjacent to and 440 nucleotides downstream of the core poly(A) site of the intronless melanocortin receptor 1 gene are critical for efficient 3' end processing." Molecular and cellular biology **27**(5): 1568-1580.
Danckwardt, S., I. Kaufmann, M. Gentzel, K. U. Foerstner, A. S. Gantzert, N. H. Gehring, G. Neu‐Yilik, P. Bork, W. Keller, M. Wilm, M. W. Hentze and A. E. Kulozik (2007). "Splicing factors stimulate polyadenylation via USEs at non‐canonical 3′ end formation signals." The EMBO Journal **26**(11): 2658-2669.
Dantonel, J.-C., K. G. K. Murthy, J. L. Manley and L. Tora (1997). "Transcription factor TFIID recruits factor CPSF for formation of 3′ end of mRNA." Nature **389**(6649): 399-402.
Darty, K., A. Denise and Y. Ponty (2009). "VARNA: Interactive drawing and editing of the RNA secondary structure." Bioinformatics **25**(15): 1974-1975.
de Vries, H., U. Rüegsegger, W. Hübner, A. Friedlein, H. Langen and W. Keller (2000). "Human pre‐mRNA cleavage factor IIm contains homologs of yeast proteins and bridges two other cleavage factors." The EMBO Journal **19**(21): 5895-5904.
Decorsiere, A., A. Cayrel, S. Vagner and S. Millevoi (2011). "Essential role for the interaction between hnRNP H/F and a G quadruplex in maintaining p53 pre-mRNA 3'-end processing and function during DNA damage." Genes & Development **25**(3): 220-225.
Decorsière, A., A. Cayrel, S. Vagner and S. Millevoi (2011). "Essential role for the interaction between hnRNP H/F and a G quadruplex in maintaining p53 pre-mRNA 3'-end processing and function during DNA damage." Genes & development **25**(3): 220-225.
Delaney, K. J., R. Xu, J. Zhang, Q. Q. Li, K.-Y. Yun, D. L. Falcone and A. G. Hunt (2006). "Calmodulin interacts with and regulates the RNA-binding activity of an Arabidopsis polyadenylation factor subunit." Plant physiology **140**(4): 1507-1521.
Denome, R. M. and C. N. Cole (1988). "Patterns of polyadenylation site selection in gene constructs containing multiple polyadenylation signals." Molecular and Cellular Biology **8**(11): 4829-4839.
Deo, R. C., J. B. Bonanno, N. Sonenberg and S. K. Burley (1999). "Recognition of Polyadenylate RNA by the Poly(A)-Binding Protein." Cell **98**(6): 835-845.
Derti, A., P. Garrett-Engele, K. D. MacIsaac, R. C. Stevens, S. Sriram, R. Chen, C. A. Rohl, J. M. Johnson and T. Babak (2012). "A quantitative atlas of polyadenylation in five mammals." Genome Research **22**(6): 1173-1183.

DeZazzo, J. D., J. E. Kilpatrick and M. J. Imperiale (1991). "Involvement of long terminal repeat U3 sequences overlapping the transcription control region in human immunodeficiency virus type 1 mRNA 3' end formation." Molecular and cellular biology **11**(3): 1624-1630.
Dichtl, B., D. Blank, M. Sadowski, W. Hübner, S. Weiser and W. Keller (2002). "Yhh1p/Cft1p directly links poly(A) site recognition and RNA polymerase II transcription termination." The EMBO Journal **21**(15): 4125-4135.
Dickson, J. R., C. Kruse, D. R. Montagna, B. Finsen and M. S. Wolfe (2013). "Alternative polyadenylation and miR-34 family members regulate tau expression." J Neurochem **127**(6): 739-749.
Dominski, Z., X.-c. Yang, M. Purdy, E. J. Wagner and W. F. Marzluff (2005). "A CPSF-73 Homologue Is Required for Cell Cycle Progression but Not Cell Growth and Interacts with a Protein Having Features of CPSF-100." Molecular and Cellular Biology **25**(4): 1489-1500.
Doty, P., H. Boedtker, J. R. Fresco, R. Haselkorn and M. Litt (1959). "SECONDARY STRUCTURE IN RIBONUCLEIC ACIDS." Proceedings of the National Academy of Sciences **45**(4): 482-499.
Dower, K. E. N. and M. Rossh (2002). "T7 RNA polymerase-directed transcripts are processed in yeast and link 3??? end formation to mRNA nuclear export." RNA **8**(5): 686-697.
Dufner-Beattie, J., F. Wang, Y.-M. Kuo, J. Gitschier, D. Eide and G. K. Andrews (2003). "The Acrodermatitis Enteropathica Gene ZIP4 Encodes a Tissue-specific, Zinc-regulated Zinc Transporter in Mice." Journal of Biological Chemistry **278**(35): 33474-33481.
Dye, M. J. and N. J. Proudfoot (1999). "Terminal Exon Definition Occurs Cotranscriptionally and Promotes Termination of RNA Polymerase II." Molecular Cell **3**(3): 371-378.
Edmonds, M., M. H. Vaughan and H. Nakazato (1971). "Polyadenylic Acid Sequences in the Heterogeneous Nuclear RNA and Rapidly-Labeled Polyribosomal RNA of HeLa Cells: Possible Evidence for a Precursor Relationship." Proceedings of the National Academy of Sciences **68**(6): 1336-1340.
Eksi, R., H.-D. Li, R. Menon, Y. Wen, G. S. Omenn, M. Kretzler and Y. Guan (2013). "Systematically Differentiating Functions for Alternatively Spliced Isoforms through Integrating RNA-seq Data." PLoS Computational Biology **9**(11).
Fabrega, C., V. Shen, S. Shuman and C. D. Lima (2003). "Structure of an mRNA Capping Enzyme Bound to the Phosphorylated Carboxy-Terminal Domain of RNA Polymerase II." Molecular Cell **11**(6): 1549-1561.
Fechter, P. and G. G. Brownlee (2005). "Recognition of mRNA cap structures by viral and cellular proteins." The Journal of general virology **86**(Pt 5): 1239-1249.
Fica, S. M., N. Tuttle, T. Novak, N.-S. Li, J. Lu, P. Koodathingal, Q. Dai, J. P. Staley and J. A. Piccirilli (2013). "RNA catalyses nuclear pre-mRNA splicing." Nature **503**(7475): 229-234.
Fitzgerald, M. and T. Shenk (1981). "The sequence 5'-AAUAAA-3'forms parts of the recognition site for polyadenylation of late SV40 mRNAs." Cell **24**(1): 251-260.
Fong, N. and D. L. Bentley (2001). "Capping, splicing, and 3′ processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD." Genes & Development **15**(14): 1783-1795.
Fortes, P., Y. Cuevas, F. Guan, P. Liu, S. Pentlicky, S. P. Jung, M. L. Martínez-Chantar, J. Prieto, D. Rowe and S. I. Gunderson (2003). "Inhibiting expression of specific genes in mammalian cells with 5′ end-mutated U1 small nuclear RNAs targeted to terminal exons of pre-mRNA." Proceedings of the National Academy of Sciences **100**(14): 8264-8269.
Fox-Walsh, K., J. Davis-Turak, Y. Zhou, H. Li and X.-D. Fu (2011). "A multiplex RNA-seq strategy to profile poly(A+) RNA: Application to analysis of transcription response and 3′ end formation." Genomics **98**(4): 266-271.
Fresco, J. R., B. M. Alberts and P. Doty (1960). "Some Molecular Details of the Secondary Structure of Ribonucleic Acid." Nature **188**(4745): 98-101.

Fresco, L. D. and S. Buratowski (1996). "Conditional mutants of the yeast mRNA capping enzyme show that the cap enhances, but is not required for, mRNA splicing." RNA (New York, N.Y.) **2**(6): 584-596.
Fu, Y., Y. Sun, Y. Li, J. Li, X. Rao, C. Chen and A. Xu (2011). "Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing." Genome Research **21**(5): 741-747.
Furth, P. A., W. T. Choe, J. H. Rex, J. C. Byrne and C. C. Baker (1994). "Sequences homologous to 5' splice sites are required for the inhibitory activity of papillomavirus late 3' untranslated regions." Molecular and cellular biology **14**(8): 5278-5289.
Gehring, N. H., U. Frede, G. Neu-Yilik, P. Hundsdoerfer, B. Vetter, M. W. Hentze and A. E. Kulozik (2001). "Increased efficiency of mRNA 3′ end formation: a new genetic mechanism contributing to hereditary thrombophilia." Nature Genetics **28**(4): 389-392.
Gil, A. and N. J. Proudfoot (1987). "Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit β-globin mRNA 3′ end formation." Cell **49**(3): 399-406.
Gilmartin, G. M., E. S. Fleming and J. Oetjen (1992). "Activation of HIV-1 pre-mRNA 3' processing in vitro requires both an upstream element and TAR." The EMBO journal **11**(12): 4419-4428.
Gilmartin, G. M. and J. R. Nevins (1989). "An ordered pathway of assembly of components required for polyadenylation site recognition and processing." Genes & Development **3**(12b): 2180-2190.
Gilmartin, G. M. and J. R. Nevins (1991). "Molecular analyses of two poly(A) site-processing factors that determine the recognition and efficiency of cleavage of the pre-mRNA." Molecular and cellular biology **11**(5): 2432-2438.
Graber, J. H., C. R. Cantor, S. C. Mohr and T. F. Smith (1999). "Genomic detection of new yeast pre-mRNA 3′-end-processing signals." Nucleic Acids Research **27**(3): 888-894.
Graveley, B. R., E. S. Fleming and G. M. Gilmartin (1996). "Restoration of Both Structure and Function to a Defective Poly(A) Site by in Vitro Selection." Journal of Biological Chemistry **271**(52): 33654-33663.
Graveley, B. R., E. S. Fleming and G. M. Gilmartin (1996). "RNA structure is a critical determinant of poly(A) site recognition by cleavage and polyadenylation specificity factor." Molecular and cellular biology **16**(9): 4942-4951.
Gruber, A. R., G. Martin, W. Keller and M. Zavolan (2012). "Cleavage factor Im is a key regulator of 3' UTR length." RNA biology **9**(12): 1405-1412.
Guhaniyogi, J. and G. Brewer (2001). "Regulation of mRNA stability in mammalian cells." Gene **265**(1-2): 11-23.
Gunderson, S. I., M. Polycarpou-Schwarz and I. W. Mattaj (1998). "U1 snRNP Inhibits Pre-mRNA Polyadenylation through a Direct Interaction between U1 70K and Poly(A) Polymerase." Molecular Cell **1**(2): 255-264.
Hafner, M., N. Renwick, M. Brown, A. Mihailović, D. Holoch, C. Lin, J. Pena, J. D. Nusbaum, P. Morozov, J. Ludwig, T. Ojo, S. Luo, G. Schroth and T. Tuschl (2011). "RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries." RNA **17**(9): 1697-1712.
Hall-Pogar, T., H. Zhang, B. Tian and C. S. Lutz (2005). "Alternative polyadenylation of cyclooxygenase-2." Nucleic Acids Research **33**(8): 2565-2579.
Hall, N. (2007). "Advanced sequencing technologies and their wider impact in microbiology." Journal of Experimental Biology **210**(9): 1518-1525.
Hart, R. P., M. A. McDevitt and J. R. Nevins (1985). "Poly(A) site cleavage in a HeLa nuclear extract is dependent on downstream sequences." Cell **43**(3): 677-683.
Hazel†, P., J. Huppert‡, S. Balasubramanian‡ and S. Neidle*† (2004). "Loop-Length-Dependent Folding of G-Quadruplexes." Journal of the American Chemical Society **126**(50): 16405-16415.

Henras, A. K., C. Plisson‐Chastang, M. F. O'Donohue, A. Chakraborty and P. E. Gleizes (2015). "An overview of pre‐ribosomal RNA processing in eukaryotes." Wiley Interdisciplinary Reviews: RNA **6**(2): 225-242.
Hershey, A. D. and M. Chase (1952). "INDEPENDENT FUNCTIONS OF VIRAL PROTEIN AND NUCLEIC ACID IN GROWTH OF BACTERIOPHAGE." The Journal of General Physiology **36**(1): 39-56.
Hirose, Y. and J. L. Manley (1998). "RNA polymerase II is an essential mRNA polyadenylation factor." Nature **395**(6697): 93-96.
Ho, K. C., K. Lehman and S. Shuman (1999). "An essential surface motif (WAQKW) of yeast RNA triphosphatase mediates formation of the mRNA capping enzyme complex with RNA guanylyltransferase." Nucleic Acids Research **27**(24): 4671-4678.
Hofacker, I. L., W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker and P. Schuster (1994). "Fast folding and comparison of RNA secondary structures." Monatshefte f&#65533;r Chemie Chemical Monthly **125**(2): 167-188.
Hollerer, I., T. Curk, B. Haase, V. Benes, C. Hauer, G. Neu-Yilik, M. Bhuvanagiri, M. W. Hentze and A. E. Kulozik (2016). "The differential expression of alternatively polyadenylated transcripts is a common stress-induced response mechanism that modulates mammalian mRNA expression in a quantitative and qualitative fashion." RNA **22**(9): 1441-1453.
Hoque, M., Z. Ji, D. Zheng, W. Luo, W. Li, B. You, J. Park, G. Yehia and B. Tian (2012). "Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing." Nature Methods **10**(2): 133-139.
Hu, Y.-F. and R. Li (2002). "JunB potentiates function of BRCA1 activation domain 1 (AD1) through a coiled-coil-mediated interaction." Genes & Development **16**(12): 1509-1517.
Huang, Y. and G. G. Carmichael (1997). "The mouse histone H2a gene contains a small element that facilitates cytoplasmic accumulation of intronless gene transcripts and of unspliced HIV-1-related mRNAs." Proceedings of the National Academy of Sciences **94**(19): 10104-10109.
Huppert, J. L. and S. Balasubramanian (2005). "Prevalence of quadruplexes in the human genome." Nucleic Acids Research **33**(9): 2908-2916.
Hutchins, L. N., S. M. Murphy, P. Singh and J. H. Graber (2008). "Position-dependent motif characterization using non-negative matrix factorization." Bioinformatics **24**(23): 2684-2690.
Hwang, H.-W., C. Y. Park, H. Goodarzi, J. J. Fak, A. Mele, M. J. Moore, Y. Saito and R. B. Darnell (2016). "PAPERCLIP Identifies MicroRNA Targets and a Role of CstF64/64tau in Promoting Non-canonical poly(A) Site Usage." Cell Reports **15**(2): 423-435.
Itoh, M., I. Haga, Q. H. Li and J. i. Fujisawa (2002). "Identification of cellular mRNA targets for RNA‐binding protein Sam68." Nucleic Acids Research **30**(24): 5452-5464.
Izaurralde, E., J. Lewis, C. McGuigan, M. Jankowska, E. Darzynkiewicz and I. W. Mattaj (1994). "A nuclear cap binding protein complex involved in pre-mRNA splicing." Cell **78**(4): 657-668.
Jain, N., C. E. Morgan, B. D. Rife, M. Salemi and B. S. Tolbert (2016). "Solution Structure of the HIV-1 Intron Splicing Silencer and Its Interactions with the UP1 Domain of Heterogeneous Nuclear Ribonucleoprotein (hnRNP) A1." Journal of Biological Chemistry **291**(5): 2331-2344.
Jan, C. H., R. C. Friedman, G. J. Ruby and D. P. Bartel (2010). "Formation, regulation and evolution of Caenorhabditis elegans 3′ UTRs." Nature **469**(7328): 97-101.
Jenal, M., R. Elkon, F. Loayza-Puch, G. van Haaften, U. Kühn, F. M. Menzies, J. Vrielink, A. J. Bos, J. Drost, K. Rooijers, D. C. Rubinsztein and R. Agami (2012). "The Poly(A)-Binding Protein Nuclear 1 Suppresses Alternative Cleavage and Polyadenylation Sites." Cell **149**(3): 538-553.
Ji, X., J. Wan, M. Vishnu, Y. Xing and S. A. Liebhaber (2013). " CP Poly(C) Binding Proteins Act as Global Regulators of Alternative Polyadenylation." Molecular and Cellular Biology **33**(13): 2560-2573.
Ji, Z., J. Lee, Z. Pan, B. Jiang and B. Tian (2009). "Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development." Proceedings of the National Academy of Sciences **106**(17): 7028-7033.

Johnson, J. M., J. Castle, P. Garrett-Engele, Z. Kan, P. M. Loerch, C. D. Armour, R. Santos, E. E. Schadt, R. Stoughton and D. D. Shoemaker (2003). "Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays." Science (New York, N.Y.) **302**(5653): 2141-2144.
Jun, H. U., C. S. Lutz, J. Wilusz and B. I. N. Tian (2005). "Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation." RNA **11**(10): 1485-1493.
Kaida, D., M. G. Berg, I. Younis, M. Kasim, L. N. Singh, L. Wan and G. Dreyfuss (2010). "U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation." Nature **468**(7324): 664-668.
Kalendar, R., C. M. Vicient, O. Peleg, K. Anamthawat-Jonsson, A. Bolshoy and A. H. Schulman (2004). "Large Retrotransposon Derivatives: Abundant, Conserved but Nonautonomous Retroelements of Barley and Related Genomes." Genetics **166**(3): 1437-1450.
Kaufmann, I., G. Martin, A. Friedlein, H. Langen and W. Keller (2004). "Human Fip1 is a subunit of CPSF that binds to U‐rich RNA elements and stimulates poly(A) polymerase." The EMBO Journal **23**(3): 616-626.
Keller, W., S. Bienroth, K. M. Lang and G. Christofori (1991). "Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal AAUAAA." The EMBO journal **10**(13): 4241-4249.
Kikin, O., L. D'Antonio and P. S. Bagga (2006). "QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences." Nucleic Acids Research **34**(suppl 2).
Kim, S., J. Yamamoto, Y. Chen, M. Aida, T. Wada, H. Handa and Y. Yamaguchi (2010). "Evidence that cleavage factor Im is a heterotetrameric protein complex controlling alternative polyadenylation." Genes to Cells **15**(9): 1003-1013.
Kleiman, F. E. and J. L. Manley (2001). "The BARD1-CstF-50 Interaction Links mRNA 3′ End Formation to DNA Damage and Tumor Suppression." Cell **104**(5): 743-753.
Köhler, A. and E. Hurt (2007). "Exporting RNA from the nucleus to the cytoplasm." Nature Reviews Molecular Cell Biology **8**(10): 761-773.
Kolev, N. G., T. A. Yario, E. Benson and J. A. Steitz (2008). "Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3′‐end maturation." EMBO reports **9**(10): 1013-1018.
Konig, J., K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D. J. Turner, N. M. Luscombe and J. Ule (2011). "iCLIP - Transcriptome-wide Mapping of Protein-RNA Interactions with Individual Nucleotide Resolution." Journal of Visualized Experiments : JoVE(50).
Kubo, T., T. Wada, Y. Yamaguchi, A. Shimizu and H. Handa (2006). "Knock-down of 25 kDa subunit of cleavage factor Im in Hela cells alters alternative polyadenylation within 3′-UTRs." Nucleic Acids Research **34**(21): 6264-6271.
Kühn, U., M. Gündel, A. Knoth, Y. Kerwitz, S. Rüdel and E. Wahle (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." Journal of Biological Chemistry **284**(34): 22803-22814.
Kwon, C. S., H. Yi, K. Eichelbaum, S. Föhr, B. Fischer, K. You, A. Castello, J. Krijgsveld, M. W. Hentze and N. V. Kim (2013). "The RNA-binding protein repertoire of embryonic stem cells." Nature Structural & Molecular Biology **20**(9): 1122-1130.
Kyburz, A., A. Friedlein, H. Langen and W. Keller (2006). "Direct Interactions between Subunits of CPSF and the U2 snRNP Contribute to the Coupling of Pre-mRNA 3′ End Processing and Splicing." Molecular Cell **23**(2): 195-205.
Lackford, B., C. Yao, G. M. Charles, L. Weng, X. Zheng, E. A. Choi, X. Xie, J. Wan, Y. Xing, J. M. Freudenberg, P. Yang, R. Jothi, G. Hu and Y. Shi (2014). "Fip1 regulates mRNA alternative polyadenylation to promote stem cell self‐renewal." The EMBO Journal **33**(8): 878-889.
Laishram, R. S. (2014). "Poly(A) polymerase (PAP) diversity in gene expression – Star‐PAP vs canonical PAP." FEBS Letters **588**(14): 2185-2197.

Laitem, C., J. Zaborowska, N. F. Isa, J. Kufs, M. Dienstbier and S. Murphy (2015). "CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II–transcribed genes." Nature Structural & Molecular Biology **22**(5): 396-403.
Lazer, G., L. Pe'er, V. Schapira, S. Richard and S. Katzav (2007). "The association of Sam68 with Vav1 contributes to tumorigenesis." Cellular Signalling **19**(12): 2479-2486.
Lenasi, T., M. B. Peterlin and M. Barboric (2011). "Cap-binding Protein Complex Links Pre-mRNA Capping to Transcription Elongation and Alternative Splicing through Positive Transcription Elongation Factor b (P-TEFb)." Journal of Biological Chemistry **286**(26): 22758-22768.
Lengyel, P., J. F. Speyer and S. Ochoa (1961). "SYNTHETIC POLYNUCLEOTIDES AND THE AMINO ACID CODE." Proceedings of the National Academy of Sciences **47**(12): 1936-1942.
Levitt, N., D. Briggs, A. Gil and N. J. Proudfoot (1989). "Definition of an efficient synthetic poly(A) site." Genes & Development **3**(7): 1019-1025.
Lewis, J. D. and E. Izaurralde (1997). "The role of the cap structure in RNA processing and nuclear export." European journal of biochemistry / FEBS **247**(2): 461-469.
Li, H.-D., R. Menon, G. S. Omenn and Y. Guan (2014). "The emerging era of genomic data integration for analyzing splice isoform function." Trends in Genetics **30**(8): 340-347.
Li, M., Y. Zhang, Z. Liu, U. Bharadwaj, H. Wang, X. Wang, S. Zhang, J. P. Liuzzi, S.-M. Chang, R. J. Cousins, W. E. Fisher, C. F. Brunicardi, C. D. Logsdon, C. Chen and Q. Yao (2007). "Aberrant expression of zinc transporter ZIP4 (SLC39A4) significantly contributes to human pancreatic cancer pathogenesis and progression." Proceedings of the National Academy of Sciences **104**(47): 18636-18641.
Li, W., J. Y. Park, D. Zheng, M. Hoque, G. Yehia and B. Tian (2016). "Alternative cleavage and polyadenylation in spermatogenesis connects chromatin regulation with post-transcriptional control." BMC Biol **14**: 6.
Li, W., B. You, M. Hoque, D. Zheng, W. Luo, Z. Ji, J. Park, S. I. Gunderson, A. Kalsotra, J. L. Manley and B. Tian (2015). "Systematic Profiling of Poly(A)+ Transcripts Modulated by Core 3' End Processing and Splicing Factors Reveals Regulatory Rules of Alternative Cleavage and Polyadenylation." PLOS Genetics **11**(4).
Li, X.-Q. and D. Du (2013). "RNA Polyadenylation Sites on the Genomes of Microorganisms, Animals, and Plants." PLoS ONE **8**(11).
Lian, Z., A. Karpikov, J. Lian, M. C. Mahajan, S. Hartman, M. Gerstein, M. Snyder and S. M. Weissman (2008). "A genomic analysis of RNA polymerase II modification and chromatin architecture related to 3′ end RNA polyadenylation." Genome Research **18**(8): 1224-1237.
Lianoglou, S., V. Garg, J. L. Yang, C. S. Leslie and C. Mayr (2013). "Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression." Genes Dev **27**(21): 2380-2396.
Licatalosi, D. D., A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, J. C. Darnell and R. B. Darnell (2008). "HITS-CLIP yields genome-wide insights into brain alternative RNA processing." Nature **456**(7221): 464-469.
Lim, K., L. Ferraris, M. E. Filloux, B. J. Raphael and W. G. Fairbrother (2011). "Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes." Proceedings of the National Academy of Sciences **108**(27): 11093-11098.
Lim, L. and E. S. Canellakis (1970). "Adenine-rich Polymer associated with Rabbit Reticulocyte Messenger RNA." Nature **227**(5259): 710-712.
Lin, Q., S. J. Taylor and D. Shalloway (1997). "Specificity and Determinants of Sam68 RNA Binding: IMPLICATIONS FOR THE BIOLOGICAL FUNCTION OF K HOMOLOGY DOMAINS." Journal of Biological Chemistry **272**(43): 27274-27280.
Lin, Y., Z. Li, F. Ozsolak, S. Kim, G. Arango-Argoty, T. T. Liu, S. A. Tenenbaum, T. Bailey, P. A. Monaghan, P. M. Milos and B. John (2012). "An in-depth map of polyadenylation sites in cancer." Nucleic Acids Research **40**(17): 8460-8471.

Liu, Y., W. Hu, Y. Murakawa, J. Yin, G. Wang, M. Landthaler and J. Yan (2013). "Cold-induced RNA-binding proteins regulate circadian gene expression by controlling alternative polyadenylation." Scientific Reports **3**: 2054.
LÓPez-Lastra, M., A. Rivas and M. RrÍA (2005). "Protein synthesis in eukaryotes: The growing biological relevance of cap-independent translation initiation." Biological Research **38**(2-3): 121-146.
Luco, R. F., Q. Pan, K. Tominaga, B. J. Blencowe, O. M. Pereira-Smith and T. Misteli (2010). "Regulation of Alternative Splicing by Histone Modifications." Science **327**(5968): 996-1000.
Lukong, K. E. and S. Richard (2003). "Sam68, the KH domain-containing superSTAR." Biochimica et Biophysica Acta (BBA) - Reviews on Cancer **1653**(2): 73-86.
Madhani, H. D. and C. Guthrie (1992). "A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome." Cell **71**(5): 803-817.
Makarova, O. V., E. M. Makarov and R. Lührmann (2001). "The 65 and 110 kDa SR‐related proteins of the U4/U6 • U5 tri‐snRNP are essential for the assembly of mature spliceosomes." The EMBO Journal **20**(10): 2553-2563.
Mandel, C. R., Y. Bai and L. Tong (2008). "Protein factors in pre-mRNA 3′-end processing." Cellular and Molecular Life Sciences **65**(7-8): 1099-1122.
Mandel, C. R., S. Kaneko, H. Zhang, D. Gebauer, V. Vethantham, J. L. Manley and L. Tong (2006). "Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease." Nature **444**(7121): 953-956.
Mangone, M., A. Manoharan, D. Thierry-Mieg, J. Thierry-Mieg, T. Han, S. D. Mackowiak, E. Mis, C. Zegar, M. R. Gutwein, V. Khivansara, O. Attie, K. Chen, K. Salehi-Ashtiani, M. Vidal, T. T. Harkins, P. Bouffard, Y. Suzuki, S. Sugano, Y. Kohara, N. Rajewsky, F. Piano, K. C. Gunsalus and J. K. Kim (2010). "The Landscape of C. elegans 3′ UTRs." Science **329**(5990): 432-435.
Maniatis, T. and R. Reed (2002). "An extensive network of coupling among gene expression machines." Nature **416**(6880): 499-506.
Markham, N. R. and M. Zuker (2008). "UNAFold." Methods in molecular biology (Clifton, N.J.) **453**: 3-31.
Martin, G., A. R. Gruber, W. Keller and M. Zavolan (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**(6): 753-763.
Marzluff, W. F., P. Gongidi, K. R. Woods, J. Jin and L. J. Maltais (2002). "The Human and Mouse Replication-Dependent Histone Genes." Genomics **80**(5): 487-498.
Marzluff, W. F., E. J. Wagner and R. J. Duronio (2008). "Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail." Nature Reviews Genetics **9**(11): 843-854.
Masamha, C. P., Z. Xia, J. Yang, T. R. Albrecht, M. Li, A.-B. Shyu, W. Li and E. J. Wagner (2014). "CFIm25 links alternative polyadenylation to glioblastoma tumour suppression." Nature **510**(7505): 412-416.
Masuda, A., J.-i. Takeda, T. Okuno, T. Okamoto, B. Ohkawara, M. Ito, S. Ishigaki, G. Sobue and K. Ohno (2015). "Position-specific binding of FUS to nascent RNA regulates mRNA length." Genes & development **29**(10): 1045-1057.
Mathews, D. H., W. N. Moss and D. H. Turner (2010). "Folding and Finding RNA Secondary Structure." Cold Spring Harbor Perspectives in Biology **2**(12).
Mayr, C. and D. P. Bartel (2009). "Widespread Shortening of 3′ UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells." Cell **138**(4): 673-684.
McCracken, S., N. Fong, E. Rosonina, K. Yankulov, G. Brothers, D. Siderovski, A. Hessel, S. Foster, A. Program, S. Shuman and D. L. Bentley (1997). "5′-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II." Genes & Development **11**(24): 3306-3318.

McCubrey, J. A., L. S. Steelman, W. H. Chappell, S. L. Abrams, E. Wong, F. Chang, B. Lehmann, D. M. Terrian, M. Milella, A. Tafuri, F. Stivala, M. Libra, J. Basecke, C. Evangelisti, A. M. Martelli and R. A. Franklin (2007). "Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance." Biochimica et Biophysica Acta (BBA) - Molecular Cell Research **1773**(8): 1263-1284.
McLaren, M., K. Asai and A. Cochrane (2004). "A novel function for Sam68: Enhancement of HIV-1 RNA 3′ end processing." RNA **10**(7): 1119-1129.
McLauchlan, J., D. Gaffney, L. J. Whitton and B. J. Clements (1985). "The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3′ termini." Nucleic Acids Research **13**(4): 1347-1368.
Meyer, K. D. and S. R. Jaffrey (2014). "The dynamic epitranscriptome: N6-methyladenosine and gene expression control." Nature Reviews Molecular Cell Biology **15**(5): 313-326.
Millevoi, S., F. Geraghty, B. Idowu, J. L. Y. Tam, M. Antoniou and S. Vagner (2002). "A novel function for the U2AF 65 splicing factor in promoting pre‐mRNA 3′ ‐end processing." EMBO reports **3**(9): 869-874.
Millevoi, S., C. Loulergue, S. Dettwiler, S. Karaa, W. Keller, M. Antoniou and S. Vagner (2006). "An interaction between U2AF 65 and CF Im links the splicing and 3′ end processing machineries." The EMBO Journal **25**(20): 4854-4864.
Millevoi, S. and S. Vagner (2010). "Molecular mechanisms of eukaryotic pre-mRNA 3′ end processing regulation." Nucleic Acids Research **38**(9): 2757-2774.
Miura, P., S. Shenker, C. Andreu-Agullo, J. O. Westholm and E. C. Lai (2013). "Widespread and extensive lengthening of 3' UTRs in the mammalian brain." Genome research **23**(5): 812-825.
Modem, S., K. R. Badri, T. C. Holland and T. R. Reddy (2005). "Sam68 is absolutely required for Rev function and HIV-1 production." Nucleic Acids Research **33**(4): 1420-1420.
Moore, M. J. and P. A. Sharp (1993). "Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing." Nature **365**(6444): 364-368.
Moreira, A., Y. Takagaki, S. Brackenridge, M. Wollerton, J. L. Manley and N. J. Proudfoot (1998). "The upstream sequence element of the C2 complement poly(A) signal activates mRNA 3' end formation by two distinct mechanisms." Genes & Development **12**(16): 2522-2534.
Moreno-Morcillo, M., L. Minvielle-Sébastia, C. Mackereth and S. Fribourg (2011). "Hexameric architecture of CstF supported by CstF-50 homodimerization domain structure." RNA **17**(3): 412-418.
Morris, A. R., A. Bos, B. Diosdado, K. Rooijers, R. Elkon, A. S. Bolijn, B. Carvalho, G. A. Meijer and R. Agami (2012). "Alternative cleavage and polyadenylation during colorectal cancer development." Clinical cancer research : an official journal of the American Association for Cancer Research **18**(19): 5256-5266.
Muhlrad, D., C. J. Decker and R. Parker (1994). "Deadenylation of the unstable mRNA encoded by the yeast MFA2 gene leads to decapping followed by 5'-->3' digestion of the transcript." Genes & Development **8**(7): 855-866.
Murthy, K. G. and J. L. Manley (1992). "Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus." The Journal of biological chemistry **267**(21): 14804-14811.
Murthy, K. G. and J. L. Manley (1995). "The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation." Genes & Development **9**(21): 2672-2683.
Neve, J., K. Burger, W. Li, M. Hoque, R. Patel, B. Tian, M. Gullerova and A. Furger (2016). "Subcellular RNA profiling links splicing and nuclear DICER1 to alternative cleavage and polyadenylation." Genome Research **26**(1): 24-35.
Newman, A. J., S. Teigelkamp and J. D. Beggs (1995). "snRNA interactions at 5' and 3' splice sites monitored by photoactivated crosslinking in yeast spliceosomes." RNA (New York, N.Y.) **1**(9): 968-980.

Ni, T. K. and C. Kuperwasser (2016). "Premature polyadenylation of MAGI3 produces a dominantly-acting oncogene in human breast cancer." Elife **5**.
Nicolaides, N. C., I. Correa, C. Casadevall, S. Travali, K. J. Soprano and B. Calabretta (1992). "The Jun family members, c-Jun and JunD, transactivate the human c-myb promoter via an Ap1-like element." The Journal of biological chemistry **267**(27): 19665-19672.
Nilsson, M., R. Toftgård and S. Bohm (1995). "Activated Ha-Ras but not TPA induces transcription through binding sites for activating transcription factor 3/Jun and a novel nuclear factor." The Journal of biological chemistry **270**(20): 12210-12218.
Nirenberg, M. (2004). "Historical review: Deciphering the genetic code – a personal account." Trends in Biochemical Sciences **29**(1): 46-54.
Niwa, M. and S. M. Berget (1991). "Mutation of the AAUAAA polyadenylation signal depresses in vitro splicing of proximal but not distal introns." Genes & development **5**(11): 2086-2095.
Niwa, M., C. C. MacDonald and S. M. Berget (1992). "Are vertebrate exons scanned during splice-site selection?" Nature **360**(6401): 277-280.
Niwa, M., S. D. Rose and S. M. Berget (1990). "In vitro polyadenylation is stimulated by the presence of an upstream intron." Genes & development **4**(9): 1552-1559.
Nunes, N., W. Li, B. Tian and A. Furger (2010). "A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence." The EMBO journal **29**(9): 1523-1536.
Oberg, D., J. Fay, H. Lambkin and S. Schwartz (2005). "A downstream polyadenylation element in human papillomavirus type 16 L2 encodes multiple GGG motifs and interacts with hnRNP H." Journal of virology **79**(14): 9254-9269.
Ohnacker, M., S. M. L. Barabino, P. J. Preker and W. Keller (2000). "The WD‐repeat protein Pfs2p bridges two essential factors within the yeast pre‐mRNA 3′‐end‐processing complex." The EMBO Journal **19**(1): 37-47.
Ohno, M., H. Sakamoto and Y. Shimura (1987). "Preferential excision of the 5' proximal intron from mRNA precursors with two introns as mediated by the cap structure." Proceedings of the National Academy of Sciences **84**(15): 5187-5191.
Orkin, S. H., T. C. Cheng, S. E. Antonarakis and H. H. Kazazian (1985). "Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene." The EMBO journal **4**(2): 453-456.
Orphanides, G., T. Lagrange and D. Reinberg (1996). "The general transcription factors of RNA polymerase II." Genes & Development **10**(21): 2657-2683.
Orphanides, G. and D. Reinberg (2002). "A Unified Theory of Gene Expression." Cell **108**(4): 439-451.
Ozsolak, F., P. Kapranov, S. Foissac, S. Kim, E. Fishilevich, P. A. Monaghan, B. John and P. M. Milos (2010). "Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation." Cell **143**(6): 1018-1029.
Ozsolak, F. and P. M. Milos (2010). "RNA sequencing: advances, challenges and opportunities." Nature Reviews Genetics **12**(2): 87-98.
Pan, Q., O. Shai, L. J. Lee, B. J. Frey and B. J. Blencowe (2008). "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing." Nature Genetics **40**(12): 1413-1415.
Pandey, N. B. and W. F. Marzluff (1987). "The stem-loop structure at the 3' end of histone mRNA is necessary and sufficient for regulation of histone mRNA stability." Molecular and Cellular Biology **7**(12): 4557-4559.
Parker, R. and U. Sheth (2007). "P Bodies and the Control of mRNA Translation and Degradation." Molecular Cell **25**(5): 635-646.
Paronetto, M., T. Achsel, A. Massiello, C. E. Chalfant and C. Sette (2007). "The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x." The Journal of Cell Biology **176**(7): 929-939.

Paule, M. R. and R. J. White (2000). "SURVEY AND SUMMARY Transcription by RNA polymerases I and III." Nucleic Acids Research **28**(6): 1283-1298.
Petraccone, L., E. Erra, I. Duro, V. Esposito, A. Randazzo, L. Mayol, C. Mattia, G. Barone and C. Giancola (2005). "Relative stability of quadruplexes containing different number of G-tetrads." Nucleosides, nucleotides & nucleic acids **24**(5-7): 757-760.
Proudfoot, N. J. (2011). "Ending the message: poly(A) signals then and now." Genes & Development **25**(17): 1770-1782.
Proudfoot, N. J., A. Furger and M. J. Dye (2002). "Integrating mRNA Processing with Transcription." Cell **108**(4): 501-512.
Qiu, J., R. Nayak and D. J. Pintel (2004). "Alternative polyadenylation of adeno-associated virus type 5 RNA within an internal intron is governed by both a downstream element within the intron 3' splice acceptor and an element upstream of the P41 initiation site." Journal of virology **78**(1): 83-93.
Query, C. C., M. J. Moore and P. A. Sharp (1994). "Branch nucleophile selection in pre-mRNA splicing: evidence for the bulged duplex model." Genes & development **8**(5): 587-597.
Raabe, T., F. J. Bollum and J. L. Manley (1991). "Primary structure and expression of bovine poly(A) polymerase." Nature **353**(6341): 229-234.
Reddy, T. R., W. Xu, J. K. L. Mau, C. D. Goodwin, M. Suhasini, H. Tang, K. Frimpong, D. W. Rose and F. Wong-Staal (1999). "Inhibition of HIV replication by dominant negative mutants of Sam68, a functional homolog of HIV-1 Rev." Nature Medicine **5**(6): 635-642.
Redon, C., D. Pilch, E. Rogakou, O. Sedelnikova, K. Newrock and W. Bonner (2002). "Histone H2A variants H2AX and H2AZ." Current Opinion in Genetics & Development **12**(2): 162-169.
Reese, J. C. (2003). "Basal transcription factors." Current Opinion in Genetics & Development **13**(2): 114-118.
Risitano, A. and K. R. Fox (2004). "Influence of loop size on the stability of intramolecular DNA quadruplexes." Nucleic Acids Research **32**(8): 2598-2606.
Rüegsegger, U., K. Beyer and W. Keller (1996). "Purification and Characterization of Human Cleavage Factor I Involved in the 3′ End Processing of Messenger RNA Precursors." Journal of Biological Chemistry **271**(11): 6107-6113.
Rüegsegger, U., D. Blank and W. Keller (1998). "Human Pre-mRNA Cleavage Factor Im Is Related to Spliceosomal SR Proteins and Can Be Reconstituted In Vitro from Recombinant Subunits." Molecular Cell **1**(2): 243-253.
Salisbury, J., K. W. Hutchison and J. H. Graber (2006). "A multispecies comparison of the metazoan 3'-processing downstream elements and the CstF-64 RNA recognition motif." BMC Genomics **7**(1): 1-13.
Sandberg, R., J. R. Neilson, A. Sarma, P. A. Sharp and C. B. Burge (2008). "Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites." Science **320**(5883): 1643-1647.
Schoenberg, D. R. and L. E. Maquat (2009). "Re-capping the message." Trends in Biochemical Sciences **34**(9): 435-442.
Schönemann, L., U. Kühn, G. Martin, P. Schäfer, A. R. Gruber, W. Keller, M. Zavolan and E. Wahle (2014). "Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33." Genes & Development **28**(21): 2381-2393.
Shatkin, A. J. (1976). "Capping of eucaryotic mRNAs." Cell **9**(4 PT 2): 645-653.
Shatkin, A. J. and J. L. Manley (2000). "The ends of the affair: Capping and polyadenylation." Nature Structural & Molecular Biology **7**(10): 838-842.
Sheets, M. D., S. C. Ogg and M. P. Wickens (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 research **18**(19): 5799-5805.
Shepard, P. J., E.-A. Choi, J. Lu, L. A. Flanagan, K. J. Hertel and Y. Shi (2011). "Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq." RNA **17**(4): 761-772.

Shi, Y., D. Giammartino, D. Taylor, A. Sarkeshik, W. J. Rice, J. R. Yates, J. Frank and J. L. Manley (2009). "Molecular Architecture of the Human Pre-mRNA 3′ Processing Complex." Molecular Cell **33**(3): 365-376.
Shilatifard, A., R. C. Conaway and J. Conaway (2003). "THE RNA POLYMERASE II ELONGATION COMPLEX." Annual review of biochemistry **72**(1): 693-715.
Shuman, S. and B. Schwer (1995). "RNA capping enzyme and DNA ligase: a superfamily of covalent nucleotidyl transferases." Molecular Microbiology **17**(3): 405-420.
Singh, P., T. L. Alley, S. M. Wright, S. Kamdar, W. Schott, R. Y. Wilpan, K. D. Mills and J. H. Graber (2009). "Global Changes in Processing of mRNA 3′ Untranslated Regions Characterize Clinically Distinct Cancer Subtypes." Cancer Research **69**(24): 9422-9430.
Sittler, A., H. Gallinaro and M. Jacob (1995). "The secondary structure of the adenovirus-2 L4 polyadenylation domain: evidence for a hairpin structure exposing the AAUAAA signal in its loop." J Mol Biol **248**(3): 525-540.
Smibert, P., P. Miura, J. O. Westholm, S. Shenker, G. May, M. O. Duff, D. Zhang, B. D. Eads, J. Carlson, J. B. Brown, R. C. Eisman, J. Andrews, T. Kaufman, P. Cherbas, S. E. Celniker, B. R. Graveley and E. C. Lai (2012). "Global Patterns of Tissue-Specific Alternative Polyadenylation in Drosophila." Cell Reports **1**(3): 277-289.
Sonenberg, N. (1994). "mRNA translation: influence of the 5′ and 3′ untranslated regions." Current Opinion in Genetics & Development **4**(2): 310-315.
Spies, N., C. B. Burge and D. P. Bartel (2013). "3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts." Genome Res **23**(12): 2078-2090.
Stirnimann, C. U., E. Petsalaki, R. B. Russell and C. W. Müller (2010). "WD40 proteins propel cellular networks." Trends in Biochemical Sciences **35**(10): 565-574.
Sun, J. S. and J. L. Manley (1995). "A novel U2-U6 snRNA structure is necessary for mammalian mRNA splicing." Genes & Development **9**(7): 843-854.
Takagaki, Y. and J. L. Manley (1994). "A polyadenylation factor subunit is the human homologue of theDrosophila suppressor of forked protein." Nature **372**(6505): 471-474.
Takagaki, Y. and J. L. Manley (1997). "RNA recognition by the human polyadenylation factor CstF." Molecular and Cellular Biology **17**(7): 3907-3914.
Takagaki, Y. and J. L. Manley (1998). "Levels of Polyadenylation Factor CstF-64 Control IgM Heavy Chain mRNA Accumulation and Other Events Associated with B Cell Differentiation." Molecular Cell **2**(6): 761-771.
Takagaki, Y. and J. L. Manley (2000). "Complex Protein Interactions within the Human Polyadenylation Machinery Identify a Novel Component." Molecular and Cellular Biology **20**(5): 1515-1525.
Takagaki, Y., J. L. Manley, C. C. MacDonald, J. Wilusz and T. Shenk (1990). "A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs." Genes & development **4**(12A): 2112-2120.
Takagaki, Y., L. C. Ryner and J. L. Manley (1989). "Four factors are required for 3'-end cleavage of pre-mRNAs." Genes & development **3**(11): 1711-1724.
Takagaki, Y., R. L. Seipelt, M. L. Peterson and J. L. Manley (1996). "The Polyadenylation Factor CstF-64 Regulates Alternative Processing of IgM Heavy Chain Pre-mRNA during B Cell Differentiation." Cell **87**(5): 941-952.
Taylor, S. J. and D. Shalloway (1994). "An RNA-binding protein associated with Src through its SH2 and SH3 domains in mitosis." Nature **368**(6474): 867-871.
Thomas, P. E., X. Wu, M. Liu, B. Gaffney, G. Ji, Q. Q. Li and A. G. Hunt (2012). "Genome-Wide Control of Polyadenylation Site Choice by CPSF30 in Arabidopsis." The Plant Cell **24**(11): 4376-4388.
Tian, B. and J. H. Graber (2012). "Signals for pre‐mRNA cleavage and polyadenylation." Wiley Interdisciplinary Reviews: RNA **3**(3): 385-396.

Tian, B., J. Hu, H. Zhang and C. S. Lutz (2005). "A large-scale analysis of mRNA polyadenylation of human and mouse genes." Nucleic Acids Research **33**(1): 201-212.
Tian, B. and J. L. Manley (2013). "Alternative cleavage and polyadenylation: the long and short of it." Trends in Biochemical Sciences **38**(6): 312-320.
Todd, A. K., M. Johnston and S. Neidle (2005). "Highly prevalent putative quadruplex sequence motifs in human DNA." Nucleic Acids Research **33**(9): 2901-2907.
Tripathi, V., J. D. Ellis, Z. Shen, D. Y. Song, Q. Pan, A. T. Watt, S. M. Freier, F. C. Bennett, A. Sharma, P. A. Bubulya, B. J. Blencowe, S. G. Prasanth and K. V. Prasanth (2010). "The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation." Molecular Cell **39**(6): 925-938.
Ule, J., K. Jensen, A. Mele and R. B. Darnell (2005). "CLIP: A method for identifying protein–RNA interaction sites in living cells." Methods **37**(4): 376-386.
Ule, J., K. B. Jensen, M. Ruggiu, A. Mele, A. Ule and R. B. Darnell (2003). "CLIP Identifies Nova-Regulated RNA Networks in the Brain." Science **302**(5648): 1212-1215.
Ulitsky, I., A. Shkumatava, C. H. Jan, A. O. Subtelny, D. Koppstein, G. W. Bell, H. Sive and D. P. Bartel (2012). "Extensive alternative polyadenylation during zebrafish development." Genome Research **22**(10): 2054-2066.
Vagner, S., C. Vagner and I. W. Mattaj (2000). "The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF 65 to couple 3'-end processing and splicing." Genes & development **14**(4): 403-413.
Valsamakis, A., N. Schek and J. C. Alwine (1992). "Elements upstream of the AAUAAA within the human immunodeficiency virus polyadenylation signal are required for efficient polyadenylation in vitro." Molecular and Cellular Biology **12**(9): 3699-3705.
Valsamakis, A., S. Zeichner, S. Carswell and J. C. Alwine (1991). "The human immunodeficiency virus type 1 polyadenylylation signal: a 3' long terminal repeat element upstream of the AAUAAA necessary for efficient polyadenylylation." Proceedings of the National Academy of Sciences **88**(6): 2108-2112.
Velten, L., S. Anders, A. Pekowska, A. I. Jarvelin, W. Huber, V. Pelechano and L. M. Steinmetz (2015). "Single-cell polyadenylation site mapping reveals 3' isoform choice variability." Mol Syst Biol **11**(6): 812.
Venkataraman, K., K. M. Brown and G. M. Gilmartin (2005). "Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition." Genes & Development **19**(11): 1315-1327.
Vilborg, A., M. C. Passarelli, T. A. Yario, K. T. Tycowski and J. A. Steitz (2015). "Widespread Inducible Transcription Downstream of Human Genes." Mol Cell **59**(3): 449-461.
Wahle, E. (1991). "Purification and characterization of a mammalian polyadenylate polymerase involved in the 3' end processing of messenger RNA precursors." The Journal of biological chemistry **266**(5): 3131-3139.
Wahle, E., G. Martin, E. Schiltz and W. Keller (1991). "Isolation and expression of cDNA clones encoding mammalian poly(A) polymerase." The EMBO journal **10**(13): 4251-4257.
Wang, E. T., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth and C. B. Burge (2008). "Alternative isoform regulation in human tissue transcriptomes." Nature **456**(7221): 470-476.
Wang, L., R. D. Dowell and R. Yi (2013). "Genome-wide maps of polyadenylation reveal dynamic mRNA 3′ -end formation in mammalian cell lineages." RNA **19**(3): 413-425.
Wang, Y., J. Liu, B. O. Huang, Y.-M. Xu, J. Li, L.-F. Huang, J. Lin, J. Zhang, Q.-H. Min, W.-M. Yang and X.-Z. Wang (2014). "Mechanism of alternative splicing and its regulation." Biomedical reports **3**(2): 152-158.
Warf, B. M. and A. J. Berglund (2010). "Role of RNA structure in regulating pre-mRNA splicing." Trends in Biochemical Sciences **35**(3): 169-178.

Watson, J. D. and F. H. C. Crick (1953). "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid." Nature **171**(4356): 737-738.
Weitzman, J. B., L. Fiette, K. Matsuo and M. Yaniv (2000). "JunD Protects Cells from p53-Dependent Senescence and Apoptosis." Molecular Cell **6**(5): 1109-1119.
Wells, S. E., P. E. Hillner, R. D. Vale and A. B. Sachs (1998). "Circularization of mRNA by Eukaryotic Translation Initiation Factors." Molecular Cell **2**(1): 135-140.
West, S. and N. J. Proudfoot (2008). "Human Pcf11 enhances degradation of RNA polymerase II-associated nascent RNA and transcriptional termination." Nucleic Acids Research **36**(3): 905-914.
Whitelaw, E. and N. Proudfoot (1986). "Alpha-thalassaemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to 3' end processing in the human alpha 2 globin gene." The EMBO journal **5**(11): 2915-2922.
Whyteside, A. R., A. J. Turner and D. W. Lambert (2014). "Endothelin-converting enzyme-1 (ECE-1) is post-transcriptionally regulated by alternative polyadenylation." PLoS One **9**(1): e83260.
Wickens, M. and P. Stephenson (1984). "Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3' end formation." Science (New York, N.Y.) **226**(4678): 1045-1051.
Wiestner, A., M. Tehrani, M. Chiorazzi, G. Wright, F. Gibellini, K. Nakayama, H. Liu, A. Rosenwald, K. H. Muller-Hermelink, G. Ott, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. Vose, J. O. Armitage, R. D. Gascoyne, J. M. Connors, E. Campo, E. Montserrat, F. Bosch, E. B. Smeland, S. Kvaloy, H. Holte, J. Delabie, R. I. Fisher, T. M. Grogan, T. P. Miller, W. H. Wilson, E. S. Jaffe and L. M. Staudt (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**(11): 4599-4606.
Winters, M. A. and M. Edmonds (1973). "A poly(A) polymerase from calf thymus. Characterization of the reaction product and the primer requirement." The Journal of biological chemistry **248**(13): 4763-4768.
Winters, M. A. and M. Edmonds (1973). "A poly(A) polymerase from calf thymus. Purification and properities of the enzyme." The Journal of biological chemistry **248**(13): 4756-4762.
Woychik, N. (2000). "TA-TAAA!! Transcription unveiled Mechanisms of Gene Expression. Structure, Function and Evolution of the Basal Transcriptional Machinery by Robert O.J. Weinzierl." Trends in Biochemical Sciences **25**(8): 404.
Woychik, N. A. and M. Hampsey (2002). "The RNA Polymerase II Machinery Structure Illuminates Function." Cell **108**(4): 453-463.
Xi, H., Y. Yu, Y. Fu, J. Foley, A. Halees and Z. Weng (2007). "Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1." Genome Research **17**(6): 798-806.
Xia, Z., L. A. Donehower, T. A. Cooper, J. R. Neilson, D. A. Wheeler, E. J. Wagner and W. Li (2014). "Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types." Nature Communications **5**: 5274.
Xiang, K., L. Tong and J. L. Manley (2014). "Delineating the structural blueprint of the pre-mRNA 3'-end processing machinery." Molecular and cellular biology **34**(11): 1894-1910.
Xing, H., C. N. Mayhew, K. E. Cullen, O.-K. Park-Sarge and K. D. Sarge (2004). "HSF1 Modulation of Hsp70 mRNA Polyadenylation via Interaction with Symplekin." Journal of Biological Chemistry **279**(11): 10551-10555.
Yang, C., E. Bolotin, T. Jiang, F. M. Sladek and E. Martinez (2007). "Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters." Gene **389**(1): 52-65.

Yang, F., P. Hsu, S. D. Lee, W. Yang, D. Hoskinson, W. Xu, C. Moore and G. Varani (2016). "The C terminus of Pcf11 forms a novel zinc-finger structure that plays an essential role in mRNA 3′ -end processing." RNA **23**(1): 98-107.
Yang, Q., M. Coseno, G. M. Gilmartin and S. Doublié (2011). "Crystal Structure of a Human Cleavage Factor CFIm25/CFIm68/RNA Complex Provides an Insight into Poly(A) Site Recognition and RNA Looping." Structure **19**(3): 368-377.
Yang, Q., G. M. Gilmartin and S. Doublié (2011). "The structure of human Cleavage Factor Imhints at functions beyond UGUA-specific RNA binding." RNA Biology **8**(5): 748-753.
Yao, C., J. Biesinger, J. Wan, L. Weng, Y. Xing, X. Xie and Y. Shi (2012). "Transcriptome-wide analyses of CstF64–RNA interactions in global regulation of mRNA alternative polyadenylation." Proceedings of the National Academy of Sciences **109**(46): 18773-18778.
Yasuda, M., J. Shabbeer, M. Osawa and R. J. Desnick (2003). "Fabry disease: novel alpha-galactosidase A 3'-terminal mutations result in multiple transcripts due to aberrant 3'-end formation." American journal of human genetics **73**(1): 162-173.
Zaborowska, J., S. Egloff and S. Murphy (2016). "The pol II CTD: new twists in the tail." Nature Structural & Molecular Biology **23**(9): 771-777.
Zarudnaya, M. I., I. M. Kolomiets, A. L. Potyahaylo and D. M. Hovorun (2003). "Downstream elements of mammalian pre ‐ mRNA polyadenylation signals: primary, secondary and higher ‐ order structures." Nucleic Acids Research **31**(5): 1375-1386.
Zhang, H., J. Lee and B. Tian (2005). "Biased alternative polyadenylation in human tissues." Genome Biology **6**(12).
Zhao, J., L. Hyman and C. Moore (1999). "Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis." Microbiology and molecular biology reviews : MMBR **63**(2): 405-445.

W1. http://www.nature.com/scitable/topicpage/gene-expression-14121669