Analysis of human non-canonical 3'end formation signals

by

NUNO MIGUEL DA ROCHA OLIVEIRA NUNES

LABORATORIES OF GENES AND DEVELOPMENT

DEPARTMENT OF BIOCHEMISTRY
UNIVERSITY OF OXFORD

and

CORPUS CHRISTI COLLEGE, OXFORD

DISSERTATION FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

JANUARY 2012

Abstract

Cleavage and polyadenylation are essential pre-mRNA processing reactions maturing the 3'end of almost all protein encoding eukaryotic mRNAs. Analysis of the sequences required for cleavage and polyadenylation in the human melanocortin 4 receptor (MC4R) and the human transcription factors JUNB and JUND pre-mRNAs revealed that, at least for some mammalian genes, 3'end processing of the primary transcript is independent of previously described auxiliary sequence elements located upstream or downstream of the core poly(A) sequences. The analysis of the MC4R poly(A) site, contrary to the current understanding of mammalian poly(A) sites, showed that mutations of the AUUAAA hexamer sequence had no effect on 3'end processing levels while mutations in the short DSE severely reduced cleavage efficiency. The MC4R poly(A) site uses a potent DSE and to direct maximal cleavage efficiency requires only a short upstream adenosine rich sequence. Furthermore, analysis of the endogenous A-rich human JUNB poly(A) signal validated upstream A-rich core sequences as genuine 3'end formation directing sequences in human non-canonical 3'end formation signals. The results show that a minimal human poly(A) site, similar to yeast and plants, can be defined by an adenosine rich sequence adjacent to a U/GU-rich sequence element and a cleavage site. These findings further imply that some human non-canonical poly(A) sites may be recognised via a similar DSE-dependent mechanism and may not require additional auxiliary sequence elements. Finally, results on the analysis of the EDF1 poly(A) signal show that, in a spliced environment, A-rich sequences are also 3'end formation effectors but depend on an competent upstream splicing reaction for efficient definition of the 3'end processing site.

Acknowledgments

This dissertation is dedicated to my family and the important people in my life for all the effort they have put into helping me achieve my goals. My mother Delfina, my father Júlio, my stepfather Júlio and especially to my grandparents Maria and António without whom everything in life would have been substantially more difficult.

This dissertation would not have been possible without the help and scientific knowledge of Dr. Andre Furger, Dr. Martin Dalziel and the work environment shared with Dr. Simon Haenni, Helen Sharpe, Cathy Browne and Kerstin Zechner.

Table of contents

List of figures

List of Abbreviations

A Adenine

AuxDSE Auxillary downstream sequence element

Bp Base pairs

C Cytosine

CPSF Cleavage and polyadenylation specificity factor

CstF Cleavage stimulatory factor

CTD Carboxyl-terminal domain

DNA Deoxyribonucleic acid

DNase Deoxyribonuclease

DSE Downstream sequence element

G Guanine

GFP Green fluorescent protein

GSP Gene-specific primer

HS Heat shock

Kb Kilo-base pairs

kDa Kilo-Daltons

MCS Multiple cloning site

min Minute(s)

miRNA microRNA

mRNA Messenger RNA

Nt Nucleotide(s)

OH Hydroxyl group

ORF Open reading frame

PAB II Poly(A) binding protein II

PAP Poly(A) polymerase

PCR Polymerase chain reaction

RNA pol I RNA polymerase I

RNA pol II RNA polymerase II

RNA pol III RNA polymerase III

Poly(A) Polyadenosine

Pre-mRNA Pre-messenger RNA

Pro Proline

PTB Polypyrimidine tract binding protein

R Purine

RNA Ribonucleic acid

RNase Ribonuclease

rRNA Ribosomal RNA

RT-PCR Reverse transcriptase polymerase chain reaction

sec Second(s)

Ser Serine

snoRNA Small nucleolar RNA

snRNA Small nuclear RNA

snRNP Small nuclear ribonucleoproteins

T Thymine

Thr Threonine

TMG Trimethylguanosine

TPE Telomere position effect

Tyr Tyrosine

u Unit(s)

U Uracil

U2AF U2 auxiliary factor protein

Ur Uridine-rich sequence element

USE Upstream sequence element

UTR Untranslated region

Y Pyrimidine

Overview and epistemological considerations

Precise definitions of *life* and its possible origins on Earth are, in scientific terms, contentious subjects[1-5]. Such fact has fundamental implications in terms of contemporary scientific thought.

Many, if not all, fundamental terms in science are problematic to define and therefore it is no surprise that “life”, perhaps the most abstract term in biology, is hardly definable. The primary recognition of “life” has been and still remains an essentially intuitive process for both scientists and non-scientists. “Life” is not a theoretical concept and one of the most interesting aspects of its intuitive process of recognition is the antagonistic aspect. The contrast between “living” and “non-living” rather than a precise theoretical or abstract content associated with one of these states allows a swift discrimination because, throughout evolution, this perception has been required in a multitude of biological and social situations. It is worth noting however, that such perception allow us to derive a notion, not a definition[1-3, 5]. The understanding that the fundamental object of study in biochemistry and molecular biology does not have a precise definition should clearly highlight that the primary objective of these particular subjects, and of science in general, is to formulate conjectural empirically testable hypotheses about observable phenomena rather than to reveal the truth or the essence about something[6, 7]. Scientists do not produce certainties but identify operative criteria. Science may be inspired by the search for the truth but works with hypotheses.

A further implication of the absence of a precise scientific definition of “life” is that the major theoretical possibilities that shape the general conceptions about this phenomenon are essentially philosophical. The three major concepts of life (in

the trivial sense of animals and plants) had already been constructed before the 19th century and are each associated with an outstanding philosopher[3]. Life as animation, as elaborated by Aristotle, which explains life in terms of a specific principle, the soul, which is both the ensemble of body functions as well as their coordination distinguishing living beings from all other natural beings and therefore defined as a special explanatory principle. Life as a mechanism, conceptualized by Descartes, where all the vital functions are no more than mechanisms and the living body is itself a machine. There is no real distinction between living and non-living bodies and therefore, at its limit, no requirement for a special type of explanatory principle or for terms such as “life”. Finally, life as organization, as developed by Kant in the *Critique of Judgment*, where an organized being is defined as a being where any part is both means and productive cause for the others in contrast to a machine where every part is a means relative to others but is not “produced” in any sense by them. Such a being is able to self-organize (self-maintain, self-repair and self-reproduce). This conception incorporates both animist and mechanistic ideas leading to the creation of the term *organism* which entered 19th century scientific vocabulary and still today pervades most scientific conceptions has a proof of its intellectual merit[3, 8].

The principle of a self-organized being, capable of self-maintaining, self-repairing and self-reproducing, extended through the advancements of cell biology, evolutionary theory, biochemistry and molecular biology is, still today, providing the theoretical framework underlying the three major operative criteria used to describe life in terms of its biochemical properties. As conceptualized by contemporary science the biochemical definition of a living organism rests on three basic principles. The first is the degree of chemical complexity and

organization where several thousands of different molecules contribute to form an intricate cellular structure which is the unitary base of life. The second is the ability of continuously exchanging energy and matter with the surrounding environment sustaining a state of dynamic equilibrium that enables the maintenance of the organic structure as well as to perform work. The third is the capacity of self-assembly and self-replication which is one of the most remarkable characteristics of the living state.

At the very core of all three statements stands the biochemical information coded in the genetic material of each organism.

Chapter 1

Introduction

1. Introduction

The fundamental defining structural feature of a eukaryotic cell is the enclosing of the majority of its genetic material in a discrete subcellular organelle denominated *Nucleus*. The nucleus of eukaryotes is characterized by an organization which reveals itself through the presence of distinct and complex higher-order functional structures[9, 10]. Nuclear compartments are conceptualized as being self-organizing entities generated in a cooperative manner by a multitude of stereospecific short-lived interactions which are promoted and stabilized by molecular crowding and, at least to some extent, might display a degree of fractal organization[11-16].

The human genome, which contains ~28.000 genes and 3.2 billion base pairs, is hierarchically compacted ~400.000 fold into chromatin fibers, chromosome domains and eventually chromosomes so it can fit a nuclear volume of ~1000μm^3 [17-20]. Packaging of DNA into chromatin is an extremely efficient way of solving the spatial constraints posed by a crowded and confined environment such as the eukaryotic nucleus. At its simplest level linear eukaryotic DNA does not exist as a naked molecule but packaged into an extremely organized and compact nucleoprotein structure known as

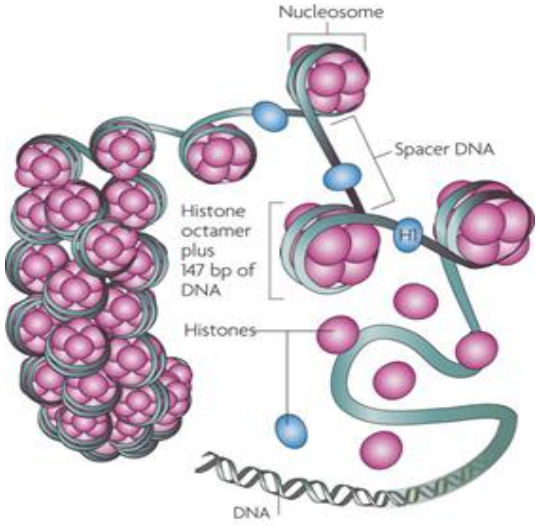


Figure 1: The nucleosomes are the basic units of chromatin. Each nucleosome is composed of approximately 147 bp of DNA wrapped around an octamer of histones (two copies of H2A, H2B, H3 and H4). Histone H1 binds to the DNA that links the two nucleosomes. (From: Luisa M. Figueiredo, George A. M. Cross & Christian J. Janzen. Nature Reviews Microbiology 7, 504-513 ,2009)

chromatin. The *nucleosome* is the basic architectural unit of the chromatin fiber and consists of 146 or 147 bp of DNA wrapped around an octamer core of basic proteins containing two copies each of four histone proteins: H2A, H2B, H3 and H4[21, 22]. Linker DNA joins nucleosomes into a beads-on-a-string like fiber at a 10 nm diameter. Association of the "linker histone" H1 with both nucleosomes and DNA promotes formation of the 30 nm thick fiber (Figure 1). Higher-order fibers of various diameters, whose precise *in vivo* geometry is unknown, are then formed by compaction of the primary fibers onto itself[18, 20]. The next level of organization consists in the folding of this chromatin fiber into subchromosomal domains of ~1Mb in size[23, 24] which, in turn, are folded to give rise to the interphase chromosomes[20]. These exist in a de-condensed state as chromosome territories[23] which are defined as dynamic nuclear compartments occupied by particular chromosomes showing the ability of intermingling with each other[25, 26]. Such chromosome territories are generally radially positioned in the nucleus, with most low gene density chromosomes located close to the periphery and high gene density chromosomes located near the center of the nucleus[9, 27].

In interphase nuclei of mammals and other eukaryotes, the most conspicuous form of structural genome organization however is not the chromosome territory but rather the organization of chromosomes into two large-scale chromatin states: *Heterochromatin* and *Euchromatin.* The two organizational states differ mainly in their DNA sequence composition, degree of condensation, epigenetic marking and, to a certain extent, in transcriptional activity[9]. Heterochromatin is generally characterized by more compact chromatin fibers which remain condensed throughout the cell cycle, AT-rich DNA sequences, high degree of repeated sequence elements, lower gene density, hypoacetylated

nucleosomes enriched in H3K9me3 and H3K27me3 and association with adaptor proteins: heterochromatin protein 1 (HP1)[28-31]. Heterochromatin can be further subdivided into constitutive and facultative heterochromatin. Constitutive heterochromatin is characterized by large arrays of repeated sequence elements that are mainly inactive and are organized into hypoacetylated nucleosomes enriched in H3K9me3 which is bound by HP1. Facultative heterochromatin is generally characterized by H3K27me3 and can be described as the developmentally regulated transition of euchromatic regions into heterochromatic-like features, which in its extreme form can inactivate an entire chromosome as is the case in X-chromosome inactivation in mammalian females[32]. Euchromatin is characterized by generally more open chromatin fibers, higher gene density, hyperacetylated nucleosomes enriched in H3K4me3 and in general increased gene activity[28-31]. Average heterochromatin is about 1.4-fold more condensed than euchromatin[31]. The ability of genes to be activated is not necessarily lost when chromatin is packed in more compact fibers and, conversely, inactive genes when placed close to active genes in an open chromatin environment might still stay inactive[29]. The present view, suggested by extensive work, is that heterochromatic domains provide an environment which is accessible but not favourable for the expression of euchromatic sequences and vice-versa[9].

Chromatin domains provide the nuclear landscape in which genetic information, through highly regulated mechanisms, must become accessible to support complex DNA-dependent multistep processes such as Transcription, DNA repair and DNA Replication.

1.1. Eukaryotic gene expression

Eukaryotic gene expression is a highly complex multi-step process. Three DNA-dependent RNA polymerases (RNA pol I, II, III) are responsible for transcribing the genes in the eukaryotic nucleus. RNA pol I transcribes most of the ribosomal RNAs (28S, 18S, and 5.8S rRNAs), RNA pol II transcribes messenger RNAs (mRNAs) as well as some small nuclear RNAs including snRNAs and RNA pol III synthesizes 5S rRNA, transfer RNAs (tRNAs) and some snRNAs[33, 34]. All three enzymes present conserved structures and each consist of ten core subunits surrounded by a number of polymerase-specific subunits[35, 36]. Even though the combined activities of RNA pol I and RNA pol III exceed 80% of the total RNA synthesis in growing cells, the main focus has been on RNA pol II as it transcribes the broadest variety of sequences and all protein encoding genes. Furthermore RNA pol II genes constitute the largest proportion of the transcribed genome[33, 34].

In order to activate gene expression, transcriptional activator proteins promote local chromatin decondensation allowing access of the general transcription factors (GTFs) to their target promoter sequences[37]. Unlike in the prokaryotes, eukaryotic RNA polymerases are not able to recognize the promoters of their target genes on their own. They rely on a series of accessory factors known as general transcription factors to direct them[38, 39]. These protein factors recognize conserved sequences, like the “TATA” box or “initiator” sequences on their target genes and create a protein platform onto which the RNA polymerases are recruited thus enabling transcriptional initiation. Following assembly and initiation, transcriptionally engaged RNA polymerases then proceed to precisely copy the template into a precursor messenger RNA (pre-mRNA) molecule in a phase known as elongation. This process is highly regulated and controlled by the

binding of elongation factors to the initial transcription complex. Finally, in the case of RNA pol II driven transcription, elongation termination will be triggered by the recognition of a functional 3'end formation site (pA) located near the end of the transcript[33].

Although Transcription is an essential step at the very beginning of eukaryotic gene expression, it's not the only regulated process. The resulting precursor messenger RNA molecules are subjected to processing reactions: Capping, Editing, Splicing and 3'End Formation. During pre-mRNA processing, the primary transcript is modified and matured into an export and translation competent mRNA molecule. The final mature mRNA is subsequently exported into the cytoplasm and translated into proteins by the ribosomes. All these steps are highly regulated and play a fundamental part to precisely control gene expression[40-42].

1.2. Interconnections between Transcription and pre-mRNA Processing Reactions

The initial approach to study mRNA maturation was essentially based on *in vitro* experiments where each of the reactions was studied in isolation. The ability to reproduce each of the reactions in *in vitro* reconstituted systems of purified pre-mRNA substrates led to the discovery of the protein machineries and the finding of sequences on the pre-mRNAs that are essential for the processing reactions[43, 44]. However, this also led to the misconception that pre-mRNA processing is independent and temporally separated from transcription. In recent years though, it became clear that the processing reactions are tightly interlinked with the

transcription process and highly dependent on the components of the transcription machinery. *In vivo*, the processing reactions are functionally connected and it is now widely accepted that transcription, capping, splicing and 3'end formation occur co-transcriptionally in highly organized "factories" (Figure 2) and that the coupling of these processes enhances both the efficiency and the scope of regulation[41, 45-56]. Indications that optimal pre-mRNA processing is dependent on transcription by RNA pol II arose from early observations that the transcription of a RNA Pol II gene when driven by a RNA Pol I or a RNA Pol III promoter leads to total or partial repression of capping, splicing and polyadenylation[57-59]. Furthermore the observation that 3'end processing factors are present at promoters (for example, the cleavage and polyadenylation specificity factor (CPSF) interacts with the general transcription factor TFIID bound to the core promoter[60] and the cleavage and polyadenylation stimulatory factor (CstF) interacts with the transcriptional co-activator PC4[61] are additional compelling evidence for the cotranscriptionality of pre-mRNA processing.

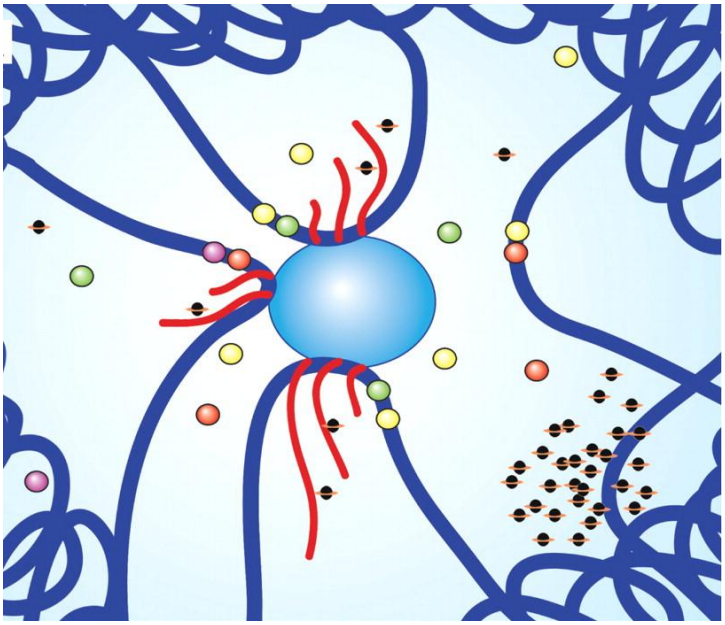


Figure 2: Schematic diagram of transcription of multiple genes at a nuclear RNAPII transcription factory. RNAPII factory shown as central blue circle with three transcribing genes and their associated transcription factors (small colored circles). Nascent transcripts are shown in red, chromatin is dark blue, and splicing components are depicted as small black circles with orange halo. (From: Chakalova, L. & Fraser, P. Organization of transcription. Cold Spring Harb Perspect Biol 2, (2010)).

1.3 Integration of Transcription and Processing through the RNA Polymerase II Carboxyl Terminal Domain (CTD)

The molecular engine of the transcriptional apparatus is RNA pol II[44] which is a multimeric protein whose 12 subunits are highly conserved throughout eukaryotes[33]. The largest subunits of RNA pol I, II and III are highly homologous but one particular domain at the carboxyl terminus of this subunit is only found in RNA pol II[51]. This carboxyl terminus domain (CTD) is highly conserved in RNA pol II of different eukaryotes[62, 63] being positioned just outside the overall globular RNA pol II 3D structure just below the RNA exit channel[64] (Figure 3). The CTD of RNA pol II is composed of imperfect tandem repeats of the heptad consensus sequence Tyr[1]-Ser[2]-Pro[3]-Thr[4]-Ser[5]-Pro[6]-Ser[7] (YSPTSPS) which show a variable number of repeats in different species: 26 in yeast, 42 in flies and 52 in mammals[65-67]. The

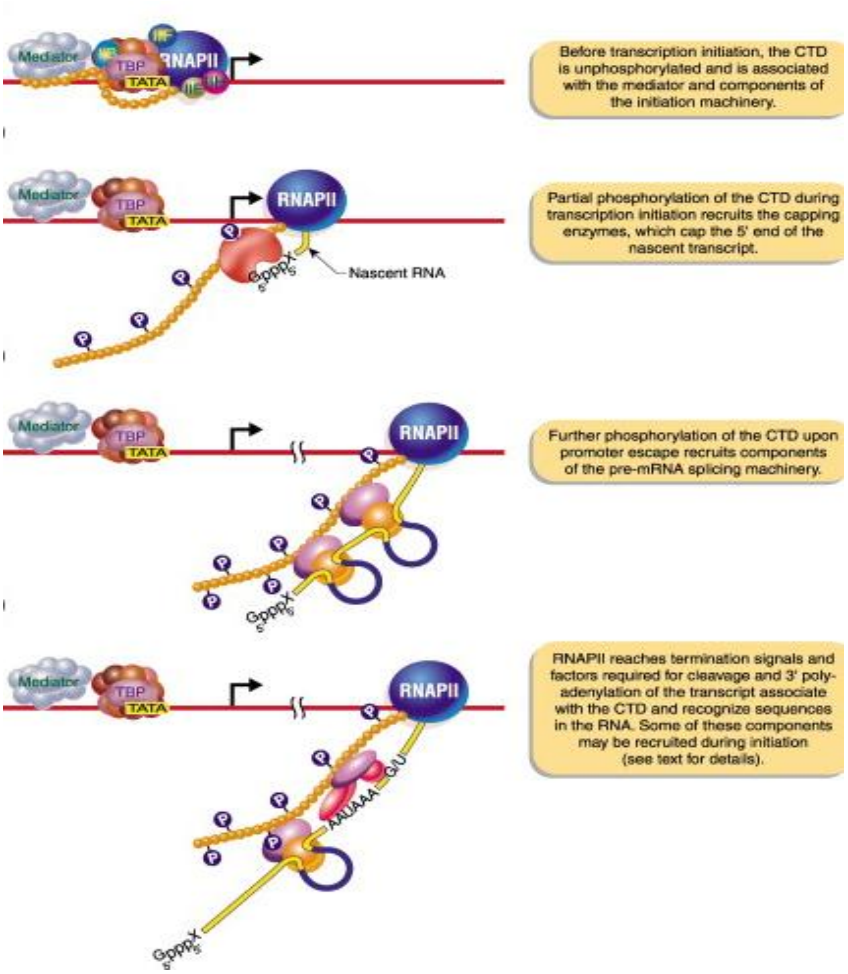


Figure 3: C-Terminal Domain (CTD) of RNA Polymerase II coordinates transcription and pre-mRNA processing. (From: Orphanides, G., Reinberg D. (2002). Cell. 108(4): 439-51)

CTD can be differentially phosphorylated on five of these seven residues and, most importantly, on all three serine residues located in these heptapeptide structures[68]. The reversible phosphorylation states of the serine residues depend on which of the three main cycles of transcription RNA pol II is engaged. Serine5 is phosphorylated during early elongation (approximately 150bp downstream of the transcription start site (TSS)) by the TFIIH subunit Kin28 in budding yeast and CDK7 in metazoans[69]. Serine2 , on the other hand, is predominantly modified during elongation and up until transcription termination via the kinase Ctk1 in budding yeast and CDK9 , which is associated with the positive transcription elongation factor b (P-TEFb), in metazoans[70]. The transition from a Ser5 phosphorylation state to a Ser2 phosphorylation state is thought to be controlled by a CTD phosphatase known as Rtr1[71]. Finally, Serine7 seems to be phosphorylated by Bur1 during early elongation (about 50bp downstream of the TSS) and remains in this state until transcription terminates[72]. Serine7 phosphorylation has been linked with intron removal as well as with the recruitment of the pA site-independent termination factor Nrd1[73] and is required for snRNA expression[74, 75]. Upon transcription termination, the CTD is dephosphorylated by phosphatases such as Fcp1, which enables reassembly of RNA pol II with the pre-initiation complex (PIC) for re-initiation of transcription[76].

Numerous protein factors required for efficient transcription elongation and termination[77] as well as for pre-mRNA processing and export bind the CTD according to its serine phosphorylation pattern[70]. It appears as though the three repeated CTD serine residues are phosphorylated depending on the gene promoter structure and this is thought to allow the control of pre-mRNA synthesis and processing[77]. The CTD enhances the efficiency of all mRNA processing

reactions (capping, splicing, 3'end formation and RNA editing) *in vitro* and deletion of CTD repeats results in severe processing defects[45, 50, 78, 79]. Conversely, the capping reaction appears to be essential for transcription re-initiation and elongation, the splicing reaction has been shown to have a stimulatory effect on transcription and 3'end formation has been shown to strongly impact at the level of transcription[47, 68, 80]. The differential regulatory phosphorylation patterns of the CTD integrate the different phases of the transcriptional cycle with the corresponding pre-mRNA processing events[67, 77, 81].

1.4 Processing Reactions

In order to produce a fully mature eukaryotic mRNA, all primary RNA pol II transcripts must undergo three major nuclear modifications or *Processing reactions*: 1) Capping: the 5'end of each emerging pre-mRNA is modified by the addition of a 7-methylguanosine cap structure. 2) Splicing: intervening sequences (introns) are excised via a complex splicing mechanism. 3) 3'end formation: 3'ends of all pre-mRNAs are generated by an endonucleolytic cleavage reaction followed by the polymerisation of a non-templated poly(A) tail with the exception of replication-dependent histone genes where endonucleolytic cleavage occurs but is not followed by non-templated poly(A) tail polymerisation[45, 82].

1.5. Capping

One unique feature of RNA pol II nascent transcripts (20-30 nucleotides long) is that their 5'ends are co-transcriptionally modified[83, 84] by the addition of a m^7G(5')ppp(5')N cap structure[85, 86]. 5'capping protects mRNA from 5'-3'

exonuclease degradation[87], is essential for proper translation initiation[88] and has been shown to be essential for efficient gene expression and thus for cell growth[89].

The first mRNA processing factors to be recruited to the phosphorylated RNA pol II CTD during the transcription cycle are the capping enzymes (CEs): RNA 5'-triphosphatase (CE-RTp), RNA guanylyltransferase (CE-RGt) and RNA (guanine-7)-methyltransferase (CE-RMt)[51]. In mammals both the triphosphatase and the guanylyltransferase activities are on the same bifunctional polypeptide at the N-terminus and at the C-terminus respectively while in yeast these activities are performed by a heterodimer encoded by two different genes *CET1* and *CEG1*. Both in mammals and in yeast the methyltransferase activity is carried out by a different individual protein[37].

During early initiation of transcription RNA pol II transcribes 20-30 nucleotides and then pauses due to the association of the negative elongation factor (NELF) and DRB-sensitivity inducing factor (DSIF) to the transcriptional

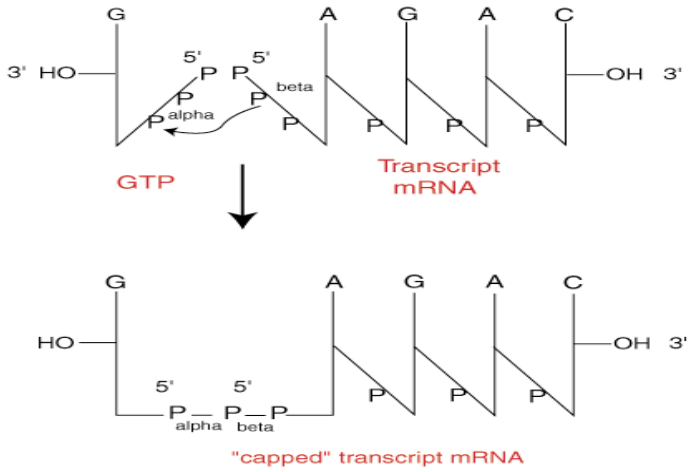


Figure 4: Capping reactions catalyzed by CE-RTp and CE-RGt

complex. Subsequently CE associate and cap the pre-mRNA[90]. After capping is completed, P-TEFb phosphorylates Ser2 of the CTD, releases the polymerase

from its arrest and allows it to enter a higher processive state. In mammals and *C. elegans*, the CE performs the first two of the three major reactions involved in capping[91]: the RNA 5' triphosphatase hydrolyzes the triphosphate of the 5' nucleotide from the nascent transcript to a diphosphate through removal of the γ-phosphate of the first nucleotide; subsequently the RNA guanylyltransferase catalyzes the fusion of guanylate (GMP) through a unusual 5'-5' triphosphate linkage between the α-phosphate of the GMP and the β-phosphate of the first nucleotide. Finally the RNA (guanine-7)-methyltransferase methylates the N7 position of the terminal guanine completing the cap structure[37, 45, 92]. Additionally, the two nucleotides directly adjacent to the cap may be methylated at the 2' hydroxyl position, which has been shown to occur in most eukaryotes with the exception of *Saccharomyces cerevisiae*[93]. The cap structure is then recognized by the cap binding complex (CBC) which is composed of two evolutionarily conserved proteins, CBP20 (20kDa) and CBP80 (80kDa), that bind as a heterodimer to the m^7G cap shortly after its formation[94, 95]. The CBC retains its association with the mRNA throughout transcription, processing and nucleocytoplasmic export[94, 96]. Upon export through the nuclear pore complex the cap binding proteins are substituted by the cytoplasmic translation initiation factor eIF-4E[91]. The CBC bound to the cap structure protects mRNAs against 5'-3' exonuclease activities and enhances the splicing of the first intron[95, 97]. The cap structure also promotes 3'end processing, facilitates mRNA cytoplasmic transport and assists translation[98, 99]. Additionally it has recently been revealed that the 5'cap may possess previously unreported roles within eukaryotic cells. In *Arabidopsis thaliana* the CBC appears to be involved in alternative splicing as well as processing of microRNAs[100, 101] and a connection between the CBC and alternative splicing has

been established in mammals[102]. In the cytoplasm the cap and poly(A) binding proteins mediate mRNA circularization enhancing translation initiation[103].

1.5.1. Integration of Transcription and Processing - Capping

The findings that RNA pol II coding regions transcribed by RNA pol III[104] and RNA pol I[105] were not capped pointed to the possibility of capping reaction catalytic enzymes being directly targeted to RNA pol II transcription complexes. This specificity is understood to be achieved through the RNA pol II CTD interactions, which Pol I and Pol III lack, since it was shown that both mammalian and yeast CEs interact with the phosphorylated form of RNA pol II (RNA pol IIO)[106, 107] and truncations of the CTD reduce the cellular levels of mRNAs with 5'caps[90]. During early elongation, after RNA pol II has synthesized 20-30bp of pre-mRNA, both NELF and DSIF bind to the polymerase and the phosphorylation of the CTD at Ser5 residues by the cyclin H dependent kinase (CDK7-cyclinH, a component of the basal transcription factor TFIIH) is sufficient to trigger the recruitment and stimulation of the capping enzymes via association with the Spt5 subunit of DSIF and promote capping[108, 109]. The promoter proximal halting of RNA pol II induced by NELF and DSIF is thought to allow time for efficient capping. In yeast the RTp-RGt complex and RMt bind directly and independently to the phosphorylated CTD and the phosphorylation of Ser5 at the promoter by a subunit of TFIIH (Kin28) is necessary for their recruitment[51, 110, 111]. The removal of the Ser5 phosphates from the CTD during early elongation is correlated with dissociation of CEs from the elongating polymerase downstream of the promoter[51, 110, 111] pointing out the importance of the CTD phosphorylation pattern in the coordination of capping and transcription. Upon phosphorylation of the CTD Ser2 residues and of Spt5 by a different cyclin dependent kinase (CDK9) the elongation

complex is assembled and transcription resumes[54, 108]. Interestingly, *in vitro* the recruitment of the CE to RNA pol II may aid lifting the NELF induced repression of transcription, thus allowing elongation to take place[112] and providing a quality control for pre-mRNAs only permitting 5' capped transcripts to be further synthesised. Contrastingly, in budding yeast, where there are two separate enzymes responsible for the RNA triphosphotase (Cet1) and guanyltransferase (Ceg1) reactions, it has been shown that Cet1 may in fact repress the re-initiation of transcription[113].

The factors associated with transcription also play an important role in capping. The binding of the CE to the Ser5 phosphorylated CTD has been shown to increase the affinity of guanylyltransferase to GTP nearly two-fold[109].

In budding yeast, Cet1 and Ceg1 can directly interact with the Ser5 phosphorylated CTD[111], and the cap-methyltransferase (Abd1) has been detected at the very 3' of nascent transcripts due to its interaction with the CTD[111]. Abd1 may also play a role in promoter clearance and/or early elongation[114]. Additionally, on some promoters, it appears that Ceg1 may enhance elongation[113, 115]. Furthermore, the TFIIH-associated kinase Mcs6 has been shown to recruit the pTEF-b/Pcm1 (mRNA cap methyltransferase) complex, thus linking capping with transcription elongation[116].

1.6. Splicing

The open reading frame (ORF) of a typical mammalian gene is interrupted by non-coding sequences (introns) which on average are 3000 base pairs (bp) in length. The average size of a coding sequence (exon) which contains the

information for the encoded functional product is smaller and around 150bp long[117]. In order to generate a mature and functional mRNA molecule with a continuous open reading frame, introns must be precisely removed and the exons must be fused and ligated. This is accomplished through a splicing reaction that relies both on consensus sequences within the pre-mRNA sequence as well as on a complex multicomponent system – the *Spliceosome* –. The Spliceosome consists of five catalytic small ribonucleoproteins (snRNPs) U1 snRNP, U2 snRNP, U4 snRNP, U5 snRNP, U6 snRNP (each consisting of the corresponding uridine-rich snRNA: U1 snRNA, U2 snRNA, U4 snRNA, U5 snRNA and U6 snRNA, bound by eight Sm proteins (with the exception of the U6 snRNP) as well as particle-specific proteins[118] plus a large number of proteins associated with the pre-mRNA, forming a large complex[119-123]. Spliceosome-associated proteins play essential roles in splicing regulation, including alternative splicing. They include serine/arginine rich (SR) proteins which function as general activators of exon definition and heterogeneous nuclear ribonucleoproteins (hnRNPs) which are the best characterized group of splicing silencers[124]. Furthermore, a minor spliceosome consisting of U11, U12, U4atac, U5 and U6atac snRNPs has been shown to splice non-canonical introns[125].

The pre-mRNA sequence consensus elements consist of the 5'exon-intron junction or splice site (5'SS:▼) which, in mammals, reads the consensus sequence 5'- *AG▼GURAGU*-3' (R: purine; Y: pyrimidine; N: any nucleotide), the 3'splice site (3'SS:▼) which defines the intron end 5'-*YAG▼RNNN*-3' and the branch point (BP) site 5'-*CURA*$^{2'OH}$*Y*-3'. The BP site contains an Adenosine (highly conserved in budding yeast, degenerated in higher eukaryotes) which lies approximately 100 nucleotides upstream of the 3'SS and between the BP site and the 3'SS there is a

variable stretch of pyrimidines called the polypyrimidine tract[45, 126]. The removal of introns by splicing involves two *trans*-esterification steps resulting in a spliced RNA and introns that form a lariat structure (Figure 5A). The first step involves the nucleophilic attack of the 2'OH of the branchpoint Adenosine to the phosphodiester bond of the 5'splice site resulting in a free 5'exon with a 3'hydroxyl group and a lariat shaped molecule comprising the intron sequences and the 3'exon. In the second step the 3'hydroxyl group of the 5'exon attacks the phosphodiester linkage at the 3'splice site and this second *trans*-esterification reaction results in the fusion of the two exon sequences and the release of the lariat shaped intron[45, 127].

The Spliceosome complex is composed of multiple subunits and involves

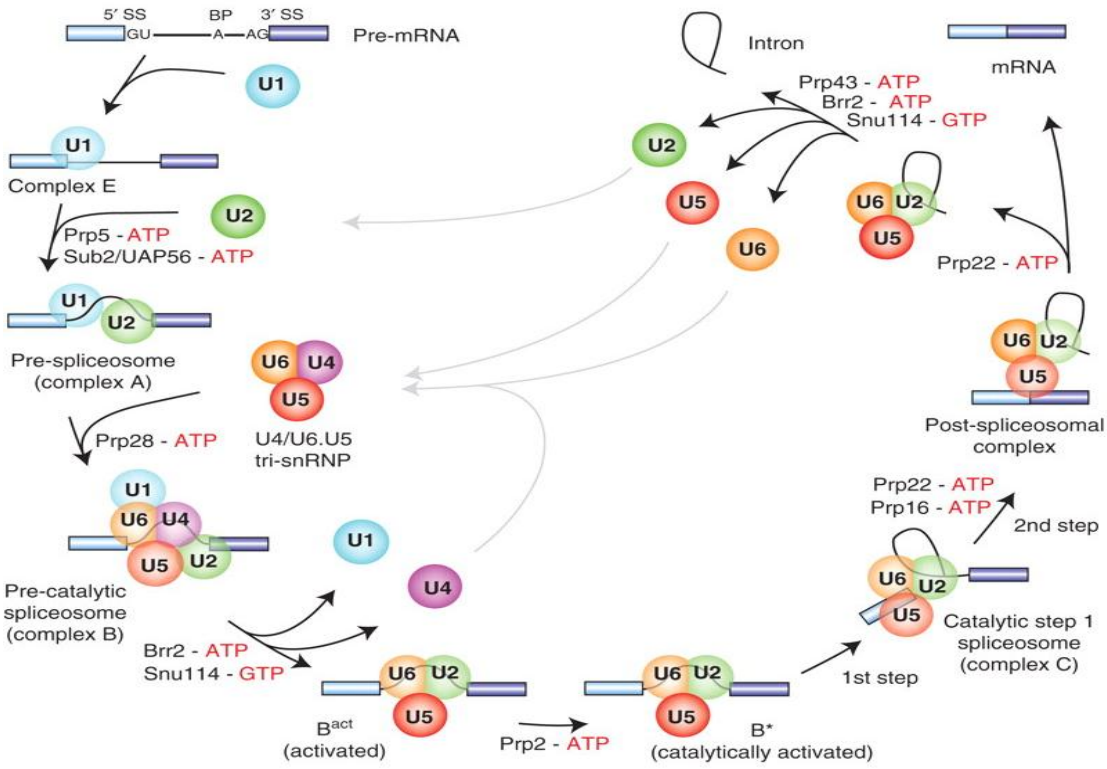


Figure 5A: Pre-mRNA splicing by the U2-type spliceosome. Canonical cross-intron assembly and disassembly pathway of the U2-dependent spliceosome. (From: Will, C.L. & Luhrmann, R. Spliceosome structure and function. Cold Spring Harb Perspect Biol 3 (2011))

more then 300 polypeptides[120, 128, 129]. The Spliceosome facilitates splicing and the first step of its assembly is the interaction of the U1 snRNP with the 5' splice site and recognition of the polypirimidine tract and of the AG of the 3' splice site by the dimeric U2 Auxiliary Factor (U2AF) by $U2AF_{65}$ and $U2AF_{35}$ subunits respectively. These interactions help the recruitment of the U2 snRNP and of the Branchpoint Bridging Protein (SF1/mBBP) to the branchpoint. U2 snRNP base pairs with the branchpoint sequences leading to the bulging of the branchpoint adenosine needed for the first nucleophilic attack. Subsequently U4 snRNP, which does not interact directly with the pre-mRNA, bridges the connection between U5 snRNP and U6 snRNP and recruits these two factors into the spliceosome. Upon displacement of the U1 snRNP from the 5' splice site, U6 snRNP interacts with these sequences and with the U2 snRNP generating a catalytic core that places the 5'SS and the branchpoint in close proximity creating the necessary conditions for the first nucleophilic attack[45, 130, 131]. Finally the U5 snRNA loop interacts with both 5' and 3' exon sequences and its role is thought to be the positioning of the two exons in close proximity so the second nucleophilic attack can occur at this stage[105]. Each time one intron is removed the spliceosome is believed to undergo a certain level of recycling. However, the mechanism of spliceosome recycling between successive introns in a given transcript remains an interesting and contentious open question[132].

Recognition of splice sites is a complex and highly regulated process which can occur via intron or exon definition. In exon-defined splicing, the 3' splice site of an upstream intron and the 5' splice site of a downstream intron interact across and adjacent exon. Alternatively, intron-defined splicing occurs on the splice sites of the same intron. It is believed that the gene structure dictates which model of

splicing is used and both require the essential regulatory action of SR proteins. In vertebrate genes, which generally consist of short exons and long intronic sequences, exon-defined intron removal is predominantly observed. Conversely, fission yeast genes, which contain short intronic sequences, are mostly spliced via an intron-defined splicing mechanism (reviewed in[133]).

Studies have shown that spliceosome may assemble prior to any interaction with the pre-mRNAs[134]. It has been observed that these associations can occur in the absence of actively transcribed pre-mRNA in Cajal bodies and nuclear speckles[135-137]. However, it is argued that the formation of these holo-spliceosomes may be transient and may never be functional[138].

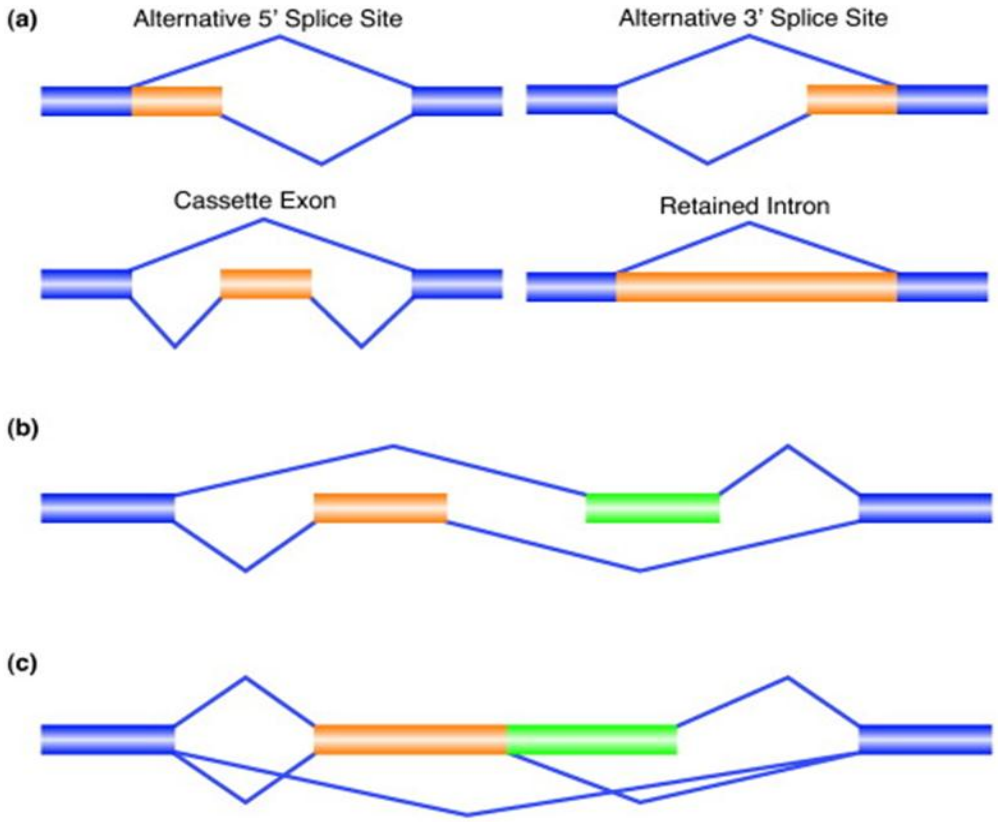


Figure 5B: Types of alternative splicing. (a) The basic building blocks of all types of splicing events are depicted. Alternative 5′ splice sites; alternative 3′ splice sites; cassette exons; retained introns. (b) An example of a mutually exclusive splicing event which is built from two cassette exons. (c) An example of a complex splicing event containing a cassette exon and an alternative 5′ splice site. (From: Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463, 457-463 (2010))

An essential role of splicing is to produce diverse pre-mRNA isoforms which can then be translated into functionally distinct proteins. This process is known as alternative splicing, involves the recognition and use of different splice sites (Figure 5B) and is controlled by splicing regulators such as SR proteins. Permutation of exons allows the synthesis of mRNA molecules with alternative sequences and thus proteome expansion. The developmental complexity of mammals is in part attributed to alternative splicing and it is believed that 95-100% of human pre-mRNAs that contain sequences corresponding to more than one exon are processed to yield multiple mRNAs[139-141].

1.6.1. Integration of Transcription and Processing – Splicing

Pre-mRNA splicing, in contrast with 5'capping which is tightly coupled with transcription reinitiation[51] and with 3'end formation which is closely linked to transcription termination[142, 143], can proceed either co-transcriptionally[144-146] or post-transcriptionally after transcript release from the DNA template[145-151]. The distinction is fundamentally relevant since co-transcriptional splicing allows functional integration of the transcription and pre-mRNA processing machineries allowing them to modulate each other. In contrast, post-transcriptional splicing might allow additional regulatory mechanisms to operate or couple splicing with other downstream events such as mRNA export[42, 144]. From a theoretic-logical point of view co-transcriptional splicing is likely to predominate for most introns in eukaryotic cells. It provides efficient recognition of splice sites as they emerge from the elongating RNA pol II, offers splicing regulation through transcription factors and transcription associated epigenetic regulators and, last but not least, constitutes an important mechanism to facilitate transcription[42, 144].

Several lines of evidence support the co-transcriptional splicing mechanism for most constitutive splicing events. Co-transcriptional splicing was initially observed by electron microscopy in *Drosophila* embryos and looped RNAs attached to chromatin were revealed[152-154]. Furthermore spliced mRNAs have been shown to be associated with mechanically dissected or biochemically fractionated chromatin[145, 146], spliced RNAs are also detected at their gene loci by RNA *in situ* hybridization with splice-junction probes[155] and, in transcriptionally synchronized cells, introns are removed from the nascent RNA before the completion of transcription[156]. Additionally, *in vivo* recruitment of splicing factors and spliceosome assembly occur while the nascent RNA is still attached to chromatin via RNA pol II (co-transcriptionally) in both yeast[157-160] and mammalian cells[161] which shows that splicing is initiated co-transcriptionally even if, in some cases, it might be post-transcriptionally concluded[147]. Recently, transcriptional pausing was also show to be linked to co-transcriptional splicing[162-164] but the extent to which the recruitment of spliceosomal components is influenced by direct interactions with RNA pol II remains unclear[157]. However, two additional important concepts have emerged from the works on coupling between splicing and transcription: 1) promoter identity and RNA pol II processivity influence splicing outcome[165, 166] which potentiates RNA splicing regulation by epigenetic strategies since both chromatin remodelling factors and specific chromatin marks have been implicated in the modulation of alternative splicing[167-169]; 2) co-transcriptional splicing can feedback into transcription[170-172] because although it is possible that specific splicing factors could have dual independent roles, recent work seems to reinforce the notion of mutual influence since splice sites can function as a checkpoint for elongating RNA pol II complexes[164].

In vitro, spliceosome assembly and splicing efficiency have also been shown to be activated by the presence of the CTD with the activation effect being dependent on the presence of complete exons with both 3' and 5' splice sites on the precursor RNA substrates strongly suggesting a mechanism of interaction between the CTD and splicing factors dependent on exon definition[173]. CTD truncation causes inefficient splicing in mammalian cells and inhibition of colocalization of splicing factors with transcription sites[81, 174]. Consequently the CTD is thought to enhance coding sequence (exon) definition and stabilize combinatorial interactions of various factors bound to the exon and so facilitate assembly of the spliceosome[126]. However, considering the observation that capping stimulates splicing[42] and that the CTD is necessary for the recruitment of the CEs[61] the interpretation of these results should be cautious. A number of results seem to argue that the CTD plays a direct role in splicing independently of its function in transcription and capping[175]. Several studies showed that the CTD of RNA pol II is indeed required for the recruitment of splicing factors to sites of transcription[94, 174, 176-179]. Additionally, recent work shows that direct interactions between the CTD with U1snRNP and various SR proteins also seem to occur[180] although the validity and significance of the latter associations is still being debated[181]. The recently discovered interaction of the U2AF65-PRP19 complex and the CTD seems to suggest that this region of RNA pol II is responsible for the activation of splicing[182] even though the observation that fusion proteins consisting of the body of either mammalian RNA pol III or bacteriophage T7 pol and the RNA pol II CTD were incapable of promoting capping or splicing of pre-mRNAs[183] seems to point that the CTD alone may not be competent in directing intron excision. One reason however, why, at least in the case of T7 RNA pol,

transcription does not support coupled pre-mRNA processing might be that this polymerase elongates several times faster than RNA pol II[132]. Additional evidence for CTD functional relevance comes from studies on the CTD phosphorylation pattern and related effects. *In vitro*, anti-CTD antibodies and CTD peptides can inhibit splicing and expression of phosphorylated CTD peptides has a similar effect on mammalian cells *in vivo*[178, 184]. Furthermore, *in vitro*, the hyperphosphorylated form of RNA pol II (RNA pol IIO) significantly enhances the rate and frequency of spliceosome assembly on a variety of pre-synthesized transcripts while the hypophosphorylated RNA pol II (RNA pol IIA) inhibits these reactions suggesting an interaction dependent on the CTD phosphorylation state[42, 54, 79, 173]. This interaction probably results from RNA pol IIO-dependent stimulation of early spliceosome assembly by, possibly, facilitating the binding of snRNPs to the nascent transcript[42] but it is also worth noting that, plausibly, not all intron-containing transcripts will be equally sensitive to the presence of the CTD in splicing assays[94]. Interestingly, RNA pol II CTD seems to be able to influence alternative splicing. Inclusion of a fibronectin alternative exon was tested in cells carrying a mutant RNA pol II version lacking the CTD, with the results showing that the alternative exon was more likely to be retained when compared to transcripts synthesised by wild type RNA Pol II[185]. Currently, two models try to explain this phenomenon. The recruitment model hypothesizes that the lack of the CTD may prevent SR proteins from binding to RNA pol II which would inhibit alternative splicing[185]. However, evidence against direct interactions of SR proteins with the CTD has been provided[181]. The elongation model, is supported by observations that the alternative exon is more likely to be included in transcripts synthesised by a mutant form of RNA pol II, which exhibits a lower rate of

elongation[165]. This slower pace of transcription is thought to allow more time for the recognition of weaker splice sites. At present, it is believed that both the recruitment and the elongation mechanisms work together to influence alternative splicing[186].

Further evidence for the tight interrelation between transcription and splicing comes from observations which show that splicing can positively influence transcription at the promoter. The presence of a promoter-proximal splice site has been shown to increase the levels of pre-mRNA synthesis in transgenic mice[187] and in HIV[188]. Moreover a connection between promoter structure and alternative splicing has also been suggested. Inhibition of promoter-associated transcription factor binding by site directed mutagenesis in the fibronectin promoter controlling the expression of an α-globin/fibronectin minigene showed that the EDI exon was more likely to be incorporated in mRNA transcripts generated from the mutant promoter vector, thus indicating that the structure of the promoter has an influence on alternative splice site choice[189]. It was further shown that the sequence of the promoter can affect the activation of the SR protein SF2/ASF, which is involved in alternative splicing[166] and hypothesized that the promoter itself may be responsible for recruiting these splicing enhancers since it has been observed that the transcriptional co-activator p52 directly binds to SF2/ASF[190].

Interestingly, transcription initiation and elongation also appear to be influenced by splicing. Higher levels of initiation factors such as TFIID and TFIIH are bound to the promoter of transcripts containing a functional 5' splice site, when compared to pre-mRNAs with a mutated version of this sequence [170] and U2 snRNP appears to be required for efficient RNA pol II elongation[172]. The

interaction of the U2 snRNP with the elongation factor TAT-SF1 is also thought to allow a more efficient assembly of the spliceosome[172].

Importantly, the excision of introns seems to be additionally regulated on the chromatin level. Recent studies show that exon sequences are more likely to be associated within a nucleosome than introns and it is hypothesized that this arrangement acts as an obstacle, slowing down the RNA Pol II elongation rate and thus allowing the recognition of splice sites linked to exon borders[191, 192]. Exon-related pausing has also been observed at terminal exons in *S. cerevisiae* genes[164], which is thought to allow sufficient time for intron excision before RNA pol II termination occurs at the end of a gene[162]. Additionally, histone modifications may direct the recruitment of splicing factors such as PTB to the transcription complex, and may thus play a role in alternative splicing[168].

Finally, the association of 5' and 3' splice sites with RNA pol II known as exon tethering, is another example of a close interaction between splicing and transcription. It has been shown that efficient splicing and transcription of the upstream and downstream exons can proceed even in the presence of a co-transcriptional cleavage event within the intermediate intron[193]. Recently, contradictory results have been presented and exon tethering still remains a contentious subject[194].

1.7. 3'End Formation: Cleavage and Polyadenylation

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 subsequently subjected to polyadenylation resulting in mature transcripts with uniform poly(A) tails consisting of an average of 200 adenylate

residues. The exception to this generalization are replication-dependent metazoan histone genes which undergo a 3'end cleavage event but not the non-templated polymerisation of a poly(A) tail[45].

Cleavage and polyadenylation define a critical biochemical feature in eukaryotic gene expression promoting transcription initiation, transcription termination, terminal intron removal, nuclear cytoplasmic transport, translation initiation and stability of the mRNA[42, 45, 64, 195-197].

The 3'ends of nascent pre-mRNAs are processed via a multi-step mechanism. This process is initiated by the recognition of *cis* elements that constitute the polyadenylation (poly(A)) sequence[64, 198]. The actual processing reaction consists of two steps: endonucleolytic cleavage of the nascent transcript downstream of the poly(A) hexamer sequence and poly(A) polymerase (PAP)-dependent polymerisation of the polyadenylate tail onto the 3'OH end of the 5'cleavage product[199].

1.7.1. Integration of Transcription and Processing - 3'end formation

The 3'end formation reactions are also tightly linked with RNA pol II. Initial indications of this interaction were obtained from studies showing that in cells expressing CTD-truncated RNA pol II, pre-mRNA 3'end cleavage appears to be strongly inhibited[81]. Furthermore, in the same study, subunits of both CPSF and CstF were seen to specifically bind a CTD affinity column. Additionally, three of the four subunits that compose CPSF have been shown to be associated with TFIID (basal transcription factor) and to be transferred, after initiation, to the

phosphorylated form of RNA pol II (pol IIO) travelling along the gene with RNA pol II during elongation[60]. Recent investigations suggests that CPSF is recruited to the transcription complex at the promoter, and then remains associated with the elongating RNA pol II up until 3'end formation at the end of the gene[200, 201]. Presently, however, there is conflicting information whether CPSF does indeed bind directly to the CTD or rather associates with other factors within the elongation complex[175]. Despite some unresolved questions, it has been shown that CTD length is a critical feature for activating cleavage and polyadenylation, the requirement being for a CTD with more than 26 heptad repeats in order to function efficiently[202]. Additional evidence for tight links between RNA pol II and 3'end formation factors came from experiments showing that CFI binds to RNA pol II at the promoter region and remains associated with the transcription complex until 3'end formation occurs[203]. This evidence has recently been further supported by the observation that the Pcf11 subunit of the cleavage and polyadenylation factor CFIA and the termination factor Rtt103 directly interact with Ser2 phosphorylated CTD in budding yeast[77, 204]. Such interaction is thought to ensure that the action of these factors remains confined to 3'end processing sites[205]. Interestingly, *in vitro*, the addition of both hypo- and hyperphosphorylated CTD can instigate 3'end formation in the absence of transcription[78]. Additionally, inactivation of CTK1, a Ser2/5 CTD kinase, disrupts 3'end formation[206]. Taken together these observations suggest that the interaction between some 3'end processing factors and the CTD may not always correlate with the phosphorylation state of this RNA pol II subunit.

Both CPSF and the 64 kDa subunit of CstF are, according to some studies, thought to be recruited to RNA pol II at the promoter. In mammalian cells, CPSF is

loaded onto the body of RNA pol II via the transcription factor TFIID during transcription initiation[60] and CstF64 associates with the CTD-bound transcriptional co-activator PC4 being this association conserved between the homologs of these two proteins in yeast[61]. Furthermore, in mammalian cells, it has also been shown that phosphorylation of the Ser65 residue of the transcription factor TFIIB recruits CstF to both the promoter and the 3'end formation site[207]. Interestingly, the transcriptional activator GAL4-VP16 was observed to stimulate co-transcriptional 3'end processing by recruiting the transcription elongation PAF1 complex to the DNA template *in vitro*[208]. An additional protein recruited at the promoter which is involved in 3'end formation is Ssu72, a component of the yeast cleavage/polyadenylation factor (CPF) complex. This protein works as a CTD Ser5 phosphatase during transcription initiation. Notably it has also been reported to be required for both 3'end processing and termination of RNA pol II transcription[55, 209]. Early recruitment of 3'end formation factors to the transcription complex, highlighted by the examples described above, may occur to allow the possibility of efficient 3'end formation at newly transcribed, functional poly(A) sites but the presence of these factors at both the very 5' and 3'end of genes may also be evidence, at least for some genes, of the presence of gene loops, which are formed across the ends of a gene via interaction of promoter and poly(A) site sequences[210]. This hypothesis would be mechanistically logical, linking 3'end formation to transcription re-initiation as it has recently been shown for the β-globin and HIV poly(A) sites as well as for knockdown of the 3'end processing factor Pcf11. This study shows that, *in vivo*, mutation of the β-globin and HIV poly(A) sites as well as knockdown of the 3'end processing factor Pcf11

decreased the occupancy rates of RNA pol II, TFIIB and CDK9 at the respective promoters[196].

Release of mature transcripts from the transcription complex also appears to be interconnected with pre-mRNA processing events. Pre-mRNAs that have been cleaved but not polyadenylated due to a PAP mutation, are retained at the transcription complex and are only released upon activation of exosome factors[211]. Interestingly, in a coupled *in vitro* system, it has also been shown that cleaved pre-mRNAs remain attached to RNA pol II via interaction with the cleavage and polyadenylation apparatus and are only released upon polyadenylation and, in some cases, concomitant splicing[212].

Notably, the 3'end processing event also strongly influences transcription. Near the 3'end of genes, immediately after the polymerization of the AAUAAA hexamer sequence, transcription appears to pause[213]. The decrease in transcription velocity seems to be independent of the DSE, the RNA pol II CTD and CstF, thus indicating that this form of pausing does not have a role in transcriptional termination[200, 213]. The arrest in elongation is believed to occur through a conformational change in CPSF caused by the recognition of the hexamer. The altered CPSF then acts on the body of RNA pol II, causing it to pause. This may allow time for the displacement of CPSF from the body of RNA pol II triggering the interaction of this 3'end processing factor with CTD-bound CstF[200].

Transcription termination is also intimately coupled with 3'end formation reactions[45, 64, 78]. Efficient termination depends on the presence of a functional polyadenylation signal in eukaryotes and both budding and fission yeast[214-216]. In *S. cerevisiae* the mutation of cleavage factors Rna14, Rna15 and Pcf11 disrupted

transcription termination in the *CYC1* gene, whereas mutated forms of polyadenylation factors had no such effect[205]. Removal of a functional poly(A) site from a mini-chromosome in *Xenopus laevis* oocytes, led to RNA pol II read around this circular DNA and failed to terminate[217]. Notably, when 3'end formation sequences were inserted efficient termination was restored. It was also observed that cleavage at the poly(A) site was not necessary for transcription to halt. Furthermore the 5'-3' exonuclease enzyme, which is a termination factor termed Rat1 in yeast and Xrn2 in mammals, has been shown to associate with the transcription machinery towards the 3'end of a gene like conventional 3'end formation factors[218, 219]. Rat1 binds to the CTD interacting domain protein Rtt103 although this interaction is not essential to its recruitment[219]. After 3'end cleavage Rat1/Xrn2 may be passed from the RNA pol II to the uncapped 5'end of the nascent RNA and upon the degradation of this transcript it might trigger the release of the polymerase from its DNA template concluding the transcription termination step[218, 219].

1.8. *Cis*-acting elements in mammals

3'end formation is directed by characteristic sequence elements (core poly(A) signals) within the transcript. In mammals these *cis*-acting elements are:

(1) the conserved hexanucleotide sequences AAUAAA or AUUAAA are present in ~70% of the human genes[220, 221] positioned 10-30 nucleotides upstream of the cleavage/polyadenylation site[82]. Several results have conclusively established that these hexanucleotide sequences are essential for cleavage and poly(A) tail polymerisation[82, 222]. The consensus sequence –AAUAAA- was initially revealed by a comparison of nucleotide sequences preceding poly(A) sites in

several mRNAs[223] and is one of the most highly conserved sequences known[224]. Although, it is now estimated that in ~30% of the human genes fully functional sequences that differ in 1 or more nucleotides from the conserved hexameric signal are present upstream of the poly(A) cleavage site[221, 225]. However, any single base substitution in the conserved sequence appears to strongly inhibit cleavage and polyadenylation[226-228]. The only highly functional exception found was the AUUAAA variant of the canonical sequence that is able to *in vitro* direct cleavage and polyadenylation with 66% and 77% efficiency respectively, relative to wild-type[227]. This variant is now considered as the second most common (~15%) poly(A) signal sequence[225]. Poly(A) sites that differ in 2 or more nucleotides from the canonical sequence are associated with alternative polyadenylation or with tissue-specific polyadenylation[82]. The hexamer-like sequences found in the near upstream region (NUE) in plants and the positioning element (PE) in yeast are often degenerated to little more than A-rich sequences[82, 229] in contrast with mammalian hexamer sequences which are generally highly intolerant to sequence alterations[226, 227].

(2) a more diffuse and less conserved U-rich and/or GU-rich downstream element (DSE) of variable length located up to 60 nucleotides downstream of the poly(A) hexamer[225]. The two main types that have been described are the U-rich element which is a short run of U residues[82, 230, 231] and the GU-rich element that presents the consensus sequence YGUGUUYY (Y= pyrimidine) downstream of the poly(A) site and was found in several genes[82, 232]. The poly(A) signal may have both elements acting synergistically[82, 231], have only one of the types[82, 230, 233] or none of them[82]. The proximity of a DSE to a poly(A) site can affect the cleavage site position and the efficiency of cleavage because the strength of a DSE is

dependent not only on its U and GU composition but also on its relative position to the cleavage site[82, 231, 233-235]. Point mutations and small deletions have, usually, weak effects on the DSE and large deletions are required to abolish function of these elements probably because of intrinsic redundancy[82].

(3) the cleavage and polyadenylation site is mainly determined by the distance between the upstream sequence AAUAAA and the DSE[82, 235]. The sequences around the cleavage and polyadenylation site are not conserved but an Adenosine is found at 70% of the cleavage sites of vertebrate mRNAs, the preference being A>U>C>G, the penultimate nucleotide is often an C residue (59% of analyzed genes) and therefore a CA dinucleotide often defines the cleavage and poly(A) addition site for several genes[82, 199].

(4) other auxiliary sequences were also found that are capable of modulating the efficiency of 3'end formation either in a positive or in a negative way[98, 236]:

Upstream positive elements

One type of enhancer sequence element is located upstream of the AAUAAA hexanucleotide, hence named upstream sequence element (USE), generally uridine rich but with no consensus sequence and is found primarily in viral poly(A) sites such as adenovirus[82], adeno-associated virus[237], human immunodeficiency virus type 1 (HIV-1)[238, 239], simian virus 40 (SV40) late[82] and human papillomavirus[240]. USEs have also been recently identified in human genes such as the lamin B2 gene and the complement factor C2 gene which are examples of cellular transcripts where USEs are critical in the modulation of the processing efficiency through polypyrimidine binding protein (PTB) or the 64kDa subunit of the cleavage stimulatory factor respectively (CstF-64)[241-243]. Further

examples are the cyclo-oxygenase 2 transcript which is alternatively polyadenylated and the use of the proximal poly(A) site is controlled by three short uridine-rich USEs[244] and an USE with several copies of the consensus sequence UGUAN that has recently been reported to alter the cleavage efficiency of both canonical and non-canonical poly(A) sites. An additional class of USEs was originally identified by their ability to promote the transport of intronless mRNAs and, subsequently, to promote polyadenylation as well[245] and since they are relevant in the context of the current dissertation they will be further discussed in later sections.

Downstream positive elements

Auxiliary enhancing sequences that are positioned downstream of the core poly(A) signal are less frequently reported due to the difficulty in establish definite boundaries for the core processing site DSE[82]. Despite this fact, a well characterized example comes from the identification of a 14 nucleotide G-rich region –GGGGGAGGUGUGGG- downstream of U-rich element of the SV40 late polyadenylation signal that has a positive influence on processing efficiency[246-248]. The GGGA tetranucleotide functions as a minimal protein recognition site for the members of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family[249]. *In vitro* experiments showed that hnRNP H/H' could interact with this sequence element and stimulate cleavage if the element is positioned immediately downstream of the core poly(A) site[246-248]. This type of stimulatory G-rich sequences positioned immediately adjacent of the core poly(A) signal have been predicted at the 3'ends of cellular genes[220, 248, 250]. Work with the human papillomavirus L1 early poly(A) site has also unravelled a regulatory sequence element consisting of six GGGU motifs positioned 174 nucleotides downstream of

the mentioned poly(A) site that regulates tissue- and stage-specific expression of the early and late viral genes[251]. These GGGU motifs were shown to interact with hnRNP H and this interaction proved to be essential for the efficient use of the early processing site[251]. The hnRNP H expression levels were determined in lower and upper level epithelium cells and inversely correlated with the degree of cell differentiation which is hypothesized to be associated with the alternative use of the early and late poly(A) site during the different stages of the viral infection[251]. Two G-rich regulatory elements located downstream of the core poly(A) site of the intronless human melanocortin receptor 1 gene were also shown to be critical for efficient 3'end processing[234] and a recent example as been reported for the p53 pre-mRNA[252]. Another G-rich sequence, the artificial MAZ4 site, was also shown to enhance poly(A) cleavage and transcription termination efficiency, *in vitro* and *in vivo*, by a yet unknown mechanism[67, 253, 254]. Further examples of these auxiliary sequence elements that are non G-rich are given by a pyrimidine rich sequence in the calcitonin-calcitonin gene related peptide pre-mRNA which stimulates poly(A) site use[255].

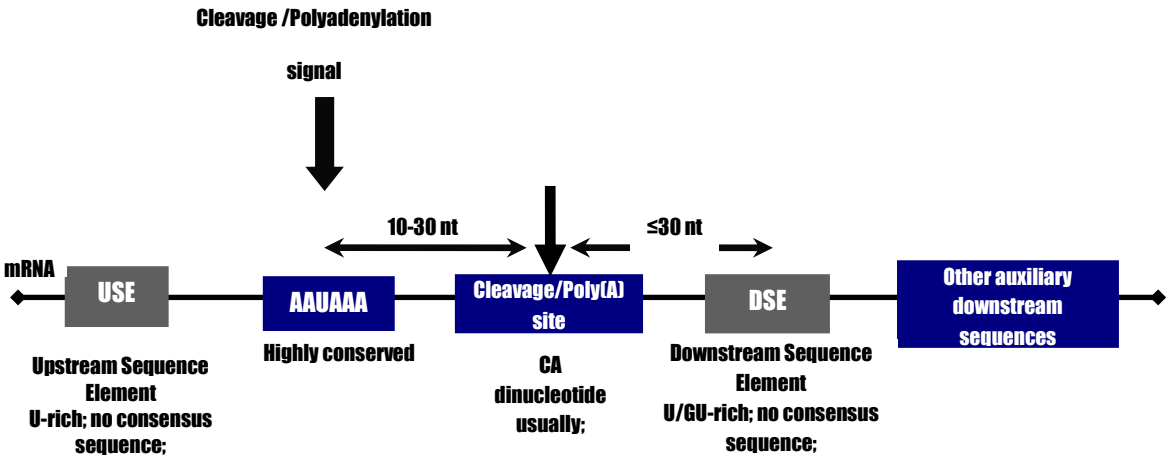


Figure 6: Schematic representation of a mammalian core poly(A) signal and auxiliary sequences

Upstream negative elements and Downstream negative elements

Negative regulatory elements were also found upstream of the U1A poly(A) site and of the human papillomavirus late poly(A) site[82, 256] and downstream of the promoter proximal HIV-1 poly(A) site[257]. A rather interesting observation is that the positioning of the SELEX identified, RNA pol II CTD consensus binding sequence –ACCCACACC- downstream of a core poly(A) signal results in a dramatic decrease of *in vitro* cleavage activity in a CTD-dependent manner which points that the CTD not only interacts with different proteins during the transcription cycle but interacts also with the transcript and this interaction seems to prevent 3' end formation and transcription termination[258].

Secondary structure of pre-mRNA

In addition to the sequences described, which are thought to work as recognition sites for factors that stabilize the polyadenylation complex, the secondary structure of sequences flanking the poly(A) site in the pre-mRNA should also be considered as an important feature of the 3'end formation mechanism[82]. The human leukaemia virus (HTLV-I) transcripts polyadenylation relies on the formation of a stem loop structure to bring the two distant core poly(A) sequences into a proximal position[259]. The formation of the TAR stem-loop in the HIV-1 transcripts allows an upstream positioned sequence enhancer to affect cleavage at a the core poly(A) site[260]. The secondary structure of the poly(A) site flanking sequences is also an important functional characteristic of poly(A) sites in HIV-1[261-264], SV40 late poly(A)[264, 265] and in the human, mouse and hamster secretory IGM poly(A) sites[266]. Further data supporting this aspect of 3'end formation is given by works which demonstrate that sequences chosen for optimal function as USEs by selection amplification were found to keep the

AAUAAA hexamer in an extensive open region[82, 267, 268] and similar functions are proposed for sequences beyond the DSE of several genes[247].

1.9. The cleavage and polyadenylation protein complex

The formation of mRNA 3'ends involves the recognition and interaction of the above described core sequences with a complex set of multimeric protein factors (Figure 7). Co-transcriptional cleavage and polyadenylation are directed by a large multi-protein complex which in humans *in vivo* can incorporate more than eighty proteins[269]. The key subunits of this complex are evolutionarily highly conserved. In mammals these are four multi-protein components: the cleavage and polyadenylation specificity factor (CPSF), the cleavage stimulation factor (CstF), cleavage factor I and II (CFI_m, $CFII_m$) and the poly(A) polymerase (PAP). Assembly of the 3'end processing complex is, in mammals, initiated by

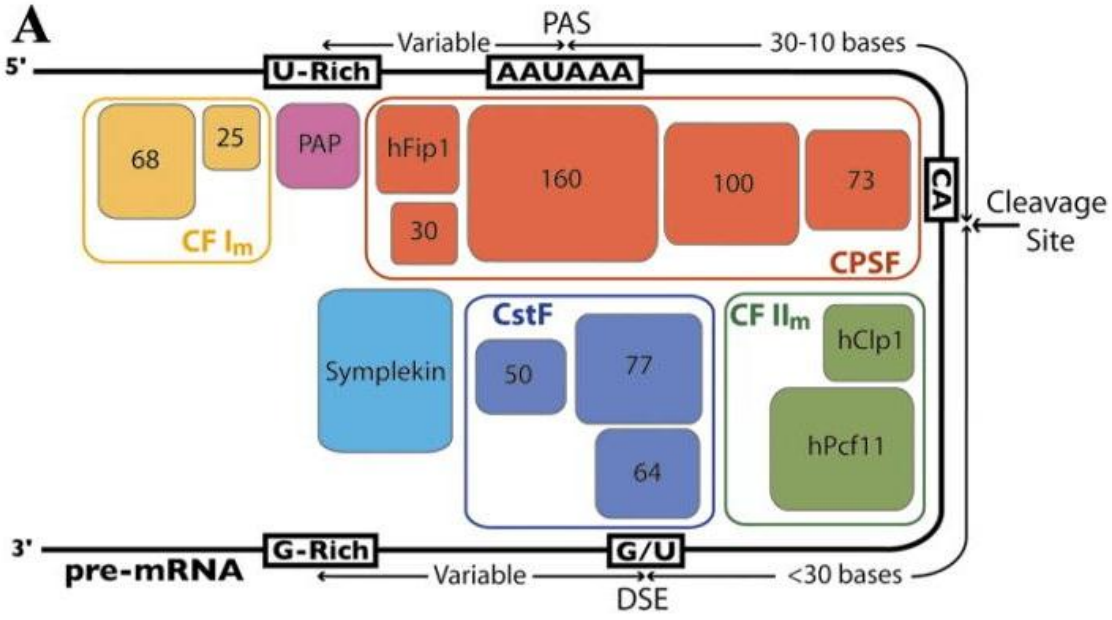


Figure 7: Schematic drawing of the pre-mRNA 3'-end processing complex in mammals. The cis elements in the pre-mRNA are also indicated. (From: Mandel, C.R., Bai, Y. & Tong, L. Protein factors in pre-mRNA 3'-end processing. *Cell Mol Life Sci* **65**, 1099-122 (2008))

cooperative interaction of CPSF and CstF with specific core sequences on the pre-mRNA[45, 220].

(a) Cleavage and Polyadenylation Specificity Factor (CPSF): required for both cleavage and polyadenylation reactions; consistent with these functions recognizes the AAUAAA hexamer essential for both reactions[45]. All six nucleotides of the hexamer are necessary for binding and RNAs as short as 10 nucleotides can be bound specifically to CPSF pointing out that the recognition of the sequence seems to be independent of any secondary structure[82]. Purified CPSF bind weakly to the hexamer but this interaction can be greatly enhanced by the cooperative interaction with CstF bound to the DSE[45, 198]. CPSF is a large protein complex composed of five different subunits CPSF-160 (160kDa), CPSF-100 (100kDa), CPSF-73 (73kDa), CPSF-30 (30kDa) and hFip1p (65-80kDa)[227, 270].

(a.1) CPSF-160 subunit interacts with the AAUAAA hexamer supporting the idea that this subunit is crucial for the recognition of this sequence, but the interaction of this subunit alone is less specific than the one observed with intact CPSF which suggests that the participation of the other subunits increases the specificity of the recognition and enhances the strength of the binding[45, 82]. The CPSF-160 subunit interacts specifically with the 77kDa subunit of CstF and with PAP[271] forming stable complexes on the RNA precursor[198].

(a.2) CPSF-100 and CPSF-73 subunits are thought to close contact the precursor and therefore increase the specificity and the strength of CPSF binding to the RNA[82]. Recent works indicate CPSF-73 subunit as the pre-mRNA 3'end processing endonuclease. CPSF-73 subunit shows strong, sequence-independent endonuclease activity that functions with specificity in 3'end processing of both histone and polyadenylated pre-mRNAs[272-275]. The crystal

structure of the CPSF-73 subunit shows the presence of two zinc ions at the active site and the cleavage activity of the whole complex is dependent on the presence of zinc ions[272-274]. Mutations which disrupt zinc binding at the active site abolish the endonuclease activity[273, 274].

(a.3) CPSF-30 subunit has 6 nucleic acid binding motifs and its probable function is to cooperate with CPSF-160 in the recognition of RNA substrates and, through interactions with PABII, stabilize the polyadenylation complex[82].

(a.4) hFip1p subunit interacts with PAP and has an arginine-binding motif that preferentially binds U-rich sequence elements on the mRNA contributing to the CPSF-mediated stimulation of the PAP activity[270]. hFip1 runs as a diffuse band in SDS-PAGE, with apparent molecular masses between 65 and 80 kDa[197].

(b) Cleavage stimulatory Factor (CstF): heterotrimeric protein with subunits of 77kDa (CstF-77), 64kDa (CstF-64) and 50kDa (CstF-50)[276, 277]. CstF interacts with the DSE through CstF-64[278] and is necessary for cleavage but not for polyadenylation[279, 280] although it can stimulate poly(A) addition on substrates with a CstF binding site upstream of the AAUAAA hexanucleotide[241]. The binding of CstF to the DSE greatly enhances the affinity of CPSF for the AAUAAA and vice versa[276].

(b.1) CstF-77 subunit presents several repeated motifs predicted to mediate protein-protein interactions consistent with the fact of CstF-77 being the middle subunit bridging CstF-64 and CstF-50 with the three arranged in a linear way[82, 281]. The CstF-77 subunit interacts with the CPSF-160 subunit probably contributing to the mutual stabilization of the CPSF-Cstf-RNA complex[271].

(b.2) Cstf-64 subunit contains a RNA-binding domain (RBD) with an RNA recognition motif (RRM) which recognizes U/GU-rich sequences and strongly binds double uridine residues within this sequence[278, 282, 283].

(b.3) CstF-50 subunit is required for cleavage contains several transducin repeated motifs which have been shown to mediate protein-protein interactions in other proteins[82].

(c) Cleavage factors I_m and II_m (CF I_m and CF II_m): multisubunit complexes essential to direct the cleavage of pre-mRNA[45]. Upon dephosphorylation of either CF I_m or CF II_m , 3'end cleavage is inhibited and therefore both proteins appear to be essential for 3'end processing[284].

(c.1) CF I_m was purified to near homogeneity and is an RNA-binding factor consisting of three major polypeptides of 25kDa (CF I_m-25), 59kDa (CF I_m-59), 68kDa (CF I_m-68) and possibly a fourth one of 72kDa (CF I_m-72) which increases the stability of the CPSF-RNA complex thus suggesting that it may also interact with CPSF and contribute for the overall stability of the 3'end formation complex[82, 285, 286]. CF I_m-25 has no known motifs while CF I_m-68 has a domain organization strongly reminiscent of that from spliceosomal SR proteins[82, 285] which would agree with the involvement in the regulation of poly(A) site selection[64, 287]. Interestingly recombinant CF I_m-25 and CF I_m-68 can be assembled *in vitro* and can replace purified CF I_m in cleavage assays which taken together with the fact that preliminary studies in the sequence of CF I_m-59 show great similarity with CF I_m-68 suggest that either CF I_m-59 is a degradation product of CF I_m-68 or that CF I_m exists as heterodimers of CF I_m-25-68 or CF I_m-25-59 [41,130]. The interaction of CF I_m with the RNA substrate is thought to be an early

step in the assembly of the 3'end formation complex which facilitates the recruitment of the other factors[82, 285].

(c.2) CF II_m can be purified in two fractions CF IIA_m and CF IIB_m of which only one fraction, CF IIA_m, is essential for the cleavage reaction. In humans consists of two subunits, hPcf11 and hClp1. One of the components of CF IIA_m, the hClp1 component, has been shown to interact with CPSF and CF I_m suggesting that it bridges these two 3' end processing factors within the cleavage complex. The CF IIB_m fraction has a stimulatory effect in the cleavage reaction[45, 288].

(d) Poly(A)-polymerase (PAP): consists of a single polypeptide but exists in several alternatively-spliced forms[289-291] and catalyzes the addition of adenosine residues to the 3' ends of pre-mRNAs[82]. PAP is specific for the utilization of ATP[290] but has no specificity for the RNA substrate[82]. The C-terminal domain of PAP interacts with hFip1 and CPSF-160[270]. PAP is recruited to the processing complex through interactions with CPSF-160[271]. PAP is required for the cleavage of most mammalian pre-mRNAs[82].

(e) Poly(A)-binding protein (PABPN1): is a 33kDa protein[82]. Although CPSF and PAP are sufficient for poly(A) addition to a precleaved RNA substrate, the rapid elongation and control of the poly(A) tail length (limited to ~200 residues) requires the presence of PABPN1[292, 293]. Once a short poly(A) tail has been polymerized by PAP, PABPN1 binds to this tail and forms a quaternary complex with CPSF, PAP and the RNA substrate which transiently stabilizes the binding of PAP to the 3'end of the RNA and supports processive synthesis of a long poly(A) tail in a single rapid step[292-294].

(f) Symplekin is an additional factor that can be found as part of a large complex containing CPSF and CstF and its thought to be involved in the assembly of the polyadenylation machinery[295]. Symplekin has been shown to interact with the heat shock transcription factor 1 (HSF1) and this interaction functions as a mechanism to enhance the recruitment of polyadenylation factors to the heat shock protein (HSP) gene[296].

Additionally, over eighty proteins are implicated in 3'end processing highlighting the complexity of the 3'end formation reaction[199, 236].

1.10. Mechanism of 3'end formation: Assembly, Cleavage and Polyadenylation

The initial step of the assembly of a functional cleavage/polyadenylation complex is, probably, the recognition of the sequence signals AAUAAA and DSE on the precursor mRNA by CPSF and CstF respectively in a process assisted by CF I_m[64, 82]. CPSF, which is essential for both cleavage and polyadenylation, binds to the AAUAAA hexamer through its CPSF-160 subunit assisted by CPSF-30 and possibly by CPSF-100[82, 271]. CstF, which is required only for the cleavage step, binds to double uridine residues[283] in the U/GU rich region of the DSE through its RNA recognition motif (RRM) located in the CstF-64 subunit[282] (Figure 7). Both individual interactions are weak and are stabilized by cross factor interaction of CPSF-160 and CstF-77[82]. The final component of this initial complex is RNA pol II whose CTD is necessary for efficient splicing and polyadenylation *in vivo*[78]. CPSF-CstF interactions define the precise region where the cleavage site must lie and

the formation of a cleavage-competent complex requires the additional recruitment of CF II_m and PAP with the later probably interacting with CPSF-160 at this point[82]. Once this complex is formed cleavage occurs co-transcriptionally, preferentially at a CA dinucleotide[235] located 10-30 nucleotides downstream of the hexamer and probably catalyzed by the CPSF-73 subunit[273]. Cleavage is probably followed by several significant rearrangements where degradation of the 3' cleavage product and dissociation of CstF, RNApol II, CFI_m and $CFII_m$ occurs[82]. The polyadenylation step is then initiated by the recruitment of PAP to the AAUAAA-containing substrate through its interaction with CPSF-160 and this complex proceeds to slowly polymerize an adenosine tract of ~10 nucleotides that creates the PABPN1 binding site[292, 293]. Once PABPN1 is bound, the interactions between CPSF, PAP and PABPN1, cooperatively stimulate poly(A) polymerase such that a complete poly(A) tail is synthesized in one processive event, which terminates at a length of approximately 250 nucleotides[293, 297]. Stimulation by CPSF is disrupted when the poly(A) tail reaches a length of approximately 250 nucleotides, and this terminates processive elongation. PABPN1 measures the length of the tail and is responsible for disrupting the CPSF-poly(A) polymerase interaction[293].

1.11. Alternative polyadenylation (APA)

Over half of human genes encode multiple mRNA isoforms corresponding to transcripts with 3'UTRs of differential lengths[40, 298]. Interestingly, this differential 3'UTR length has been shown to influence mRNA stability and translation by providing an alternative miRNA binding landscape[299]. Analyses of APA at the genome scale revealed a widespread biological role in humans, mice, worms, yeast, plants and algae with genes encoding multiple transcripts ranging from 10-

15% in *S. cerevisiae* to ~54% in humans[298, 300, 301]. The number of 3'ends mapped for orthologous genes with alternative p(A) sites shows a high degree of similarity between mouse and humans indicating that these have been actively selected in evolution[298, 302]. The majority of tissue-specific and non-canonical poly(A) sites does not show conservation[302]. Further bioinformatics analysis also showed that while the canonical AAUAAA sequence predominates in genes with a single poly(A) site, the less conserved variants occur frequently in genes with multiple poly(A) sites, generally in a promoter proximal position, whereas the canonical sequence tends to appear downstream of the variant sites[221, 298]. Selection of alternative poly(A) sites characterized by suboptimal hexamer variants is, at least in part, defined by the presence of sequence motifs able to compensate for the absence of a consensus hexanucleotide such as strong CstF binding sites with higher U and GU content[303].

Cell growth, differentiation and development are physiological conditions which have a strong impact on differential processing at multiple poly(A) sites[301]. Analysis of 42 human tissues revealed a considerable degree of APA with tissues from retina, placenta, blood and ovary showing increased likelihood of using proximal poly(A) sites and samples from bone marrow, uterus, brain and nervous system showing increased usage of distal sites[141, 304]. Furthermore, genome-wide analysis of APA suggests a pattern that correlates proliferation and differentiation states of cells with transcript 3'UTR length. States of increased proliferation, dedifferentiation and disease seem to be associated with general shortening of 3'UTRs while late developmental stages and cell differentiation states tend to present longer 3'UTRs[141, 305-309]. Additionally, cancer cells, which constitute an highly relevant subset of proliferative cells, display, when compared to normal

(pro-B) cells, a generalized pattern of 3'UTR shortening (although some elongated 3'UTR transcripts have also been identified) concomitant with upregulation of a number of mRNAs encoding 3'end processing factors with the most pronounced differences at the level of CPSF160 and CstF64 subunits[306]. Further data shows upregulation of CstF and CPSF subunits, RBBP6[236] and symplekin during the generation of iPS (induced pluripotent stem) cells correlating with general 3'UTR shortening and downregulation of the same factors in differentiated embryonic tissues where longer 3'UTRs were observed[308]. Finally, a classical example of alternative 3'end processing is the regulation of immunoglobulin M heavy chain (IgM H-chain) synthesis during B cell differentiation in the DT40 chicken B cell line. Early in differentiation, pre-B and B cells produce membrane-bound form mRNA by removing the upstream poly(A) site by splicing and by using the downstream poly(A) site. In contrast, at the final stage of differentiation, plasma cells primarily synthesize large amounts of secreted form mRNA by use of the upstream poly(A) site. The concentration of CstF-64 subunit increases during activation of B cells, and this is sufficient to switch IgM heavy chain mRNA expression from the membrane-bound form to the secreted form[310, 311]. Taken together, all these observations suggest a model where changes in expression and/or stoichiometry of 3'end processing factors result in alternative poly(A) site selection through, for example, increased recognition of suboptimal poly(A) sites[301]. Further levels of complexity and integration of APA regulation are also suggested by observations that, both in yeast and in humans, the 3'region near the poly(A) site shows a low nucleosome density[312-314] although this might be an microccocal nuclease related artifact .

Observation that 3'UTRs harbour miRNA target sites and/or other regulatory sequences such as AU-rich elements and shortening of 3'UTR length directly correlates with increased levels of the protein generated from the shortened mRNA transcript[305, 306, 315, 316] strongly suggest that APA has a fundamental role in the definition of variant mRNA transcripts stability and localization and ultimately in protein expression levels[301]. Usage of alternative poly(A) sites, like splicing and alternative splicing, confers to a particular gene further degrees of plasticity. This gives the possibility of enhanced regulation of the pool of existing transcripts in response to a wider variety of stimulus and physiological situations. Combination of transcriptome diversity and proteome diversity expands the range of biological possibilities available to a specific organism.

1.12. Interconnections between pre-mRNA processing reactions

After discovery and identification of the pre-mRNA processing reactions it soon became evident that all three events are interconnected and are capable of mutually enhancing each other[45].

1.12.1. Interactions between capping and splicing

Experiments conducted *in vitro* with HeLa cells nuclear extract revealed that replacement of the standard m^7GpppN transcript cap with the cap analogue m^7GDP led to a significant disruption of splicing[317]. Further support to these observations came from *in vitro* experiments with mammalian cell extract depleted of the CBP80 subunit of CBC where splicing failed to occur[95] and from results obtained with fission yeast cell extract where it was shown that the CBC is able to

enhance the binding of the U1 snRNP to a 5' splice site[318]. Interestingly, these observations were limited to the splicing of a cap-proximal intron and were not observed in downstream introns in a mammalian *in vitro* splicing system suggesting that the stimulatory action of capping on splicing is spatially limited[319]. The stimulation of intron excision via the CBC seems not to be restricted to higher eukaryotes. When the guanylyltransferase activity, which is required for efficient pre-mRNA capping, was inactivated in budding yeast cells, splicing of pre-mRNAs was strongly inhibited[320]. Recent experiments on *SUS1* pre-mRNA, which is a spliced transcript encoding a protein required for histone ubiquitination in budding yeast, and the MADS box transcription factor Flowering Locus C gene in *Arabidopsis* also show that, at least for some transcripts, precise and efficient pre-mRNA splicing requires the presence of CBC components[321, 322]. Interestingly, the CBC has also been suggested to positively influence alternative splicing by facilitating the binding of P-TEFb and SF2/ASF to a splicing minigene in HeLa cells[102].

1.12.2. Interactions between capping and 3'end processing

The interaction between capping and 3'end formation was first observed in early SV40 viral expression. *In vitro,* addition of the m7GpppG cap sequence to the 5'end of a SV40 minigene was observed to enhance cleavage and polyadenylation of the corresponding transcripts[323]. Furthermore, *in vitro*, an increased addition of m7GpppG inhibited pre-mRNA 3'end cleavage thus indicating that out-competition of CBC from the pre-mRNA cap sequence can repress 3'end formation[324]. Interestingly, it appears that only cleavage and not polyadenylation is affected by the depletion of the cap structure[324].

1.12.3. Interactions between splicing and 3'end processing

Since the open reading frames (ORFs) of most eukaryotic protein encoding genes are interrupted by introns, the synthesis of a contiguous uninterrupted and translatable mRNA requires that these intervening non-coding sequences are precisely excised by the splicing machinery. The excision of these introns is not an isolated mechanism and it has become increasingly clear in recent years that splicing not only produces an mRNA with an uninterrupted ORF but also influences the efficiency of transcription, translation, RNA localization and RNA degradation[45, 64, 126, 325, 326]. In addition to this, splicing has long been known to influence the efficiency of processing with this interconnection resulting in a splicing dependent increase in gene expression[326]. Excision of terminal introns from pre-mRNAs has an essential stimulatory role in the poly(A) cleavage reaction of intron-containing genes and expression constructs subject to artificial removal of intronic sequences show, in general, a dramatic reduction in cleavage efficiency at poly(A) sites[327, 328]. Interestingly, the inhibition of 3'end formation via the absence of capping was lifted by insertion of a functional 3' splice site and polypyrimidine tract upstream of the poly(A) site[324]. The explanation for the molecular mechanism underlying these observations has become clearer with the observation that components of the splicing machinery can directly interact with proteins of the cleavage and polyadenylation complex. The splicing factor U2AF plays a crucial role in this process since its 65kDa subunit was shown to directly contact the Poly(A)-polymerase and influence cleavage efficiency *in vivo* and *in vitro*[329-331]. The splicing factor SRm160 has also been reported to interact with the cleavage and polyadenylation specificity factor (CPSF) and stimulate cleavage and polyadenylation[332, 333], the U1 snRNP-specific protein U1A has been linked

with CPSF-160[334] and it was also shown that direct interactions between CPSF subunits and the U2snRNP affect both splicing and 3'end processing[335]. Importantly, not all splicing factors affect 3'end processing in a positive way since there are reported examples where the proximal presence of a 5' splice site and the associated U1 snRNP can inhibit cleavage and/or polyadenylation[336, 337]. Further support to these observations has recently been uncovered in a genome-wide study where knockdown of U1 snRNA via addition of antisense morpholino oligonucleotides led to premature 3'end formation at a cryptic poly(A) site in HeLa cells, thus highlighting the function of the U1 snRNP in preventing premature pre-mRNA 3'end processing at cryptic, intronic poly(A) sites[338].

Notably, in *S. pombe*, the unpolyadenylated 3'ends of the telomerase RNA gene *TER1* were precisely mapped to terminal 5' splice site of the gene, and thus appear to be formed by partial *cis*-splicing[339]. The functional relevance and implications of this observation remain to be explored but, for this particular gene, the spliceosome seems to be responsible for 3'end formation.

Both stimulatory and inhibitory examples presented above highlight the close connection between splicing and 3'end processing and raise the question of how do intronless genes bypass the absence of the splicing effects on the 3'end processing of their pre-mRNAs.

1.13. Intronless genes: how to bypass the absence of splicing

The percentage of intronless genes in the human genome is thought to be around 5%[340] and the proteins encoded by this type of gene are responsible for diverse and crucial functions since they range from interferon-alpha-coding genes, histone-coding genes, oncogenes or transcription factors to transmembrane

protein coding genes. Information on the efficient expression of human intronless genes that are not subject to the effects of intron splicing is relatively scarce mainly due to the fact that most of our knowledge about the expression of uninterrupted genes comes from studies on the expression of intronless viral transcripts[341-345]. An interesting image begins to emerge from the work available pointing towards the assumption that the absence of the splicing effect in the expression of intronless genes is, in general, bypassed through the inclusion of specific *cis*-acting sequences, such as upstream sequence elements (USEs) in the mRNA, often associated with enhanced expression of intronless transcripts. These elements facilitate the processing and export of the transcripts enabling efficient intron-independent gene expression through binding alternative cellular factor(s)[341-349] . Examples of this type of sequences and mechanics come from observations like the Herpes simplex virus type 1 (HSV-1) intronless thymidine kinase pre-mRNA that revealed the presence of a specific pre-mRNA processing enhancer (PPE) which stimulates 3'end formation and possibly nuclear cytoplasmic export, through a molecular mechanism that seems to be dependent on the interaction between the cellular heterogeneous nuclear ribonucleoprotein (hnRNP) L and the PPE[341, 346, 349]. *Cis* post-transcriptional regulatory elements (PREs) were also described in Hepatitis B viral transcripts[342, 343] and the Woodchuck Hepatitis RNAs[344] that prove to have similar functions. Furthermore, the polyadenylated but non-protein encoding nuclear retained RNA from the Kaposi sarcoma-associated herpes virus also revealed an analogous element that is able to confer efficient expression to a intronless globin mRNA by enhancing 3'end processing[345] and a cellular functional homologue of the above described viral sequences was found in the intronless mouse Histone H2a gene[245, 348]. In the

latter case the interaction between this sequence and the splicing factors 9G8 and Srp20 was shown to be determinant in the stimulation of 3'end processing and nuclear cytoplasmic export[347]. Recently the intronless human *c-jun* gene also revealed to contain an element which may enhance 3'end formation and cytoplasmic accumulation by a yet unknown mechanism[346]. Further information came from the analysis of U2AF mutants in Drosophila that after treatment with siRNA against U2AF showed expected nuclear accumulation of unspliced pre-mRNA but additionally also showed unexpected retainment of a significant number of intronless gene transcripts in the nucleus[350]. The fact that in purified nuclear ribonucleoprotein complexes the *Drosophila* large subunit of U2AF (dU2AF^{50}) is found to be associated with a large number of intronless transcripts and that the vast majority of these intronless genes possess putative U2AF binding sites led to the proposal that U2AF affects the export of both intron containing and intron-free mRNAs by possibly recruiting mRNA export factors such as UAP56 and NXF1[350]. A question that remains to be determined is whether the effect of U2AF depletion is solely at the level of nuclear cytoplasmic export or if the lack of U2AF affects the expression of these genes at the cleavage and polyadenylation step given the close relationship between U2AF and processing.

The multiple examples presented above seem to indicate that, in intronless genes, *cis*-acting enhancing elements within the pre-mRNA assume an essential role in the stimulation of efficient gene expression bypassing the absence of a splicing mechanism and allowing maximal expression of either intron-free viral or intron-free cellular transcripts.

1.14. Aim of the study

The aim of the present work is to explore and clarify if previously unidentified 3'end formation regulatory sequence elements, such as the viral elements (PPE, PRE) described in the previous section, are present within the 3'UTR and 3'Flank sequences of human intronless genes that compensate both for the lack of a defined DSE and/or the lack of stimulation by terminal intron removal which, in spliced genes, has a strong positive effect in 3'end formation efficiency.

Chapter 2

Materials and Methods

2. Materials and Methods

2.1. Bacterial Strains

E. coli. DH5α competent cells were used for plasmid cloning. *E. coli* XL1-Blue competent cells were used for TA-vector cloning.

2.1.1. Culture and maintenance of bacterial strains

The strains described above were grown in LB media (10 g/l bactotryptone, 5 g/l yeast extract, 10 g/l NaCl, pH 7.5) at 37°C in a shaking incubator. Where ampicillin was required, it was added to autoclaved media to a final concentration of 50-100µg/ml.

2.1.2. Preparation of chemically competent *E. coli* cells

10 ml of overnight DH5α culture was used to inoculate 500 ml ψ broth (20 mM MgSO4, 10 mM NaCl, 5mM KCl, 2% tryptone, 0.5% yeast extract, pH adjusted to 7.6 with KOH), being subsequently grown at 37°C until an OD of 0.5 at UV_{600nm} was reached. The culture was chilled on ice for 10 min and spun down at 4000 g for 10 min. The cell pellets were then re-suspended in 100 ml ice cold TFB-1 (30 mM $KC_2H_3O_2$, 100 mM RbCl, 10 mM $CaCl_2$. 50 mM $MnCl_2$, 15% glycerol, pH adjusted to 5.8 with acetic acid) and incubated on ice for 30 min. After this incubation, a second centrifugation at 4000 g for 10 min was conducted and the pelleted cells were re-suspended in ice cold TFB-2 (10 mM MOPS, 75 mM $CaCl_2$, 10 mM RbCl, 15% glycerol, pH adjusted to 6.8 with NaOH), incubated on ice for 15 min, aliquoted into 150µl aliquots and snap frozen with liquid nitrogen. Competent cells were subsequently stored at -70°C.

2.2. DNA manipulation and analysis

2.2.1. Polymerase chain reaction

DNA sequences used in these work were amplified via PCR. Primers were designed using Vector NTI® software (Invitrogen, 2005) and were synthesised by Sigma or MWG.

Taq polymerase (NEB) was used to perform the PCR, however, when proof-reading was required, the *Pfu* polymerase (Fermentas) was used. PCR conducted with *Taq* or *Pfu* polymerase was executed at 94°C for 4 minutes (min) and then 94°C for 1 min (denaturation), 50-60°C for 1 min (annealing) and 72°C for 1 min (extension) for 30 cycles. PCR reactions were prepared containing 200 µM of each dNTP, 10x *Taq* DNA polymerase buffer (NEB: 100 mM Tris-HCl pH 8.4, 500 mM KCl, 15 mM MgCl2), 1.5U of *Taq* or *Pfu* polymerase and 1 µM of each primer. The annealing temperature was altered according to the melting temperatures of the primers used. All PCRs were performed using a Biometra T3000 Thermocycler.

2.2.2. DNA sequencing

DNA sequencing was carried out by Geneservice at the Department of Biochemistry at the University of Oxford. Sequences were aligned and analysed using the Vector NTI® software (Invitrogen, 2005).

2.2.3. Agarose gel electrophoresis

DNA fragments and plasmids were run on agarose gels for visualization. Gels were prepared by melting 0.8-2% w/v of agarose (Bioline) in 1x TBE buffer (100 mM Tris, 100 mM boric acid and 1 mM EDTA) and by adding 0.5 µg/ml

ethidium bromide. Samples were mixed with xylene cyanol/bromophenol blue loading dye prior to loading. Mix Fast Ruler DNA ladder (Fermentas) was loaded as a marker. Gels were run in 1x TBE buffer. DNA fragments were visualized under UV_{320nm} light.

2.3. Plasmids – design, description and cloning

2.3.1. TA vector preparation

The procedure for the construction of a TA vector is described in Marchuk[351]. 10-20µg of pUC18 vector was digested with SmaI (NEB) at 25°C for 2 hours (h) in a reaction mix containing 20 mM Tris-acetate, 50mM $KC_2H_3O_2$, 10mM $Mg(C_2H_3O_2)_2$ and 1 mM dithiothreitol (DTT) and subsequently inactivated at 65°C for 15 min. The digested plasmids were phenol/chloroform extracted, ethanol precipitated and subsequently incubated with *Taq* polymerase (NEB) in a reaction containing 20mM Tris-HCl pH8.4, 50mM KCl, 1.5 mM $MgCl_2$, 2mM dTTP at 72°C for 3 h. Finally plasmids were phenol/chloroform extracted, ethanol precipitated, re-suspended in H_2O and stored at -20°C.

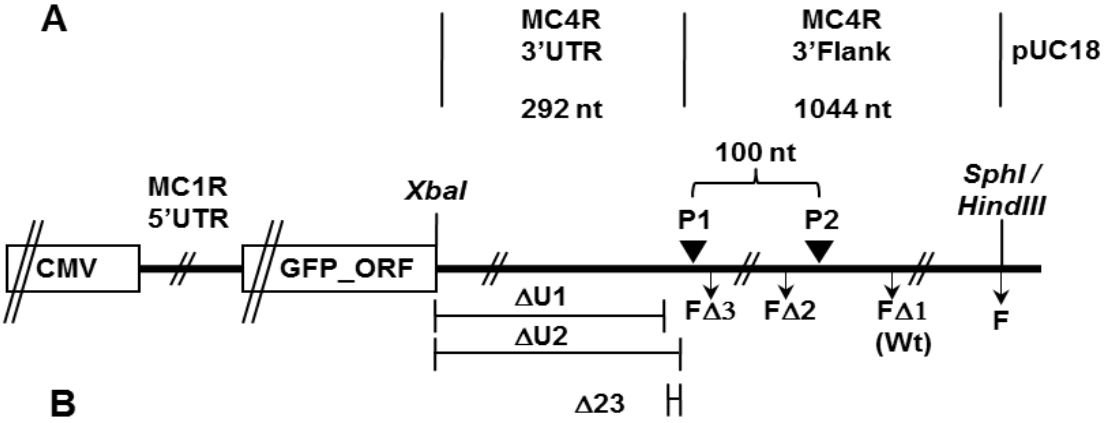
2.3.2. Reporter plasmid construction – MC4R, JUNB, JUND and EDF1

MC4R reporter gene constructs

The MC4R reporter constructs retain the Cytomegalovirus (CMV) promoter, the MC1R 5'UTR and the GFP ORF of the MC1R-pUC18 reporter described in[234]. MC4R 3'UTR and flanking sequences, amplified from HeLa cells genomic DNA, replace the 3'UTR and flanking regions of the MC1R reporter and were inserted into the original MC1R plasmid using *XbaI-SphI*, *XbaI-HindIII* or *HindIII-HindIII* restriction sites depending on the specific constructs.

(1) MC4R Wt plasmid, 3'Flank and 3'UTR deletion clones (primer sequences: Appendix, Table 1)

The 3'Flank deletion clones F through FΔ3 (Figure 8) were made by *Pfu*



atggggacagagcacgcaatataggaacatgcataagagacttttcactcttaccctacctgaatattgtacttctgcaac
agctttctcttccgTGTAGggtactggttgagaatatccattgTGTAAatttaagcctatgatttttaatgagaaaaaatg
cccagtctcTGTATtatttccaatgtcatgctacttttttggccataaaatatgaatctatgttataggtTGTAGgcactgtg
gatttacaaaaagaaaagtcctt**ATTAAA**agcttaacaatgtctc**C**ttcgtgttattcataagcattggacactttgcgtgc
tttcgtaacatagaaatcagagcctc**ATTAAA**catattctaataaatgctttattatattatattccaccattgaaatgtag
agagttcattctagcagttaagggaaaatatttaaagaatagatgtataatcatttttaaaaatatcatcactgaaatttcaagt
aatttaattaaaggtttgttcatcctccctgtgcagaagtagaaatgaagctcctctagagagaaaaacaggtgctgaaaa
aagagagatgctgagttagaaaactatgtgtacttctcagacacagaaagattcatctcccttagcaaacagcctgatattt
cacacaggcaattcggtttggggtgggtgggaaaaggcagctatagggatgaaagtgtcctggaaaacctaagttgcca
tgtcaagctaaatttcctgttaccaaattatccctaaacaatattaataaaacagagaatggggatttatgccccctttaaaca
aggattcagggtctgctgaagtgtccagaaaaggtaatattgtactgagagatttagcttcctgtttctagagaagtacagga
ctgacatcgaattctttctcatttcacatgaaaatgacttgctgtttacttattgctgatttttgaactggagtacaaatgttgattat
caaggattttaaacttcaaaaccaaattcatagcagaagtgttttctggcaatatttgtaaataactaaaataaaggtatctgt
gatgctggtacattttaagcaggaagggaggcacctaccaaaccccagaggatgtcatgcaataaactagccttccaca
gccacttcttgaagagtaaaatgttttgctccaaaataatcctacagatttggccactgagcaataaaatttgtactgtactttct
aatctgtgtgtaatttaaaataaagcaaaagaatatattaggttgaggattaccagtaaactatcttaggatttacagggtcttc
aaat

Figure 8: (A) Schematic representation of the MC4R reporter plasmid. CMV promoter and GFP_ORF are shown as open boxes. Lines across indicate regions not drawn to scale. Nucleotide lengths of cloned MC4R regions are indicated. MC4R poly(A) sites P1 and P2 are filled triangles. Main restriction enzyme sites used for cloning are depicted. Vertical arrows indicate the end of the deletion clones relative to Wt sequence. 3'UTR deletion clones (ΔU1: deletion of the first 246nt of the 3'UTR; ΔU2: deletion of the first 269nt of the 3'UTR; Δ23: deletion of the 23nt immediately upstream of the P1 hexamer). 3'Flank deletion clones (F: retains the first 1044nt of 3'flank sequences; FΔ1: retains the first 250nt of 3'flank sequences; FΔ2: retains the first 80nt of 3'flank sequences; FΔ3: retains the first 25nt of 3'flank sequences). **(B)** MC4R 3'UTR (underlined) and 3'flank nucleotide sequences. P1 hexamer and main cleavage site are shown in bold uppercase and underlined. P2 hexamer is shown in bold uppercase. TGTAN motifs present in MC4R 3'UTR are shown in uppercase and underlined.

amplification using the forward primer 4XBA paired with one of the reverse primers ENDS4, SPH1U, NS2, and FLD1 respectively. Truncated *HindIII-HindIII* (F clone) or *XbaI-SphI* (FΔ1, FΔ2 and FΔ3) amplified fragments were religated into *HindIII-HindIII* or *XbaI-SphI* digested MC4R respectively. After analysis, the MC4R FΔ1 clone was defined as the Wt (Results 3.2.4.) and is subsequently referred to as the MC4R Wt plasmid. Unless otherwise stated all subsequent clones have been engineered from this plasmid.

The 3'UTR deletion clones ΔU1 and ΔU2 were made by *Pfu* amplification of the MC4R 3'UTR using the primer pair UTRD1-SPH1U and UTRD2-SPH1U respectively. Truncated *XbaI-SphI* amplified fragments were religated into *XbaI-SphI* digested MC4R Wt plasmids. Δ23 clone was constructed by *Pfu* amplification with phosphorylated internal primers placed immediately next to each other resulting in precise deletion of the 23nt upstream of MC4R P1. The resulting two fragments were then ligated together and subsequently re-amplified using the external primer pair containing 5'-*XbaI* and *SphI*-3' restriction sites. The primers used: 4XBA-D23WtR and D23F-SPH1U. Truncated *XbaI-SphI* amplified fragments were religated into *XbaI-SphI* digested MC4R Wt plasmids.

(2) MC4R core hexamers and DSE mutation clones (primer sequences: Appendix, Table 1)

Mutations targeting core hexamers and DSE sequences were introduced by PCR amplification with specific primers using the MC4R Wt plasmid as a template and cloned into *XbaI-SphI*, *XbaI-HindIII* or *HindIII-HindIII* Wt digested vectors.

The H1h2, h1H2, h1h2 and h1h2* clones were constructed by *Pfu* amplification with phosphorylated internal primers placed immediately next to each other, resulting in no loss of sequence, save the T to C (H1h2), the G to C (h1H2),

both combined (h1h2) or both combined plus A to C (h1h2*) changes (Figure 17A, Results 3.2.4. for sequences) encoded by the primers themselves. Resulting fragments were then ligated together and subsequently re-amplified using the external primer pair containing 5'-*XbaI* and *SphI*-3' restriction sites. The primers used: for H1h2, 4XBA-HEXU1 and HEXD2-SPH1U; for h1H2, 4XBA-HEXU2 and HEXD1-SPH1U; for h1h2, 4XBA-HEXU2 and HEXD2-SPH1U; for h1h2*, 4XBA-HEXU2 and D22F-SPH1U. The resulting fragments were subsequently cloned into *XbaI-SphI* digested MC4R Wt plasmids.

The DSE mutation d2 and d4 clones were constructed by *Pfu* amplification with phosphorylated internal primers placed immediately next to each other, resulting in no loss of sequence, save the U to C changes encoded by the primers. The resulting two fragments were then ligated together and subsequently re-amplified using the external primer pair containing 5'-*XbaI* and *SphI*-3' restriction sites. The primers used: for d2, 4XBA-UU2 and UD1-SPH1U; for d4**,** 4XBA-UU3 and UD2-SPH1U. The resulting fragments were subsequently cloned into *XbaI-SphI* digested MC4R Wt plasmids. H1h2-d4, h1H2-d4 and h1h2-d4 clones were created by combination of the *HindIII-HindIII* digested d4 flank sequence with plasmids containing the respective hexamer sequences.

(3) MC4R hexamer A-rich upstream sequences mutation clones (primer sequences: Appendix, Table 1)

The 5GC, 6GCh1, 6GCh2, 7GC, 7GG clones (Figure 19A, Results 3.2.4. for sequences) were constructed by *Pfu* amplification from the Wt plasmid with phosphorylated internal primers placed immediately next to each other, resulting in no loss of sequence, save the T to C, A to C and G to C (5GC, 6GCh1, 6GCh2, 7GC) or the T to C, A to G, G to C and A to C (7GG) changes encoded by the

primers. The resulting fragments were ligated together and subsequently re-amplified using the external primer pair containing 5'-*XbaI* and *SphI*-3' restriction sites. The primers used: for 5GC, 4XBA-Bx5GCR and 5WtGCF-SPH1U; for 6GCh1, 4XBA-Bx5GCR and Bx6UHF-SPH1U; for 6GCh2, 4XBA-Bx5GCR and Bx6DHF-SPH1U; for 7GC, 4XBA-Bx5GCR and 5MtGCF-SPH1U; for 7GG, 4XBA-Bx5mR and Bx5mFmut-SPH1U. Resulting fragments were subsequently cloned into *XbaI-SphI* digested MC4R Wt plasmids.

The h1h2*-a1, h1h2*-a3, h1h2*-a4, h1h2*-a7 and h1h2*-a8 clones (Figure 22A, Results 3.2.4. for sequences) were generated by *Pfu* amplification with phosphorylated internal primers placed immediately next to each other, resulting in no loss of sequence, save the A to G changes encoded by the primers. Resulting fragments were ligated together and subsequently re-amplified using the external primer pair containing 5'-*XbaI* and *SphI*-3' restriction sites. The primers used: for h1h2*-a1, 4XBA-Abx52 and D22F-SPH1U; for h1h2*-a3, 4XBA-Abx42 and D22F-SPH1U; for h1h2*-a4, 4XBA-Abx32 and D22F-SPH1U; for h1h2*-a7, 4XBA-Abx22 and D22F-SPH1U; for h1h2*-a8, 4XBA-Abx12 and D22F-SPH1U. The resulting fragments were cloned into *XbaI-SphI* digested MC4R Wt plasmids.

(4) MC4R A-stretch and C-stretch clones (primer sequences: Appendix, Table 1)

The 17A and CStr plasmids (Figure 23A, Results 3.2.4. for sequences) were constructed by *Pfu* amplification with phosphorylated internal primers placed immediately next to each other, resulting in substitution of the MC4R P1 and the 23 nt immediately upstream by either 17 Adenosines or 15 Cytidines, respectively. The resulting fragments were ligated together and subsequently re-amplified using the external primer pair containing 5'-*XbaI* and *SphI*-3' restriction sites. The

primers used: for 17A, 4XBA-AStrR and AStrF-SPH1U; for CStr**,** 4XBA-CStrR and CStrF-SPH1U. The resulting fragments were subsequently cloned into *XbaI-SphI* digested MC4R Wt plasmids. The AStr-d4 plasmid was created by combination of the *HindIII-HindIII* digested d4 flank sequence with the respective 17A *HindIII* digested plasmid.

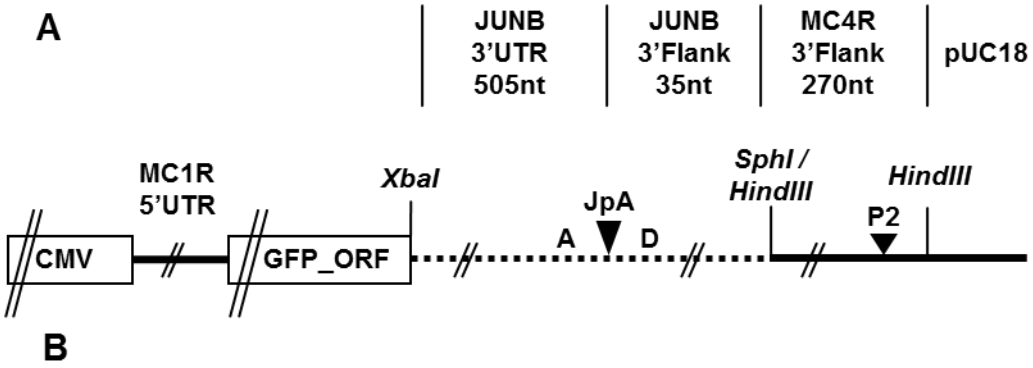
(5) MC4R UGUAN mutation clones (primer sequences: Appendix, Table 1)

The UCUAN-Wt plasmid was constructed by stepwise *Pfu* amplification with phosphorylated internal primers placed immediately next to each other, resulting in no loss of sequence, save the G to C changes encoded by the primers (Figure 24A, Results 3.2.4. for sequences). The resulting fragments were ligated and subsequently re-amplified using the same technique until four precise G to C point mutations in the four UGUAN elements located in the MC4R 3'UTR were combined in one single fragment. This fragment was amplified using the external primer pair containing 5'-*XbaI* and *SphI*-3' restriction sites. The primers used: 4XBA-CF1R, CF2F-SPH1U, 4XBA-CF3R, CF4F-SPH1U, 4XBA-CF5R, CF6F-SPH1U. The resulting fragments were subsequently cloned into *XbaI-SphI* digested MC4R Wt plasmids. The UCUAN-17A plasmid was generated by *Pfu* amplification with phosphorylated internal primers placed immediately next to each other, using the UCUAN-Wt plasmid as a template for the reverse PCR and the 17A plasmid as a template for the forward PCR. The resulting two fragments were then ligated together and subsequently re-amplified using the external primer pair containing 5'-*XbaI* and *SphI*-3' restriction sites. The primers used: 4XBA-AStUGR and AStUGF-SPH1U. The resulting fragments were subsequently cloned into *XbaI-SphI* digested MC4R Wt plasmids. The UCUAN-AStr-d4 plasmid was

created by combination of the *HindIII-HindIII* digested d4 flank sequence with the respective UCUAN-17A *HindIII* digested plasmid.

JUNB reporter gene constructs

The JUNB reporter constructs (Figure 9) retain the CMV promoter, the MC1R 5'UTR and the GFP ORF of the MC4R Wt plasmid described in the previous section. JUNB 3'UTR and flanking sequences were amplified by PCR from HeLa cells genomic DNA, fused to MC4R sequences downstream of the P1 DSE replacing the MC4R 3'UTR and P1 poly(A) sequences using *XbaI-SphI* and *HindIII-HindIII* restriction sites depending on the specific constructs. Mutations of



acgtcccctgccccttacggacacccccctcgcttggacggctgggcacacgcctcccactggggtccagggagcaggc
ggtgggcacccaccctgggacctaggggcgccgcaaaccacactggactccggccctcctaccctgcgcccagtcctt
ccacctcgacgtttacaagcccccccttccacttttttttgtatgtttttttctgctggaaacagactcgattcatattgaatataat
atatttgtgtatttaacagggaggggaagaggggcgatcgcggcggagctggccccgccgcctggtactcaagcccgc
ggggacattgggaaggggaccccgcccccctgccctcccctctctgcaccgtactgtggaaaagaaacacgcacttagt
ctctaaagagtttattttaagacgtgtttgtgtttgtgtgtgtttgttctttttattgaatctatttaagt**AAAAAAAAAA**ttggttct
ttattaatttc**T**gttgtctttttttccaagctgggagggcgggggga

Figure 9: (A) Schematic representation of the JUNB Wt reporter plasmid. CMV promoter and GFP_ORF are shown as open boxes. Lines across indicate regions not drawn to scale. Nucleotide lengths of cloned regions are indicated above the diagram. JUNB 3'UTR and 3'flanking regions are represented by a dotted line and MC4R 3'flank sequences by a solid line. Main restriction enzyme sites used for cloning are depicted. JUNB poly(A) site (JpA) and MC4R poly(A) site (P2) are filled triangles. The positions of the JUNB A-rich region (A) and DSE (D) are indicated. **(B)** JUNB 3'UTR (underlined) and 3'flank nucleotide sequences. JUNB pA A-rich sequence and main cleavage site are shown in bold uppercase and underlined.

the JUNB A-rich sequences and DSE were introduced by PCR.

(1) JUNB Wt plasmid (primer sequences: Appendix, Table 2)

The JUNB Wt plasmid (J-Wt) was constructed by *Pfu* amplification with phosphorylated internal primers placed in the JUNB 3'flank and MC4R d4 plasmid 3'flank (Figure 27A, Results 3.3. for sequences). The resulting two fragments were then ligated together and subsequently re-amplified using the external primer pair containing 5'-*XbaI* and *SphI*-3' restriction sites. The primers used: J3UXF-JFLR2 and 4R6UF-SPH1U. Resulting fragments were subsequently cloned into *XbaI-SphI* digested MC4R Wt plasmids replacing the MC4R 3'UTR and P1 poly(A) sequences. This plasmid was defined as the JUNB Wt plasmid.

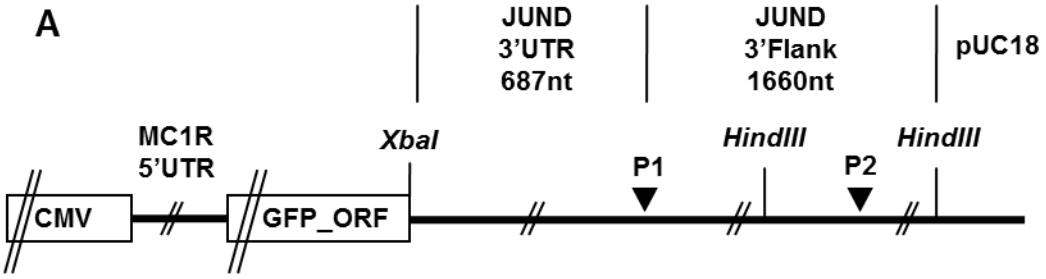
(2) JUNB A-rich core sequence and DSE mutation plasmids (primer sequences: Appendix, Table 2)

The J-Amt and J-Dmt plasmids were constructed from the J-Wt plasmid by *Pfu* amplification with phosphorylated internal primers placed immediately next to each other, resulting in no loss of sequence, save the A to C (J-Amt) or the T to C (J-Dmt) changes encoded by the primers (Figure 27A, Results 3.3. for sequence). Resulting fragments were then ligated together and subsequently re-amplified using the external primer pair containing 5'-*XbaI* and *SphI*-3' restriction sites. The primers used: for J-Amt, J3UXF-JAmtR and JDSEF-SPH1U; for J-Dmt**,** J3UXF-JAWtR and JDSEmF-SPH1U. Resulting fragments were subsequently cloned into *XbaI-SphI* digested J-Wt plasmids.

JUND reporter gene constructs

The JUND reporter constructs (Figure 10) retain the CMV promoter, the MC1R 5'UTR and the GFP ORF of the JUNB Wt plasmid described in the

previous section. JUND 3'UTR and flanking sequences were amplified from HeLa



B

gtccgcgcgcggggcgcatgcgcggccaccctccccaaggggcgggctcgcgggggggtgtcgtgggcgccccggacttg
gagagggtgcggccctggggaccccccctccccgagtgtgcccaggaactcagagagggcgcggcccccgggggattccc
ccccccccgagggtgcccaggactcgacaagctggacccccctgctcccgggggggcgagcgcatgacccccccgccctcg
cgctgcctctttccccccgcgcggccgccccgtgttgcacaaacccgcgcgtctcggctgccccctttgtacaccgcgccgcggaa
gggggctccgaggggggcgcagcctcaaaccctgcctttcctttactttactttttttttttttctttggaagagagaagaacagagtg
ttcgattctgccctatttatgtttctactcgggaacaaacgttggttgtgtgtgtgtgtgttttcttgtgttggttttttaaagaaatgggaag
aagaaaaaaaaaattctccgcccctttcctcgatctcgctccccccttcggttctttcgaccggtcccccctccctttttttgttctgttttgtt
ttgttttgctacgagtccacattcctgtttgtaatccttggttcgcccggttttctgttttc**AGTAAA**gtctcgttacgccagctcgg**C**t
ctccgcctccttcttcccccgccggggcctggcgggctgggcggggcctggttcgcttcccccctacacccccgcctttctctcct
ccctctgtccctagctcctcccctttgttggagggagcagagagcgggtgtctgagagcggcccgcacatctggcgcgcagcga
gcgcccccctagacgcgccttggcgcggggccggctgatgtccaggttggggaggggtcttggggtctttgatctgcgtttggaac
tagctaatgcccccttgcctttttgggaaggagagatttcgaggtggaaagggcacagccccagcccaaagcaggaaggaag
agacagaaacttagtaaacttttaacctttggtccttggtttgcctgtccgtataaaatggggagaacagtcgaaaaacttgccc
cataaagttgtgacgattaagagagtttatgtttacagagcagtgcttggcttacagtgtaagcttttctctttggtggatcgggggat
gctgaggcccaggacaagctgatggacccagtaccaggaggaagttggcacaggtgggatctaacccaagaggtctgcct
ccaggactgggattttccaagtatctccctgaagttcgcttctcctcctcccaaca**T**cctactccctactttgaggaaatctaag
gttcagagaggttgagcaactggcttgacatcacacagccttttaaggagactgtgcagagatagtctgaccttggactccctca
ctttcagagctttgacccgaggcttcctgcttctcctgcctgctctgccctttgctgttaagaccctgcgctttgggcttccaactcattg
gggaccctgagggatgtgaacattagtggccgaggtcacacagcaagtagggctggtggccatgtgtacctagccagccctg
ttagggagggccactgaattgtacccatttacagatgaagaaacttggagagaatggccaccgtcccttggtcatggagaaa
ggccttgccaccaacagctatgttggggttcatttgtttgagacagagtctcgctctgtcacccagactggagtgcagtggcgcca
tctaggctcactgcaacctctgcctccccgggctcaagtgattctcctgcctcagcctccaagtagctcagattacaggtgcccac
catgctcggctacttttttgtatttgtagtagagacagggtttcaccatgttggtcaagctggtctggaactcctgacctcaagtggtc
cacccgcctcggcctcccaaagtgctgggattacaggcgtgagcccccacgtccggccagggtttggtttttagctgtcacgttg
ccttgtgggaggtcatgcagacgtttcagcagggcctgaaggaacctgctgaagggaactgaggctgaggtgggagagcag
gggtgtcttgacacacaagagaccaagaaccacatctcgggcagagtgtcactggttgggagaccgggacccagcttaggtt
ccagttggcatttgtgagatctcaggcgagtcctcctcccccgatggcctctacatcctgtgctgtgagatgcgggtgtttgcatcc
agcataacatctctcaggtcagaagtcagaactgggaggtgc

Figure 10: (A) Schematic representation of the JUND Wt reporter plasmid. CMV promoter and GFP_ORF are shown as open boxes. Lines across indicate regions not drawn to scale. Nucleotide lengths of cloned regions are indicated above the diagram. Main restriction enzyme sites used for cloning are depicted. JUND annotated poly(A) sites P1 and P2 are filled triangles. **(B)** JUND 3'UTR (underlined) and 3'flank nucleotide sequences. JUND P1 hexamer sequence and main cleavage site are shown in bold uppercase and underlined. P2 cleavage site is shown in bold uppercase and underlined.

cells genomic DNA and sequentially cloned using *XbaI-HindIII* followed by *HindIII-HindIII* digested plasmids. Mutations of T-rich sequences upstream of the P1 hexamer, the P1 hexamer and P1 DSE sequences were introduced by PCR.

(1) JUND Wt plasmid (primer sequences: Appendix, Table 3)

The JUND Wt plasmid (JD-Wt) was constructed by *Pfu* amplification in a stepwise procedure. An initial fragment encompassing the JUND 3'UTR, JUND P1 and ~500nt of the JUND 3'flank sequences was amplified and cloned into a J-Wt digested plasmid using 5'-*XbaI* and *HindIII*-3' restriction sites. A second fragment containing JUND P2 and ~1000nt of JUND 3'flank sequences was subsequently amplified and cloned into the resulting plasmid using a 5'-*HindIII* and *HindIII*-3' restriction digest. The primers used: for the first fragment, JDU1XbF-JD2HindR; for the second fragment, JD3HindF-JD3HindR. This plasmid was defined as the JUND Wt plasmid (Figure 29A and 30A, Results 3.4.1. for sequences).

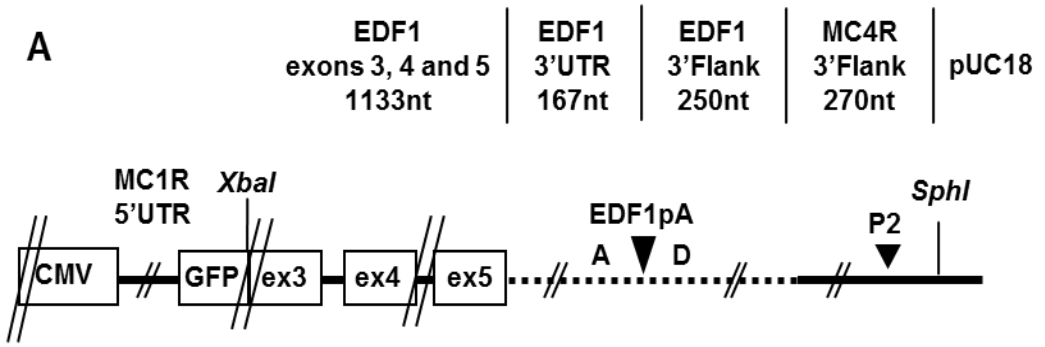
(2) JUND P1 hexamer T-rich upstream sequences, P1 hexamer and P1 DSE mutation plasmids (primer sequences: Appendix, Table 3)

The JD-FUSE, JD-NUSE, JD-Hex and JD-DSE plasmids were constructed from the JD-Wt plasmid by *Pfu* amplification with phosphorylated internal primers placed immediately next to each other, resulting in no loss of sequence, save the T to C (JD-FUSE, JD-NUSE and JD-DSE) and G, T and A to C (JD-Hex) changes encoded by the primers (Figure 29A, Results 3.4.1. for sequences). The JD-FUSE fragment was obtained by a two-step round of PCR amplification in order to introduce all the designed T to C sequence changes. Resulting fragments were ligated together and subsequently re-amplified using the external primer pair containing 5'-*XbaI* and *HindIII*-3' restriction sites. The primers used: for JD-FUSE first amplification, JDU1XbF-JDmt1TUp and JDmt1TDs-JD2HindR; for JD-FUSE

second amplification, JDU1XbF-JDmt2TUp and JDmt2TDs-JD2HindR; for JD-NUSE**,** JDU1XbF-JDp1mtUp and JDp1WtDs-JD2HindR; for JD-Hex**,** JDU1XbF-JDp1WtUp and JDp1mtDs-JD2HindR; for JD-DSE**,** JDU1XbF-JDp1DmtUp1 and JDp1DWtDs1-JD2HindR. Resulting fragments were subsequently cloned into an *XbaI-HindIII* digested J-Wt plasmid.

EDF1 reporter gene constructs

The EDF1 reporter constructs (Figure 11) retain the CMV promoter, the MC1R 5'UTR and a GFP ORF with no stop codon engineered through amplification with a reverse primer containing an in frame *XbaI* restriction site



B

acacaaagcctcgaaatcagtgcgctccagctgatctcgttccgccggttccccttggccgccagttccgttctcctcacgggcc
gaacggaacaaggggtccagcttgcgggggaccctccccagcccattcctgctgtc**AAACAAACAAAA**ccttgcaa
agc**G**ctgggttctgctggctgtttttctccctacatcgcgttcctcagatcagggtgcagaactccaccttcctgtcctggggtgtctg
cccagtagcggttgtggctcagaggcctcggggagctctgagcaccaccttatacaagggtgaggggtggaggcctgaggc
gtggaaggaagaggagcggtgtcactgggcgctgctttctttgatgcagctctgaagagaagctgggaaaagttcctgcccttg

Figure 11: (A) Schematic representation of the EDF1 Wt reporter plasmid. CMV promoter, GFP_ORF and EDF1 exons 3, 4 and 5 are shown as open boxes (EDF1 introns 3 and 4 are solid black lines intercalated with exons). Lines across indicate regions not drawn to scale. EDF1 3'UTR and 3'flanking regions are represented by a dotted line and MC4R 3'flank sequences by a solid line. Main restriction enzyme sites used for cloning are depicted. EDF1 poly(A) site (EDF1pA) and MC4R poly(A) site (P2) are filled triangles. The positions of the EDF1 A-rich region (A) and DSE (D) are indicated. **(B)** EDF1 3'UTR (underlined) and 3'flank nucleotide sequences. EDF1 pA A-rich sequence and main cleavage site are shown in bold uppercase and underlined.

replacing the stop codon and the 3nts immediately downstream, amplified from the JUND Wt plasmid described in the previous section. EDF1 exon 3, intron3, exon 4, intron4, exon 5, 3'UTR and 3'flank sequences were amplified by PCR from HeLa cells genomic DNA and fused immediately upstream of MC4R sequences downstream of the P1 DSE replacing the MC4R 3'UTR and P1 poly(A) sequences. Fragments obtained were cloned into the GFP ORF no stop codon plasmid described above using *XbaI-SphI* restriction sites. Additional constructs containing mutations of the A-rich sequence and DSE were introduced by PCR.

(1) EDF1 Wt and No4R plasmids (primer sequences: Appendix, Table 4)

The EDF1 Wt plasmid (Eex3Wt) was constructed by *Pfu* amplification with phosphorylated internal primers placed in the EDF1 3'flank and MC4R d4 plasmid 3'flank (Figure 32A, Results 3.5.2. for sequences). The resulting two fragments were then ligated together and subsequently re-amplified using the external primer pair containing 5'-*XbaI* and *SphI*-3' restriction sites. The primers used: Eex3XF-EFLR1 and 4R6UF-SPH1U. Resulting fragments were cloned into the above described *XbaI-SphI* digested GFP ORF no stop codon plasmid. This construct was defined as the EDF1 Wt.

A further construct, Eex3No4R, was engineered by *Pfu* amplification. EDF1 exon 3, intron3, exon 4, intron4, exon 5, 3'UTR and 3'flank sequences were amplified by PCR from HeLa cells genomic DNA using a primer pair containing 5'-*XbaI* and *SphI*-3' restriction sites and were not fused upstream of the MC4R d4 flank sequences. The primer pair used: Eex3XF-EFLR3. The resulting fragment was cloned into the above described *XbaI-SphI* digested GFP ORF no stop codon plasmid.

(2) EDF1 A-rich core sequence and DSE mutation plasmids (primer sequences: Appendix, Table 4)

The Eex3Amt and Eex3Dmt plasmids were constructed from the Eex3Wt plasmid by *Pfu* amplification with phosphorylated internal primers placed immediately next to each other, resulting in no loss of sequence, save the A to C (Eex3Amt) or T to C (Eex3Dmt) changes encoded by the primers. The resulting fragments were then ligated together and subsequently re-amplified using the external primer pair containing 5'-*XbaI* and *SphI*-3' restriction sites. The primers used: for Eex3Amt, Eex3XF-EAmutR and EDSEF-SPH1U; for Eex3Dmt**,** Eex3XF-EAWtR and EDSEmF-SPH1U. Resulting fragments were subsequently cloned into a *XbaI-SphI* digested Eex3Wt plasmid (Figure 32A, Results 3.5.2. for sequences).

2.4. Plasmids – Expression

2.4.1. Plasmid Transfection

Human embryonic kidney 293 (HEK293) cells were grown in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% foetal calf serum (FCS), 1% L-glutamine, and 1% penicillin-streptomycin at 37°C with 5% CO2. Cells of 50 to 70% confluence were transfected using Polyfect (QIAGEN) transfection reagent. For the transient transfections, 3µg of MC4R plasmids and 0.01µg of VA plasmid (a cotransfectional control of pUC18 vector containing the adenovirus VA I gene that is constitutively expressed by RNA polymerase III) were added to 150 µl DMEM with 25 µl Polyfect solution and incubated for 30 min at room temperature. This mix was then added to the cells in 5 ml DMEM–10% FCS and incubated at 37°C in 5% CO2. After 5h the medium was replaced with 10 ml of fresh DMEM–10% FCS.

2.5. RNA manipulation and analysis

2.5.1. Total RNA extraction from HEK293 cells

Total RNA was isolated approximately 16 h after transfection using the hot-phenol method. Volumes of 450 µl of NTE buffer (0.1 M NaCl, 10 mM Tris pH 8, 1 mM EDTA) and 50 µl of 10% SDS were added to 500 µl of acidic phenol and heated to 90°C. Subsequently, cell pellets were added to the hot-phenol mix, phenol-chloroform extracted and ethanol precipitated. Pellets were re-suspended in 100µl DNaseI buffer (10mM Tris-HCl pH 7.5, 10mM $MgCl_2$) supplemented with 5 µl DNaseI (Roche), incubated at 37°C for 60 min, phenol extracted and ethanol precipitated. The final pellets were re-suspended in R-loop buffer (80% Formamide, 400mM NaCl, 40mM Pipes pH 6.4, 1mM EDTA) and stored at -20°C.

2.5.2. Polyacrylamide gel electrophoresis

Radiolabelled RNA fragments were analysed on polyacrylamide gels. These were made with 50% urea, 6% acrylamide, 1x TBE, 0.08% ammonium persulphate (APS) and 0.08% N,N,N´,N´-tetramethylethylenediamine (TEMED). Samples were mixed with RNA loading dye (80% Formamide, 10mM EDTA, 0.25% xylene cyanol, 0.25% bromophenol blue) and denatured at 90°C for 10 min before loading. Polyacrylamide gels were run in 1x TBE buffer. Films were developed with a Compact X4 (X-graph Imaging Systems) developer.

2.5.3. *In vitro* transcription – Reactions and Riboprobes

Antisense RNase protection riboprobes were transcribed in reactions containing linearized DNA template plasmid (see paragraph below), 40mM Tris-HCl pH7.9, 6mM $MgCl_2$, 10mM DTT, 1u/µl RNaseOUT (Invitrogen), 0.5mM each

of rATP, rCTP, rGTP, 0.01mM of rUTP and [α-^{32}P]-UTP (400Ci/mmol). Reactions were carried using 50u/µl T7 or 20u/µl SP6 RNA polymerase for 2 hours at 37°C or 40°C respectively. After transcription all the *in vitro*-transcribed antisense RNA probes were gel purified on a 6% polyacrylamide gel and stored in 100 µl R-loop buffer.

BamHI-linearized pGEM4 plasmids (0.5µg) containing a 482nt *XbaI-SphI* digested fragment encompassing the MC4R Wt pA1 and pA2 sites (MC4R Wt specific probe), were transcribed by T7 RNA polymerase. Plasmids containing similar fragments encompassing the H1h2, h1H2, h1h2, d2, d4, H1h2-d4, h1H2-d4 and h1h2-d4 clone pA1 and pA2 sites were transcribed by T7 RNA polymerase generating the respective specific antisense riboprobes. *EcoRI*-linearized pGEM4 plasmids (0.5µg) containing a 472nt *EcoRI-SphI* digested fragment including the MC4R ΔU1 clone poly(A) site (3'ΔU1 probe) were transcribed by T7 RNA polymerase. *EcoRI*-linearized pGEM4 plasmids (0.5µg) containing a 449 nucleotide-long *XbaI-SphI* digested MC4R fragment covering the ΔU2 clone poly(A) site (3'ΔU2 probe) were transcribed by T7 RNA polymerase**.** *EcoRI*-linearized pGEM4 plasmids (0.5µg) containing a 340nt *EcoRI-SphI* digested fragment including 240nt of the GFP ORF and 100nt covering the MC4R pA1 and respective downstream 3'flank sequences (MC4R composite probe) were transcribed by T7 RNA polymerase. Plasmids containing a similar fragment covering the MC4R d4 pA1 and respective downstream 3'flank sequences (MC4R d4 composite probe) were also transcribed.

EcoRI-linearized pGEM4 plasmid (0.5µg) containing a 398nt *EcoRI-SphI* digested fragment encompassing 158nt of the GFP ORF and 240nt (195nt of 3'UTR, 45nt of 3'flank sequences) surrounding the JUND pA1 (JUND composite

pA1 probe) were transcribed by T7 RNA polymerase. *EcoRI*-linearized pGEM4 plasmid (0.5µg) containing a 394nt *EcoRI-SphI* digested fragment encompassing 158 nt of the GFP ORF and 236nt (190nt of 3'UTR, 46nt of 3'flank sequences) surrounding the JUND pA2 (JUND composite pA2 probe) were used for T7 RNA polymerase transcription.

EcoRI-linearized pGEM4 plasmid (0.5µg) containing a 516nt *EcoRI-SphI* digested fragment encompassing 158nt of the GFP ORF and 358nt (75nt of intron 4, 62nt of exon 5, 165nt of 3'UTR, 56nt of 3'flank sequences) surrounding the EDF1 pA site (EDF1Wt probe) were transcribed by T7 RNA polymerase. Plasmids containing similar fragments encompassing the Eex3Amt and Eex3Dmt pA sites were subjected to transcription by T7 RNA polymerase generating the respective Eex3Amt and Eex3Dmt specific antisense riboprobes.

2.5.4. RNase protection analysis

Three to 10 µl (depending on transfection efficiencies) of total RNA isolated from transfected HEK293 were added to 2 to 3 µl of 300- to 500-cps/µl [α-^{32}P]-UTP radiolabeled riboprobe in a total volume of 30 µl R-loop buffer. This hybridization mix was then denatured at 94°C for 20 min and incubated at 56°C for 15 to 18 h. RNAse digestion was carried out by adding 300 µl RNAse mix (300 mM NaCl, 10 mM Tris-HCl (pH 7.4), 5 mM EDTA supplemented with 4 µl RNAse A [10 mg/ml] and 2 µl RNAse T1 [550 U/ml]) to the hybridized samples and subsequent incubation at 30°C for 45 min. RNAse digestion was followed by Proteinase K (Roche) digestion at 37°C for 30 min, phenol-chloroform extraction, and ethanol precipitation. RNA pellets were re-suspended in 15 µl RNA loading buffer, denatured at 90°C for 10 min prior to loading and protected RNA fragments

were subsequently fractionated on a 6% polyacrylamide gel. Undigested control probes were re-suspended in a total RNA loading buffer volume of 50 µl, denatured at 90°C for 10 min before loading and a volume of 10 µl was loaded onto a gel.

2.5.5. 3' rapid amplification of cDNA ends (3'RACE)

Polyadenylated transcripts were reverse transcribed in a reaction containing 20 mM Tris-HCl pH 8.4, 50 mM KCl, 2.5 mM $MgCl_2$, 10 mM DTT, 0.5 mM of each dNTP (dATP, dCTP, dTTP, dGTP), 200u of M-MLV (Moloney-Murine Leukemia Virus) reverse transcriptase (Invitrogen), 1u/µl RNaseOUT (Invitrogen), 1-5µg total RNA and 0.5 µM oligo-dT adapter primer (AP: 5´-GGCCACGCGTCGA CTAGTAC$(T)_{17}$-3´).

Briefly, 1-5µg total RNA and 1 µM of the oligo-dT adapter primer (AP) were incubated at 70°C for 5 min. Then each dNTP, 5x M-MLV RT buffer and 1u/µl of RNaseOut were added and the reaction was kept at 37°C for 5 min. 200u of M-MLV RT were subsequently introduced and the solution was exposed to 42°C for 60 min followed by 15 min at 70°C to inactivate the enzyme. The amplification of cDNA templates was performed with 1 µM universal amplification primer (UAP; 5´-GGCCACGCGTCGACTAGTAC-3´) and 1 µM of a gene specific forward primer using *Taq* polymerase (Invitrogen). The reactions were exposed to 94°C for 2 min and then 25 cycles of 94°C for 1 min, 53°C for 1 min and 72°C for 2 min.

2.5.6. Reverse transcriptase polymerase chain reaction (RT-PCR)

For RT-PCR analysis, a fraction of total RNA isolated from cells transiently transfected with individual MC4R reporter plasmids was subjected to reverse

transcription using Superscript III reverse transcriptase (Invitrogen) and a oligo-dT (oligo-dT17) reverse primer. Resulting cDNAs were amplified by PCR and analysed either on agarose gels.

Total RNA was reverse transcribed in 20 µl reactions containing 50 mM Tris-HCl pH 8.4, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, 0.5 mM of each dNTP (dATP, dCTP, dTTP, dGTP), 200u of Superscript Reverse Transcriptase III (Invitrogen), 1u/µl RNaseOUT (Invitrogen), 1-5µg total RNA and 2.5 µM of gene specific primer (GSP) primer.

Briefly, reactions containing 1-5µg of RNA, 2.5 µM of the GSP and 200 µM of dATP, dCTP, dGTP and dTTP were incubated at 80°C for 10 min. 5x First strand buffer (Invitrogen), 10mM DTT and 1u/µl of RNaseOut (Invitrogen) were added and the reaction was kept at 55°C for 5 min. Finally, 200u of Superscript Reverse Transcriptase III (Invitrogen) were included and the reaction was incubated at 55°C for 60 min and stopped by inactivation of the reverse transcriptase at 70°C for 15 min. For all RT-PCR analysis, a negative control was included where the cDNA reaction contained no reverse transcriptase. Subsequently, 2-5 µl of cDNA were subjected to PCR with *Taq* DNA polymerase (Invitrogen). Reactions were incubated at 94°C for 2 min and then 25 cycles of 94°C for 1 min, 53°C for 1 min and 72°C for 2 min.

2.6. Quantitation of RNase Protection data

Radiolabelled RNA fragments generated by RNase Protection were quantitated using a Fujifilm Fluorescent Image Analyser FLA-3000. The data signals refer to photo stimulated luminescence (PSL) units.

2.7. Statistical analysis of data generated by RNase Protection

Data collected with the Fujifilm Fluorescent Image Analyser FLA-3000 was statistically analysed and is presented as the mean ± standard deviation of the mean (stdev). The data collected was analysed using Microsoft Office Excel 2010.

2.8. Bioinformatics analysis

Note: The methods presented in this section were not directly used by the author of the present dissertation and correspond to work developed in collaboration with Wencheng Li and Bin Tian from the Department of Biochemistry and Molecular Biology, UMDNJ-New Jersey Medical School. This bioinformatics analysis was originated from the experimental data generated during the work that is subject of the present dissertation and provides a relevant contextualization, at the level of the human transcriptome, for the gene specific results presented in the next section.

2.8.1. Poly(A) sites

The poly(A) sites in PolyA_DB2[352] were used for this study. To minimize the issue of internal priming[353], we used only poly(A) sites supported by cDNA, EST or Trace sequences with poly(A/T) tail length greater than 30. Poly(A) sites were grouped into single poly(A) sites (S); first (F), middle (M) and last (L) poly(A) sites in genes with alternative poly(A) sites located in the 3'-most exon, as previously described in Tian et al., 2005[298]. The resulting number of human poly(A) sites studied were 2,361, 1,722, 2,347 and 4,204 for F, M, L and S types, respectively. Canonical poly(A) signal was defined as either AAUAAA or AUUAAA in -10 to -40

nt region of the cleavage site (position 0). Conservation of human poly(A) sites in mouse was analyzed as previously described[354].

2.8.2. *Cis* element analysis

A-rich elements were hexamers with at least 5 adenosines excluding AAUAAA. We also required that A-rich elements did not overlap with A(A/U)UAAA in sequence. GU-rich and U-rich elements were hexamers corresponding to CDE.3 and CDE.2 elements defined in a previous study[250], respectively. The 5-mer UUUUU was added to the U-rich element group. To identify *cis* elements significantly biased to A-rich or A(A/U)UAAA poly(A) sites, the occurrence of each 4-mer in the +5 to +40 region was counted for A-rich and A(A/U)UAAA poly(A) sites. A 2×2 contingency table was created with columns for A-rich only and A(A/U)UAAA only poly(A) sites, and rows for the occurrences of the 4-mer and all other 4-mers. The significance of bias was evaluated by the Fisher's exact test.

2.8.3. Analysis of poly(A) site usage using mRNA-seq data

The mRNA-seq data were downloaded from the NCBI Gene Expression Omnibus (GEO) database. Two data sets, GSE12946 and GSE13652, were used, which were previously reported in[355] and[140]. The combined set includes 13 human tissue samples, i.e. brain, liver, heart, skeletal muscle, colon, adipose, testis, lymph node, breast, mixed human brain, Ambion human brain reference RNA, Stratagene Universal Human Reference RNA (UHR), cerebral cortex, and lung, and 5 mammary epithelial or breast cancer cell line samples, i.e. HME, BT474, MCF-7, MB435, T47D. The sequencing reads were mapped to the human genome (hg18), allowing at most two mismatches. To evaluate poly(A) site usage,

densities of reads mapped to upstream and downstream regions were compared, as illustrated in Sup_2. We used the relative usage of downstream poly(A) site (RUD) score, which is (density of downstream reads)/(density of upstream reads). A small RUD score for a poly(A) site represents high usage of the poly(A) site. Poly(A) sites were those reported in PolyA_DB2[352]. The reads mapped in the +/- 10 nt region around the poly(A) sites were not used for RUD calculation because the cleavage sites usually are not precise.

Chapter 3

Results

3. Results

3.1. Introduction

As stated above, 3'end formation is a fundamental processing step for the maturation of mRNAs in eukaryotes. Protein encoding primary transcripts are cleaved at their 3'end and, with the exception of replication-dependent metazoan histone genes, are subsequently subjected to polyadenylation[199, 225, 356].

The core poly(A) sequence is constituted by a bipartite element which, in more than 80% of human genes, is defined by an upstream conserved canonical hexamer motif recognised by CPSF (AAUAAA or AUUAAA) and a less defined downstream U/GU-rich region (DSE) contacted by CstF[220, 225, 298].

Recognition of 15-30% of human 3'end processing sites by the cleavage and polyadenylation machinery is currently poorly understood since these genes do not contain canonical hexamers[220, 225, 298]. Additional interactions between core factors and previously unidentified sequence elements can promote cleavage and polyadenylation at non-canonical 3'end processing sites and are thought to be involved in this recognition. Such interactions have been reported for CFI_m and UGUAN motifs located in the 3'UTR[203].

Interactions between components of the splicing machinery directing terminal intron removal and the cleavage and polyadenylation apparatus result in reciprocal enhancement of both processing reactions[329-331, 335].

Similar to non-canonical poly(A) sites, 3'end processing of mammalian and viral intronless primary transcripts also appear to require additional auxiliary *cis*-elements perhaps to bypass the absence of the stimulatory effect of splicing[329-331, 335]. Several pre-mRNA processing enhancer elements (PPE) located upstream of

the core poly(A) sequences have been described for viral and mammalian non-spliced genes[245, 345, 346].

The work described in the present dissertation aims to further clarify how efficient 3'end processing is achieved in human naturally intronless genes since terminal intron removal has been shown to significantly contribute for the efficiency of cleavage and polyadenylation[209, 210, 213, 227]**.** Results from the analysis of the 3'end formation reaction in the naturally intronless human Melanocortin 4 Receptor (MC4R), JUNB, JUND and for the intron-containing human EDF1 genes are presented and discussed.

The MC4R and JUNB results constitute peer reviewed published material: Nunes, N.M., Li, W., Tian, B. & Furger, A. (2010) A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence. *EMBO J* **29**, 1523-36 (see Appendix).

3.2. Melanocortin 4 Receptor (MC4R) 3' End Formation Analysis

3.2.1. G-protein Coupled Receptors (GPCRs)

Although it is thought that less than 15% of human genes lack introns in their protein-coding regions, a substantial number (maybe >80%) of mammalian G-protein coupled cell surface receptors (GPCRs) are intronless[340, 357]. The GPCRs are integral membrane proteins with a highly conserved structural feature comprising seven transmembrane α-helices, an extracellular N-terminus and an intracellular C-terminus[358]. The GPCR protein family detects extracellular signals and transduces them into the cell trough activation of guanosine diphosphate (GDP) to guanosine triphosphate (GTP) exchange on the constitutively associated heterotrimeric G-protein (G-protein activation) and thus originating an intracellular

signal (second messenger cascade) that will stimulate or inhibit biochemical modifications inside the cell[358].

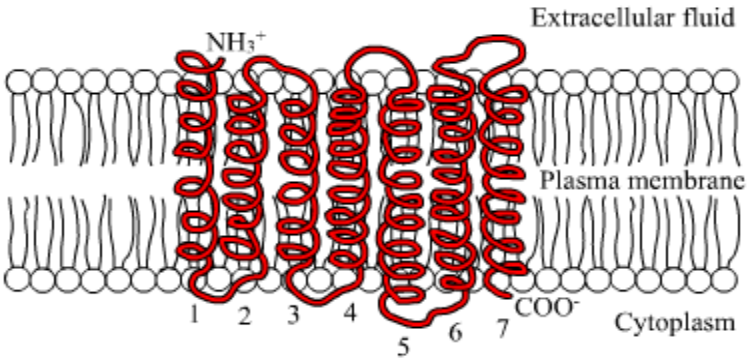


Figure 12: Seven transmembrane α-helix structure of a G-protein-coupled cell surface receptor (GPCR). (http://www.search.com/)

3.2.2. The Melanocortin Receptors

The melanocortin receptors (MCRs) belong to the rhodopsin family of human GPCR[359] and five distinct melanocortin receptor subtypes have been identified: MC1R, MC2R, MC3R, MC4R and MC5R[360-366]. The Melanocortin system consists of endogenous agonists, antagonists, the receptors and putative auxiliary proteins that regulate several physiological functions through the cAMP signalling transduction pathway. The melanocortin system is implicated in the regulation of a variety of physiological pathways including pigmentation, steroid function, energy homeostasis, food intake, obesity, cardiovascular, sexual function and normal gland regulation[367, 368].

3.2.3. Melanocortin 4 Receptor (MC4R)

MC4R is a 333 amino acid seven transmembrane protein encoded by a single exon gene localized to the human chromosome locus 18q22[369]. The MC4R gene is expressed in multiple sites in the brain including the cortex, thalamus,

hypothalamus, hippocampus and brainstem[365, 370-372]. The hypothalamus receives and integrates neural, metabolic and humoral signals from the periphery. In particular, neurons within the hypothalamic arcuate nucleus (ARC) function as primary sensors of alteration in energy stores to control appetite and energy homeostasis[373, 374]. Ablation of the MC4R gene in mice results in animals that accumulate fat to greater proportion of body weight than normal animals, hyperphagia and eventual hyperglycemia and hyperinsulinaemia[375-377]. Mutations in the MC4R gene have also been described as a cause for inherited severe human obesity[378, 379]. Furthermore, MC4R has been shown to directly play a role in increasing thermogenesis in response to dietary fat and in maintaining general activity levels[380].

3.2.4 Results - MC4R 3'end formation sequence requirements

Mapping of the MC4R 3'end processing site

A CMV promoter driven MC4R reporter plasmid containing the 5'UTR sequences from the MC1R gene and the GFP ORF was cloned containing 1.3kb of sequences located downstream of the MC4R stop codon including the 3'UTR, and 1044 nucleotides of 3'flanking sequences (see Figure 13 and Materials and Methods, section 2.3.2., Figure 8 for cloning details).

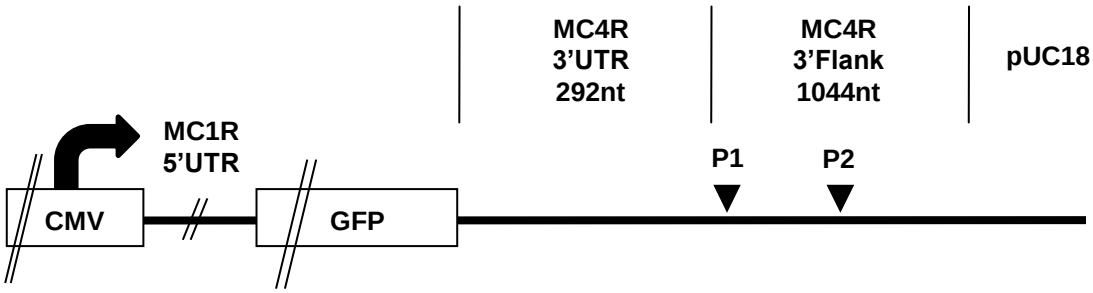


Figure 13: MC4R reporter gene. Borders between ORF, 3'UTR, 3'flank and vector backbone, are indicated by thin straight vertical lines; CMV promoter and GFP ORF are represented by open boxes, lines across indicate that regions are not drawn to scale. MC4R poly(A) sites P1 and P2 are filled triangles.

The initial experiments with the MC4R reporter construct were aimed at mapping the MC4R 3'end processing site since this had not been annotated and sequence comparison revealed at least two potential 3'end processing sites located in the first 400 nucleotides downstream of the MC4R stop codon (Figure 14: P1, P2). The MC4R reporter plasmid (containing the CMV promoter to drive efficient ligand and tissue independent transcription) was transfected into HEK 293 cells (together with a co-transfection control plasmid containing the RNA polymerase III transcribed adenovirus VA I gene) and total RNA was isolated and subsequently analysed by 3'RACE and RNAse protection (RP) (see Materials and Methods). 3'Race and RP analysis produced similar results mapping the MC4R poly(A) cleavage site to the same position 292 nucleotides downstream of the MC4R ORF (Figure 14C and D: P1). The RP analysis also showed that the first poly(A) site (P1) is efficiently used and no readthrough transcripts either processed at the second poly(A) site (P2) or RNAs that are not processed at all (rt) were detected (Figure 14D: lane 2: P2, rt). Further sequence analysis showed that an additional hexamer, albeit without a DSE, can be found overlapping the MC4R stop codon and the first nucleotides in the 3'UTR (Figure 14B: P?). Therefore an additional construct containing the MC4R 5'UTR, ORF, 3'UTR and 3'flanking sequences (Figure 14B) was created and subsequently transfected into HEK293 cells. 3'RACE analysis of total RNA isolated from these transfected cells showed that the putative hexamer overlapping the endogenous MC4R stop codon and the first nucleotides within the 3'UTR is not functional (Figure 14C: lane MC4R-ORF).

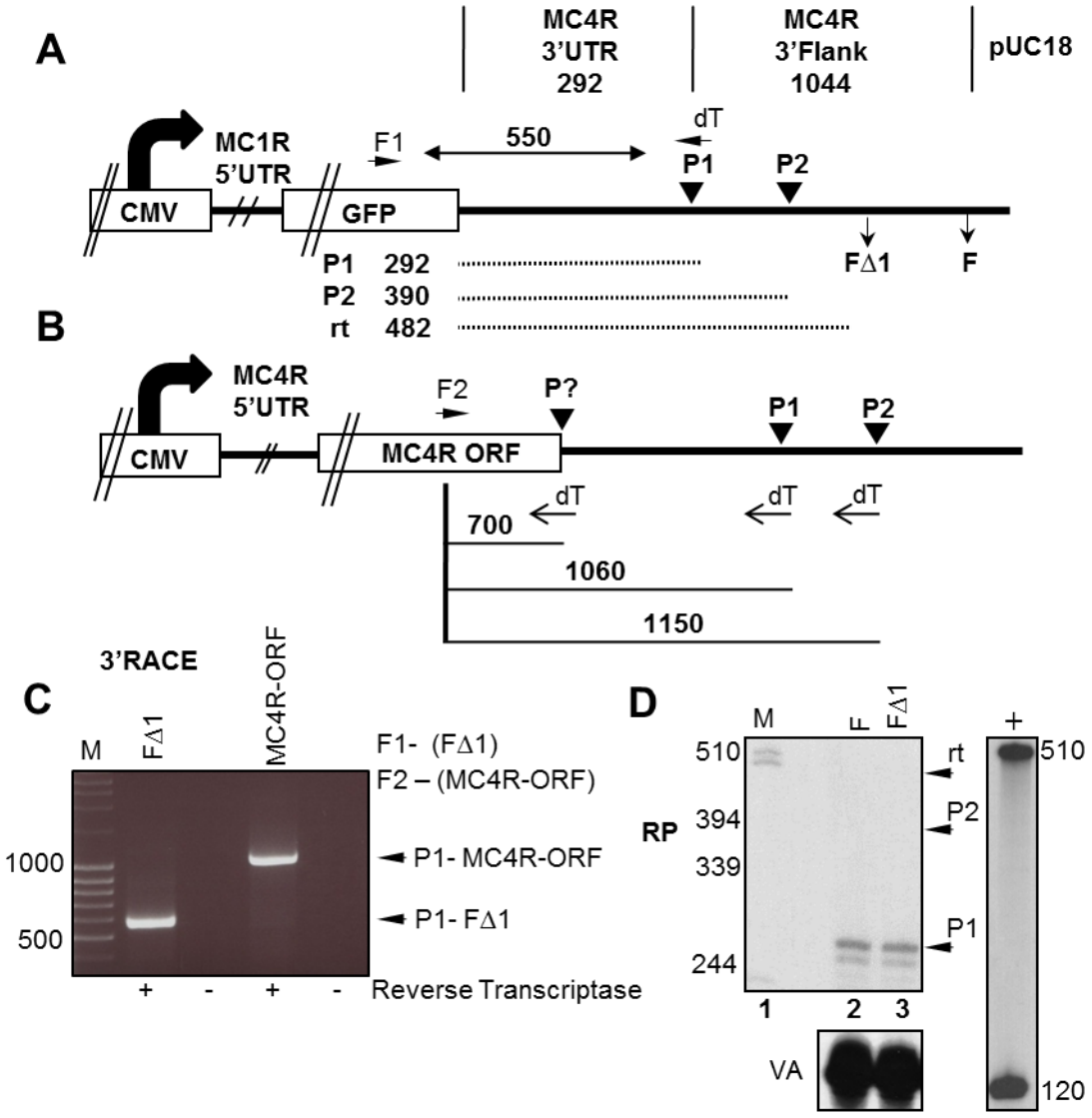


Figure 14: Analysis of poly(A) site use in MC4R reporter constructs. **(A)** Diagram depicting MC4R reporters: F and deletion of 3'flank sequences (FΔ1). Vertical arrows indicate the end of the deletion clones relative to wt sequence. MC4R poly(A) sites P1 and P2 are filled triangles. RNAse protection fragments uncleaved (rt), cleaved at P1 (P1) or cleaved at P2 (P2) are shown as dotted lines and the expected lengths are indicated. Forward primers F1 and F2 used in the 3'RACE analysis shown in C are indicated above the diagrams. **(B)** Diagram showing the MC4R ORF construct containing the MC4R 5'UTR, ORF, 3'UTR and 3' flanking sequences. The length of expected PCR products using oligo dT primers (dT) and the forward primer (F2) located in the MC4R ORF for each poly(A) site are shown. **(C)** 3'RACE analysis showing that P1 is the major cleavage and polyadenylation site used in the MC4R ORF and FΔ1 construct (Wt). Size markers are indicated. **(D)** RNAse protection analysis of total RNA isolated from HEK293 cells transiently transfected with constructs containing 1044nt (F) or 250nt (FΔ1) of 3'flank sequences. Transcripts not cleaved at P1 are indicated either as transcripts cleaved at P2 or uncleaved readthrough transcripts (rt).VA transfection control (VA) and undigested probe control (+) are shown (right panel).

MC4R pre-mRNA 3'end formation does not require additional 3'flank sequence elements

Analysis of viral and human intronless transcripts suggests that intronless pre-mRNAs may generally rely on auxiliary sequences located in the 3'UTR or 3'flanking regions to direct efficient 3'end processing[234, 245, 345, 346].

Several reporter plasmids were designed and used to identify auxiliary *cis*-

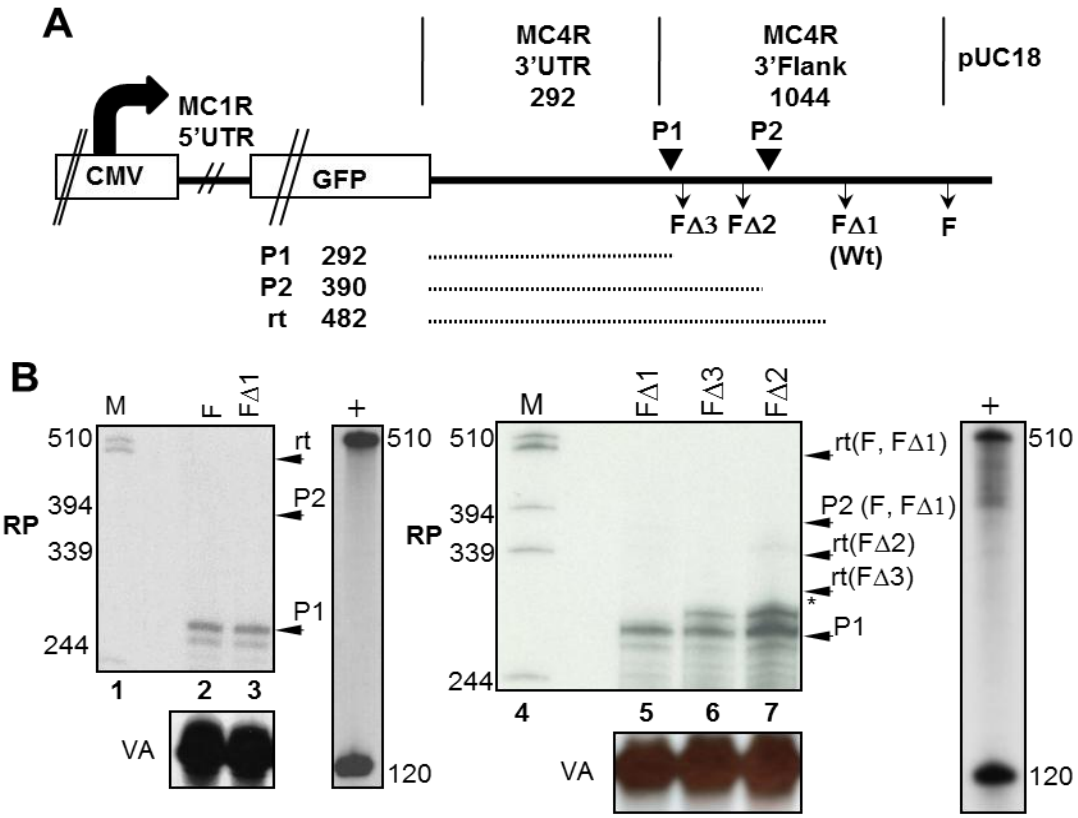


Figure 15: The MC4R poly(A) site does not require auxiliary 3'flank sequence elements. **(A)** Diagram depicting MC4R reporters: F and deletion of 3'flanking sequences (FΔ1, FΔ2 and FΔ3). Vertical arrows indicate end of the deletion clones relative to Wt sequence. RNAse protection fragments uncleaved (rt), cleaved at P1 (P1) or cleaved at P2 (P2) are shown as dotted lines and the expected lengths are indicated. **(B)** RNAse protection analysis of total RNA isolated from HEK293 cells transiently transfected with constructs containing 3'flank deletions. Transcripts not cleaved at P1 are indicated either as transcripts cleaved at P2 for (F, FΔ1) or uncleaved readthrough transcripts rt = rt(F, FΔ1), rt(FΔ1), rt(FΔ2). Alternative cleavage site use at P1 observed with plasmids FΔ2 and FΔ3 is indicated by (*). FΔ1 is subsequently referred to as wild type (Wt). VA transfection control (VA) and undigested probe control (+) are shown (right panels).

elements required for cleavage and polyadenylation of the MC4R pre-mRNA.

Three additional plasmids with gradually shorter 3'flanking sequences compared to the full length clone F (Figure 15: FΔ1, FΔ2, FΔ3 and Materials and Methods, section 2.3.2., Figure 8 for further cloning details) were constructed and analysed in order to verify the presence of potential auxiliary sequence elements in the MC4R 3'flank. An antisense riboprobe complementary to sequences overlapping P1 and P2 which results in protected bands of the same length for all transcripts cleaved at P1 was used for the RP analysis of all 3'flank clones (Figure 15: P1- 292nt). Due to 3'flank sequence deletions in the FΔ2 and FΔ3 plasmids, transcripts generated from these reporters that are not processed at P1, would result in protected readthrough bands of different lengths compared to transcripts originating from the F and FΔ1 plasmids (rt (F, FΔ1), rt(FΔ2) and rt(FΔ3)). The results presented in figure 15B clearly show that deletion of all but 25 nucleotides of the 3'flank (FΔ3, counted from the site of cleavage) or less (FΔ2, FΔ1) had no effect on the cleavage efficiency at P1 since no bands can be seen corresponding to transcripts that failed to cleave at P1 (Figure 15: compare lanes 2, 3 and lanes 5-7, rt(F, FΔ1), rt(FΔ3), rt(FΔ2) and P2 respectively). However, larger deletions of the 3'flank resulted in the appearance of an additional less intense protected band which is likely to be caused by a shift in the site of cleavage in some of the FΔ3 and FΔ2 transcripts (Figure 15B: lanes 6, 7; *). For all further experiments we used FΔ1 as the wild type reference and thus FΔ1 is subsequently referred to as Wt. Please note that FΔ1 behaves identically to the F clone as can be seen in Figure 15B. From the results described it was concluded that no sequences in the MC4R 3'flank sequences are required to direct efficient cleavage at the MC4R poly(A) site.

MC4R pre-mRNA 3'end formation does not require additional 3'UTR sequence elements

To address whether sequences located in the 3'UTR are required for

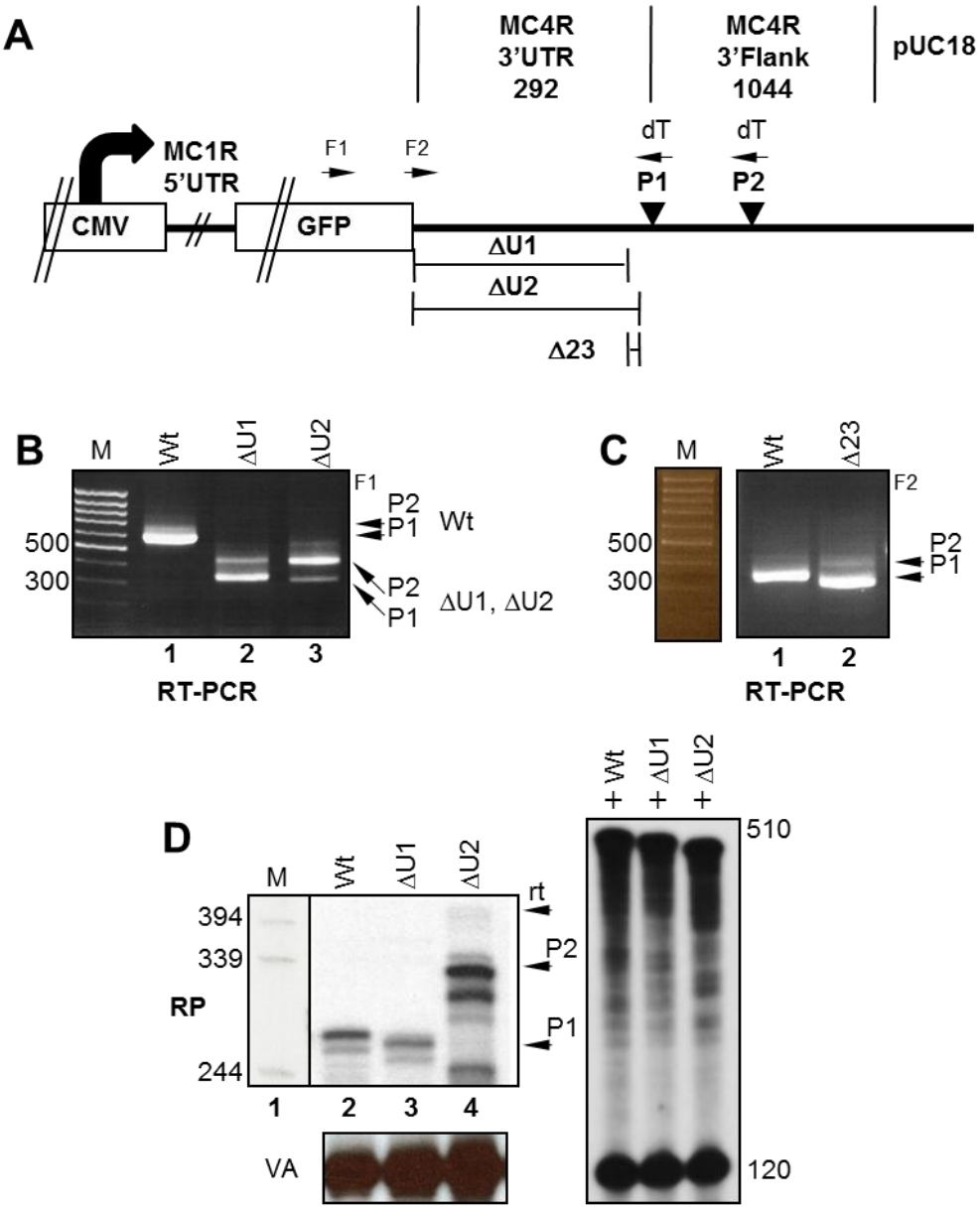


Figure 16: The MC4R poly(A) site does not require auxiliary 3'UTR sequence elements. **(A)** Diagram depicting MC4R 3'UTR deletion reporters: The regions deleted in clones ΔU1, ΔU2 and Δ23 are indicated below the graph. **(B, C)** RT-PCR analysis of 3'UTR deletion constructs. RT-PCR products corresponding to mRNAs cleaved at either P1 or P2 are indicated for Wt and 3'UTR deletion clones. Size markers are indicated. **(D)** RNAse protection analysis of total RNA isolated from HEK293 cells transiently transfected with constructs containing 3'UTR deletions (ΔU1, ΔU2). Transcripts not cleaved at P1 are indicated either as transcripts cleaved at P2 (P2) or uncleaved readthrough transcripts (rt). VA transfection control (VA) and construct specific undigested probe controls (+) are shown (right panel).

efficient 3'end processing of the MC4R pre-mRNA, two reporter plasmids were constructed. One construct which had almost all 3'UTR sequences removed, retaining only the last 23 nucleotides immediately upstream of the P1 AUUAAA hexamer and a second construct which had the AUUAAA directly fused to the GFP stop codon (Figure 16: ΔU1, ΔU2, respectively, and Materials and Methods, section 2.3.2., Figure 8 for further cloning details). Oligo-dT primed RT-PCR analysis of total RNA isolated from cells transfected with the ΔU1 reporter construct showed that ΔU1 had only a marginal effect on P1 usage as can be seen in Figure 16B and 16D. Interestingly, deletion of the entire 3'UTR (ΔU2) appeared to dramatically shift the preferred cleavage site from P1 to P2 (Figure 16B, compare lanes 1-3 and Figure 16D, compare lanes 1-3). This result suggested either the possibility of a 23 nucleotide long enhancer element located immediately upstream of the P1 hexamer or that locating P1 site close to the GFP ORF or 5'UTR sequences somehow reduces processing efficiency at the P1 poly(A) site. In order to address this, a third construct was built that retained all but the last 23 nucleotides of the 3'UTR (Figure 16A: Δ23). As can be seen in Figure 16C, precise deletion of these 23 nt did not result in a significant shift from P1 into P2 usage (compare lanes 1, 2).

From the results discussed above it was concluded that no sequences in the 3'UTR are required to direct efficient cleavage at the MC4R poly(A) site. Furthermore, these results also established P2 as an additional functional poly(A) site and that transcripts not cleaved at the P1 poly(A) site are subsequently efficiently cleaved at P2 (Figure 16D, compare lanes 1-3). The switch from a P1 cleavage event to P2 cleavage event was thus used to measure effects on poly(A) cleavage at P1.

Analysis of the MC4R P1 core poly(A) signal

The results described above strongly suggested that MC4R pre-mRNA 3'end processing is directed by nucleotides composing the core poly(A) and possibly the surrounding sequences. The functional relevance of the core sequences was addressed by focusing on the hexamer motifs. The P1 poly(A) site in MC4R contains two potential hexamers, AUUAAA and AAGAAA (see H1 and H2 in Figure 17A). Although the presence of a guanosine at position 3 in hexamers is normally considered as a potent inactivating mutation [381, 382], it has been shown to be active in at least one gene[383]. To test the functional contribution of both hexamer sequences, three reporter constructs were created with mutations destroying either (H1h2, h1H2,) or both (h1h2) hexamers (Figure 17A). HEK293 cells were transfected with these plasmids and total RNA was subsequently isolated and analysed by RP and RT-PCR. As can be seen in Figure 17, while mutations of hexamer sequences have long been known to severely impair 3'end processing[227, 228, 384], mutating the MC4R hexamer(s) had surprisingly little effect on cleavage at P1 when analysed by RP and RT-PCR (Figure 17B: compare lanes 1-4 in RP and RT-PCR panels; quantitation results are presented as a percentage of total P1 readthrough transcripts: rt(P1)=[1-P1/(P1+P2+rt)]*100). Following the unexpected results on the hexamer mutations reporter plasmids, two additional constructs were created where 2 or 4 thymidines in the DSE were substituted by cytidines (Figure 17A: d2, d4 respectively). In contrast to mutations of the hexamers, DSEs are generally described as being relatively tolerant to sequence changes[82].

Unlike the hexamer mutations results described above, substitution of 2 uridines resulted in a more than seven fold increase in P2 usage and introduction

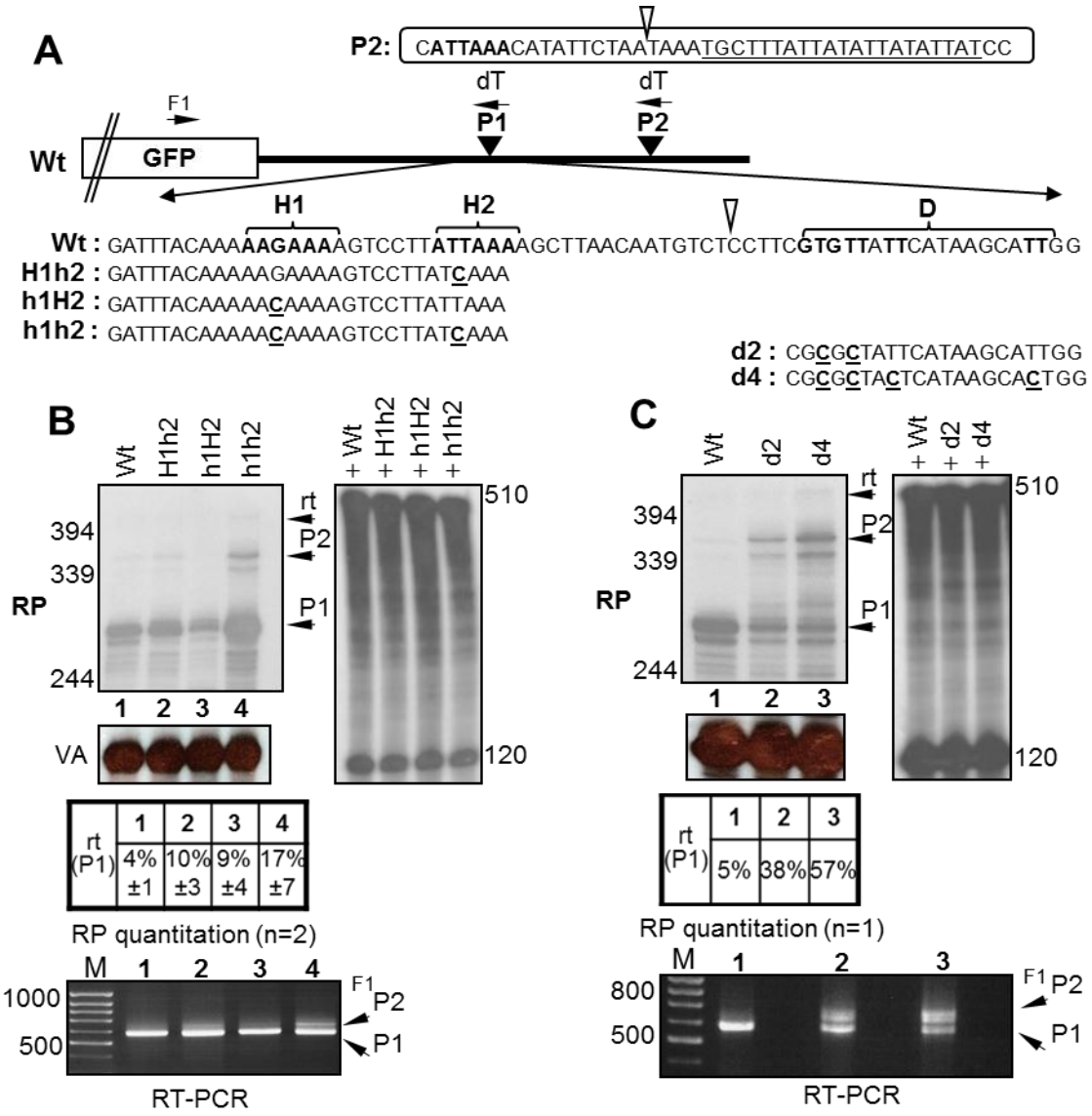


Figure 17: Mutations of the core poly(A) sequences have unexpected effects on cleavage efficiency. **(A)** Diagram of the MC4R reporter: both potential poly(A) sites are indicated by filled triangles. Potential priming of oligo-dT reverse primers (dT) at P1 and P2 and forward primer (F1) are depicted above the diagram. Sequences surrounding P2 are indicated in the box above the graph. Hexamer is in bold, DSE nucleotides are underlined, site of cleavage is indicated by the open triangle. Below the diagram is the sequence surrounding P1, two hexamers are in bold and indicated by (H1) and (H2) respectively. Capital H represents clones with wild type hexamer sequences, small caps h (h1 and/or h2) represent mutated hexamers. Open triangle marks the P1cleavage site and the DSE (D) is indicated in bold letters. Mutated nucleotides for each clone are depicted in bold and underlined below the wt sequence. **(B,C)** RNAse protection (RP, top gels) and RT-PCR analysis (bottom gels) of total RNA isolated from transiently transfected cells. Expected migration patterns of transcripts cleaved at P1, P2 or unprocessed readthrough RNA (rt) are indicated on the left of each gel. Quantitation of independent RP's as an average percentage of total P1 readthrough transcripts are shown below. VA transfection control (VA) and construct specific undigested probe controls (+) are shown (right panels).

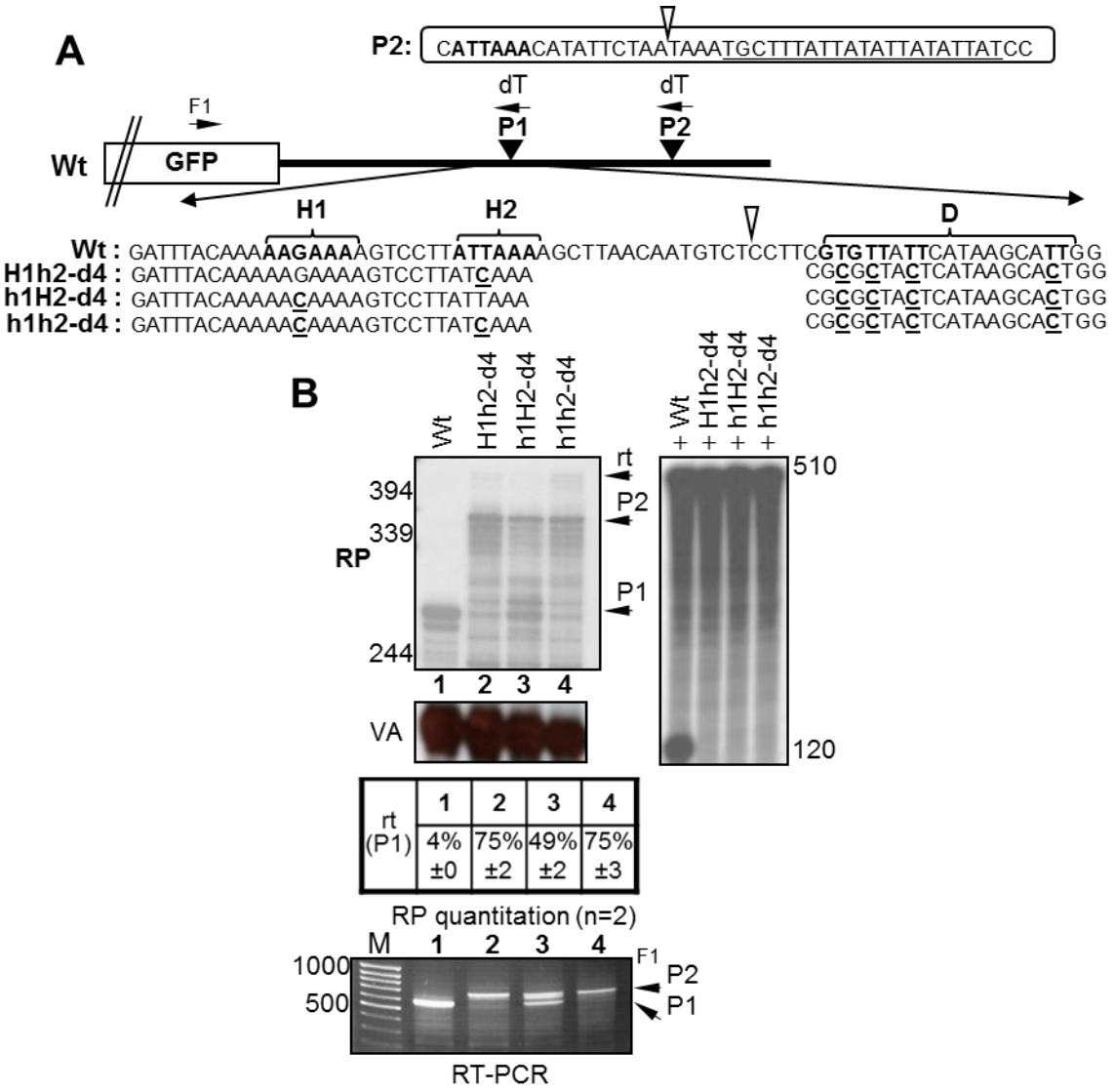


Figure 18: Mutations of the core poly(A) sequences have unexpected effects on cleavage efficiency. **(A)** Diagram of the MC4R reporter: both potential poly(A) sites are indicated by filled triangles. Potential priming of oligo-dT reverse primers (dT) at P1 and P2 and forward primer (F1) are depicted above the diagram. Sequences surrounding P2 are indicated in the box above the graph. Hexamer is in bold, DSE nucleotides are underlined, site of cleavage is indicated by the open triangle. Below the diagram is the sequence surrounding P1, two hexamers are in bold and indicated by (H1) and (H2) respectively. Capital H represents clones with wild type hexamer sequences, small caps h (h1 and/or h2) represent mutated hexamers. Open triangle marks the P1cleavage site and the DSE (D) is indicated in bold letters. Mutated nucleotides for each clone are depicted in bold and underlined below the wt sequence. **(B)** RNAse protection (RP, top gel) and RT-PCR analysis (bottom gel) of total RNA isolated from transiently transfected cells. Expected migration patterns of transcripts cleaved at P1, P2 or unprocessed readthrough RNA (rt) are indicated on the left of each gel. Quantitation of independent RP's as an average percentage of total P1 readthrough transcripts is shown below. VA transfection control (VA) and construct specific undigested probe controls (+) are shown (right panel).

of 2 further substitutions caused a significant eleven fold increase in transcripts cleaved at P2 (Figure 17C).

Combining the four U to C substitutions in the DSE with hexamer mutations resulted in a 5-7 fold increase in P2 usage compared to the hexamer mutations alone, again highlighting the importance of the DSE for the recognition of the MC4R poly(A) site (Figure 18B). From the experiments described it was concluded that the MC4R DSE is the critical core sequence element for cleavage and polyadenylation and that this DSE does not require a canonical hexamer upstream of the cleavage site to direct efficient 3'end processing.

The A-rich upstream sequence and the AUUAAA are functionally redundant in the MC4R P1 poly(A) site

To test why the AUUAAA hexamer of the MC4R P1 poly(A) site can tolerate mutations without loss of 3'end cleavage efficiency, a series of additional plasmids were engineered where several mutations were introduced into the 23 nucleotides long upstream sequence previously mentioned in the 3'UTR sequence analysis section (Figure 19: Wt underlined sequence).

Analysis of this 23 nucleotide long sequence reveals that it is very A-rich and thus resembles somewhat the PE and NUE elements found in yeast and plant 3'end processing sites. As can be seen in the oligo-dT primed RT-PCR analysis presented in figure 19B (compare lanes 1-3), reporter plasmids containing extensive substitutions of adenosines with cytidines in this sequence show no effect on P1 usage. Interestingly, a clear shift from P1 to P2 usage is observed as soon as substantial mutations in the A-rich motif are combined with a point mutation in the AUUAAA hexamer (Figure 19B: lanes 4, 5). Furthermore, a construct containing hexamer mutations and where the adenosines in the A-rich

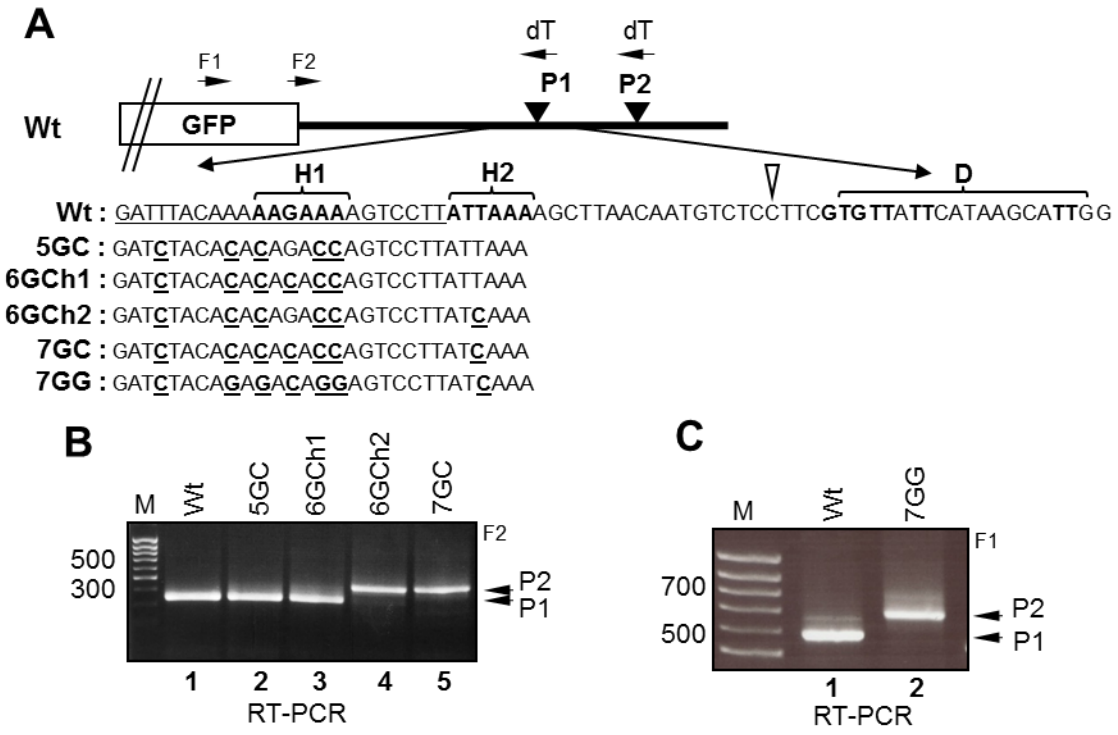


Figure 19: Mutations in the A-rich sequence and the hexamer are required to inactivate MC4R P1. **(A)** The diagram of the MC4R reporter gene and the wt sequence surrounding the P1 poly(A) site is shown, underlined letters represent 23 nucleotides long upstream sequence. The open triangle marks the site of cleavage at P1 and the DSE is indicated in bold. The changed nucleotides in each construct are shown in bold and underlined letters below the wt sequence. Forward primers (F1 and F2) and sites of potential reverse priming by oligo-dT at P1 and P2 respectively are shown above the diagram. **(B,C)** Qualitative oligo-dT primed RT-PCR analysis of total RNA isolated from HEK 293 cells transiently transfected with the wt and mutant plasmids. F1 or F2 use indicated on top right of gels.

motif were substituted by guanosines rather than cytidines (Figure 19C) could not rescue and restore efficient 3'end cleavage at P1. This observation strongly suggests that the MC4R poly(A) site contains two potential CPSF binding sites: an upstream A-rich motif and the AUUAAA hexamer sequence. 3'end cleavage and polyadenylation appears to be efficiently directed by either or both of these sequences which implies that an A-rich motif can functionally substitute for the hexamer sequence and direct cleavage and polyadenylation.

Following the A-rich sequence hypothesis enunciated above, it is worth noting that a possible molecular mechanism explaining how A-rich noncanonical poly(A) sites can be recognised by the poly(A) machinery has recently been described. The PAPOLG pre-mRNA lacks a canonical hexamer but instead contains a critical A-rich sequence upstream of the cleavage site. UGUAN sequence motifs located in the 3'UTR and the A-rich upstream sequence were shown to be critical for the recognition and function of this poly(A) site. The UGUAN motifs were shown *in vitro* to be recognised by CFI_m and this interaction facilitated the recruitment of CPSF and PAP to the pre-mRNA in the absence of a canonical hexamer[203].

Similar to PAPOLG pre-mRNA 3'end formation mechanism, the presence of four UGUAN sequences in the MC4R 3'UTR (for sequence see Materials and Methods, section 2.3.2., Figure 8B) could explain why the mutation of the AUUAAA hexamer was tolerated in the experiments described above. In order to address this question, G to C substitutions were introduced into the four UGUAN elements present in the MC4R 3'UTR both in the wild type and in the h1h2 reporter constructs (Figure 20A: U<u>C</u>UAN-Wt, U<u>C</u>UAN-h1h2). Importantly, the UCUAN sequence significantly reduces interaction with CFI_m *in vitro*[203]. Total RNA isolated from transiently transfected cells was analysed by oligo-dT primed RT-PCR. Although the experiments described in this section can not completely exclude the possibility that the introduced UGUAN elements mutations might not be sufficiently strong to inactivate them in the context of the MC4R 3'UTR despite having been positively tested in *in vitro* systems described above[203], the results of this analysis strongly suggest that mutations of the UGUAN sequence elements in the MC4R 3'UTR have no significant effects on MC4R 3'end processing efficiency

at P1 either in the presence, or the absence of a defined hexamer (Figure 20B: compare lanes 1-4: P1, P2).

From the observations described it was concluded that UGUAN elements present in the MC4R 3'UTR do not play a functional role in MC4R 3'end formation and that the MC4R poly(A) site only requires a strong DSE and an upstream A-rich region for efficient cleavage and polyadenylation.

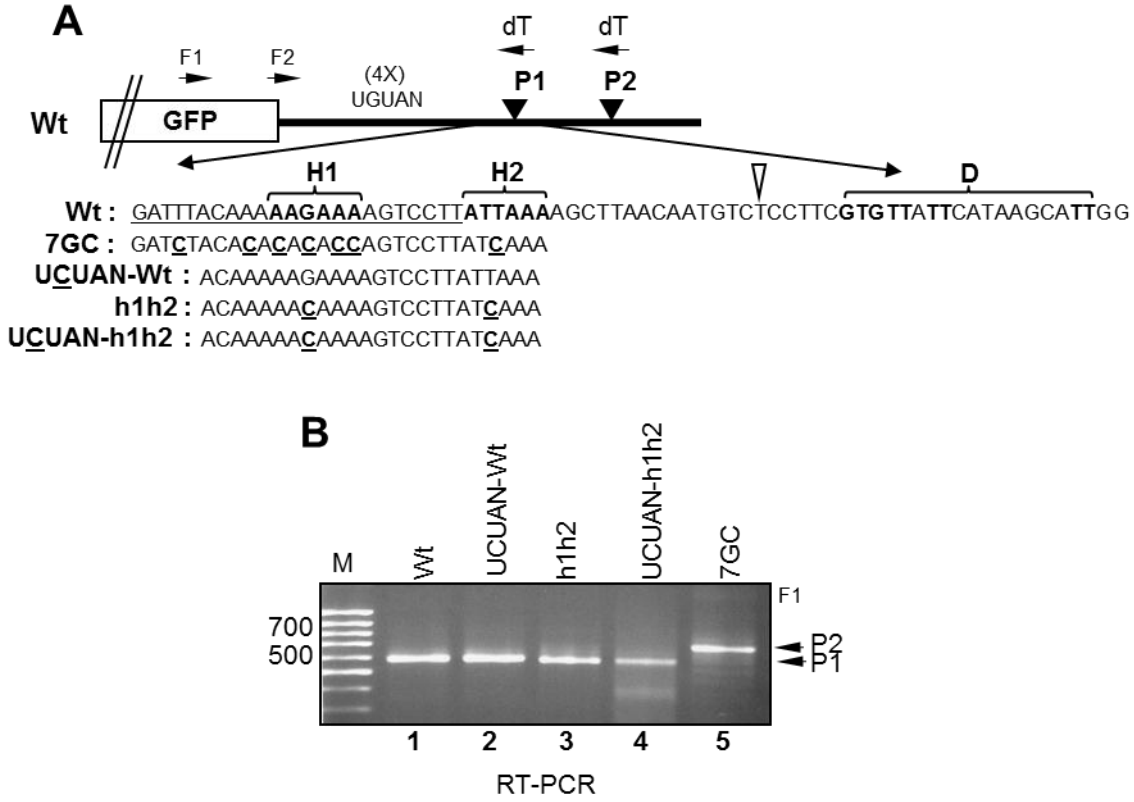
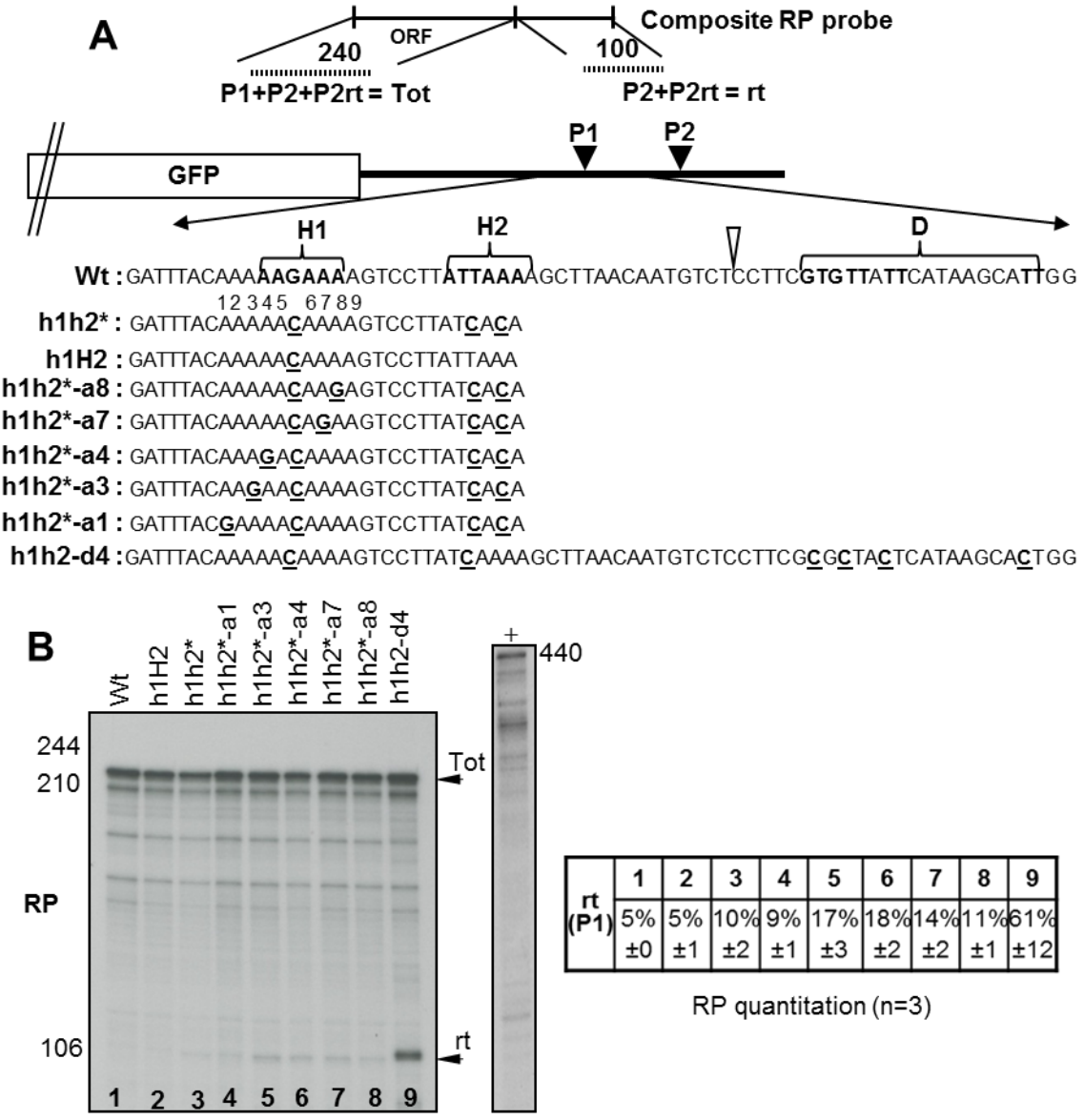


Figure 20: Mutations in the A-rich sequence and the hexamer are required to inactivate MC4R P1. **(A)** The diagram of the MC4R reporter gene and the wt sequence surrounding the P1 poly(A) site is shown, underlined letters represent 23 nucleotides long upstream sequence. The open triangle marks the site of cleavage at P1 and the DSE is indicated in bold. The changed nucleotides in each construct are shown in bold and underlined letters below the wt sequence. The four UGUAN motifs located in the 3'UTR are indicated by (4x) UGUAN in the diagram. For 3'UTR sequence details concerning UGUAN motifs refer to Figure 9B, Materials and Methods. Forward primers (F1 and F2) and sites of potential reverse priming by oligo-dT at P1 and P2 respectively are shown above the diagram. **(B)** Qualitative oligo-dT primed RT-PCR analysis of total RNA isolated from HEK 293 cells transiently transfected with the wt and mutant plasmids. F1 or F2 use indicated on top right of gels.

Efficient 3'end cleavage directed by a strong DSE downstream of a stretch of adenosines

The putative role of individual adenosines in the A-rich upstream region was further analysed by construction of additional reporter plasmids containing single A to G changes in this sequence in a double hexamer mutant (h1h2*) background (Figure 21A: h1h2*- a1-8). RP analysis of total RNA generated by HEK293 cells transfected with these constructs was conducted using a 'composite' antisense riboprobe that can be used for all mutant constructs because it does not contain regions in which sequence changes were introduced. The sequence of this antisense riboprobe corresponds to a direct fusion of the GFP ORF to the MC4R core poly(A) signal and 200nt of 3'flanking sequences (Figure 21A). This composite probe results in two protected bands, a 240 nucleotide band representing total reporter transcripts and a second 100 nucleotide long protected band that will only be present if transcripts escape cleavage at P1 and are subsequently either processed at P2 or readthrough P2 (Figure 21A). Quantitation results were calculated as: (a) $R1_{specific\ clone}$ = (rt/Tot), (b) $R2_{specific\ clone} = R1_{specific\ clone}/ R1_{Wt}$ (ratio of readthrough fold increase over Wt), and presented as a percentage based on Wt readthrough (as determined in Figure 17 and 18): (c) $R2_{specific\ clone}$ * 5% (Wt readthrough)= rt(P1).

The results obtained with this composite probe confirmed the earlier observations, where mutations of upstream hexamers in the MC4R poly(A) site had little effect on P1 usage (Figure 21B: compare lanes 1-3; Figure 22B: lanes 1, 2). Furthermore, mutations affecting the DSE or combined mutations affecting both the hexamers and the DSE simultaneously drastically reduced P1 usage which is demonstrated by the appearance of a second protected band



	1	2	3	4	5	6	7	8	9
rt (P1)	5% ±0	5% ±1	10% ±2	9% ±1	17% ±3	18% ±2	14% ±2	11% ±1	61% ±12

Figure 21: The MC4R DSE only requires an A-rich upstream sequence for efficient cleavage. **(A)** Diagram of the MC4R reporter is shown and the details are as in figure 20. The outlines of the composite RNAse protection probe are depicted above and the protected fragments are shown as dotted lines. All transcripts result in a 240nt protected band (P1+P2+P2rt) and transcripts not cleaved at P1 (P2+P2rt) give an additional protected band 100nt. The sequences surrounding P1 are shown below and the nucleotide substitutions for each clone are indicated in bold and underlined letters below the wt sequence. **(B)** RNAse protections of constructs with mutated core sequences and 17A-substitutions. The position of the protected bands is indicated by horizontal arrows at the right of the gels: total transcripts = Tot, transcripts not processed at P1 = rt. Quantitation of at least 3 independent RP's for each clone for each gel is given below the gels. Average percentage of the total transcripts that are not cleaved at P1 which, as determined in figure 17, is set to 5% for the wild type. Undigested probe control (+) is shown (right panel).

representing transcripts that are not cleaved at P1 (Figure 21B: compare lanes 1-8 with lane 9; 22B: lane 3, 4). Overall effects of mutating individual adenosines in the h1h2* background compared to h1h2-d4 clone were modest (Figure 21B: lanes 4-9). Interestingly, mutations of adenosines that disrupt the stretches of 5 or 4 consecutive adenosines closer to H1 (Figure 21A) consistently affected cleavage at P1 more severely than changing adenosines at the periphery of either stretch (Figure 21B: compare rt in lanes 4,8 with 5-7). From these results it was concluded that the adenosines in the A-rich region are unlikely to be part of a novel hexamer-like structure and that uninterrupted stretches of adenosines may be the most critical feature of this region.

To consistently address this hypothesis a reporter plasmid was created where both the A-rich region and the AUUAAA were replaced with a stretch of 17 adenosines. From this parental 17A plasmid two additional reporter constructs were created that contained G to C mutations inactivating the four UGUAN motifs present in the 3'UTR and/or with the d4 mutations inactivating the DSE. A control plasmid containing a stretch of cytidines instead of adenosines was also created (Figure 22A). The use of the composite probe for the analysis of the 17A plasmids was essential since this probe avoids an experimental artefact which can be observed with mutant specific antisense probes encompassing a long A-stretch. The U-rich sequence in such probes can hybridise with the poly(A) tails of potentially all mRNAs. This phenomenon can result in protected bands that mimic a cleavage event immediately upstream of the A-rich sequence (data not shown). A similar problem can arise from the use of oligo-dT primed RT-PCR, since miss-priming at the 17nt A-stretch by oligo-dT (17T) readily occurred and resulted in a slightly faster migrating band (Figure 22C: lane 2). Use of the composite probe, in

contrast, avoids these problems and allows accurate quantification of effects on cleavage at P1.

Analysis of total RNA isolated from cells transfected with the reporter plasmid containing the 17 adenosine substitution showed that an uninterrupted stretch of adenosines is able to direct cleavage at P1 at levels which were comparable to those observed with the wt plasmid. This is in stark contrast to effects on P1 cleavage levels observed with plasmids containing mutations in the DSE (Figure 22B: compare 'rt' in lanes 1,2,5 to lanes 3,4). Furthermore, when the A-stretch was substituted with a C-stretch or the 7GC mutations no cleavage is seen at P1 and all detectable transcripts represent mRNAs that are cleaved and polyadenylated at P2 (Figure 22C: lanes 1+2, 3+4).

In order to clarify that 3'end processing in the reporter plasmid containing the A-stretch is not dependent on the UGUAN sequences found in the 3'UTR, total RNA was analysed from cells transfected with corresponding constructs where the UGUAN motifs were changed to UCUAN (Figure 22A). The results of this analysis showed that a stretch of 17 uninterrupted adenosines positioned upstream of the MC4R DSE can direct efficient cleavage independently of the presence of upstream UGUAN sequences (Figure 23B: compare rt lanes 1-4). Contrastingly, A-stretch reporter constructs which contained mutations in the DSE resulted in a dramatic increase in transcripts that were not cleaved at P1 (Figure 23B: compare rt lanes 5-7). It is worth noting that the variability of the quantification increases notably when longer P2 cleaved transcripts are produced. It is likely that these longer mRNAs are less stable which is supported by the increased appearance of degradation products in the RP analysis as seen in figure 17C and 18B.

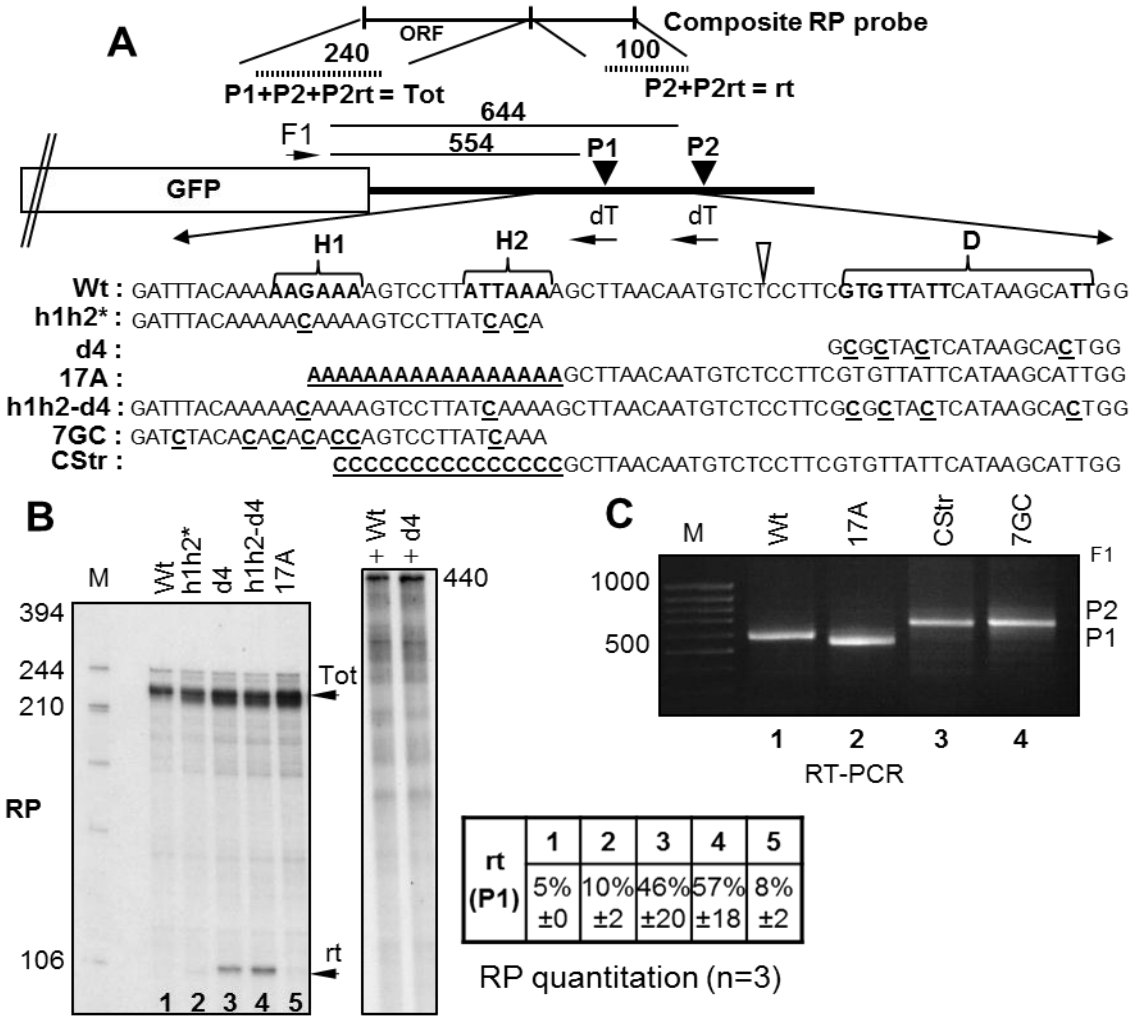


Figure 22: The MC4R DSE only requires an A-rich upstream sequence for efficient cleavage. **(A)** Diagram of the MC4R reporter is shown and the details are as in figure 20. The outlines of the composite RNAse protection probe is depicted above and the protected fragments are shown as dotted lines. All transcripts result in a 240nt protected band (P1+P2+P2rt) and transcripts not cleaved at P1 (P2+P2rt) give an additional protected band 100nt. The sequences surrounding P1 are shown below and the nucleotide substitutions for each clone are indicated in bold and underlined letters below the wt sequence. **(B)** RNAse protections of constructs with mutated core sequences and 17A-substitutions. The position of the protected bands is indicated by horizontal arrows at the right of the gels: total transcripts = Tot, transcripts not processed at P1 = rt. Quantitation of at least 3 independent RP's for each clone for each gel is given below the gels. Average percentage of the total transcripts that are not cleaved at P1 which, as determined in figure 2, is set to 5% for the wild type. Undigested probe controls (+) are shown (right panel). **(C)** Qualitative RT-PCR using an oligo-dT reverse primer. The analysis shows that oligo dT miss-priming at the A-stretch results in a shorter artefact band (lane 2) and that unlike a stretch of adenosines a stretch of cytidines is unable to direct cleavage and polyadenylation in the MC4R background.

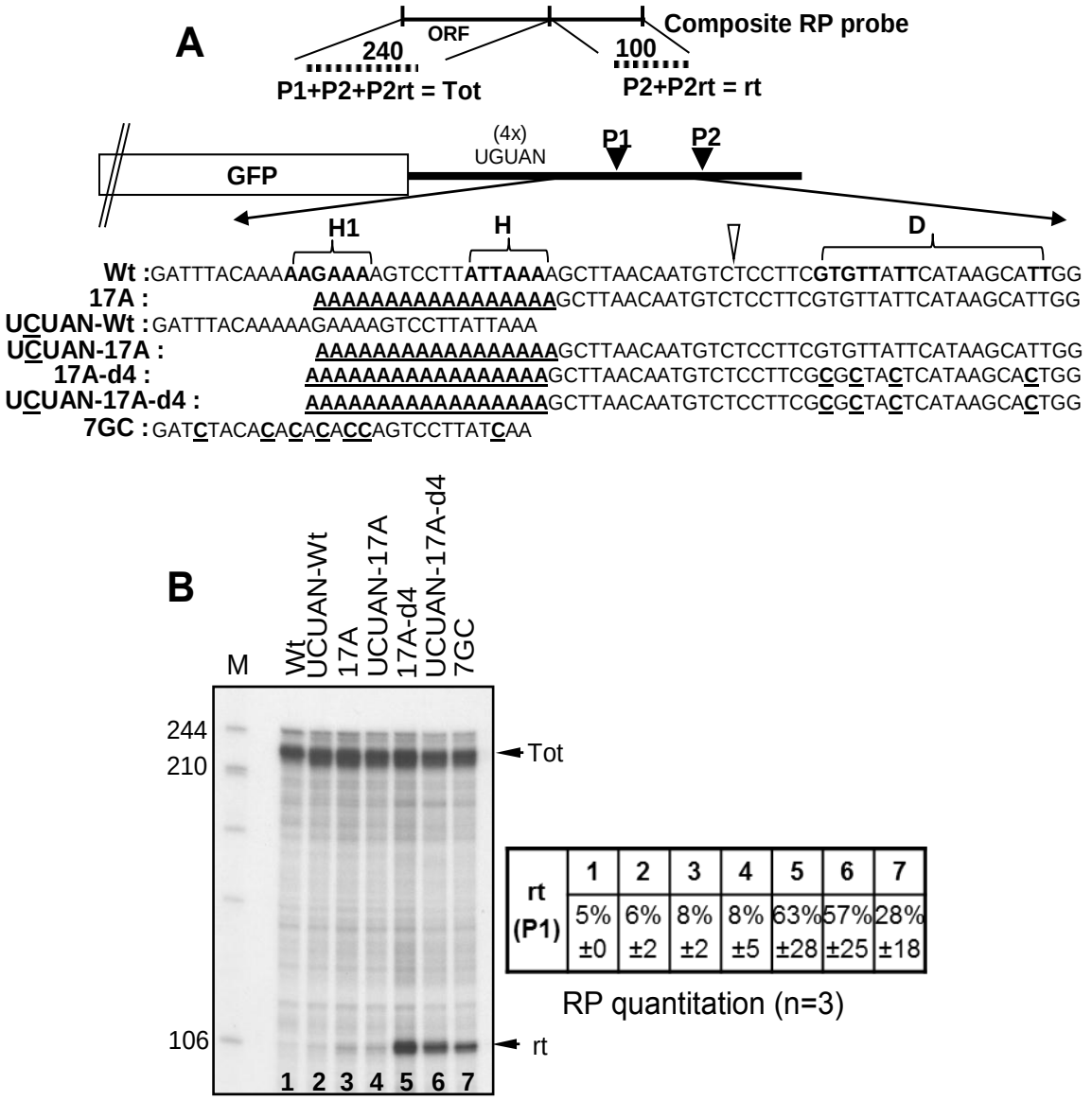


Figure 23: The MC4R DSE only requires an A-rich upstream sequence for efficient cleavage. **(A)** Diagram of the MC4R reporter is shown and the details are as in figure 3. The outlines of the composite RNAse protection probe is depicted above and the protected fragments are shown as dotted lines. All transcripts result in a 240nt protected band (P1+P2+P2rt) and transcripts not cleaved at P1 (P2+P2rt) give an additional protected band 100nt. The sequences surrounding P1 are shown below and the nucleotide substitutions for each clone are indicated in bold and underlined letters below the wt sequence. **(B)** RNAse protections of constructs with mutated core sequences and 17A-substitutions. The position of the protected bands are indicated by horizontal arrows at the right of the gels: total transcripts = Tot, transcripts not processed at P1 = rt. Quantitation of at least 3 independent RP's for each clone for each gel is given below the gels. Average percentage of the total transcripts that are not cleaved at P1 which, as determined in figure 2, is set to 5% for the wild type.

From these results it was concluded that the MC4R poly(A) signal is constituted by two distinct core upstream sequence elements, the ATTAAA hexamer and the region formed by the consecutive stretch of adenosines. These two regions share a degree of redundancy being able to, either cooperatively or independently, direct efficient 3'end formation in conjunction with the MC4R DSE. Furthermore the results presented in Figure 22, lane 5 and Figure 23, lane 3 clearly show that a strong DSE and a consecutive stretch of 17 adenosines are sufficient to direct efficient 3'end cleavage.

Note: The results presented in the next section correspond to work developed by Wencheng Li and Bin Tian from the Department of Biochemistry and Molecular Biology, UMDNJ-New Jersey Medical School. The results discussed above led to the formulation of the hypothesis that a particular subset of human genes might direct 3'end formation through core upstream sequence elements composed of consecutive stretches of adenosines possibly in conjunction with well defined DSEs. This hypothesis was suggested to Wencheng Li and Bin Tian as the parameters to carry the bioinformatics analysis described in the next section providing relevant contextualization, at the level of the human transcriptome, for the gene specific results presented in the previous section of the present dissertation.

The author of the present dissertation did not take part in the bioinformatics analysis.

3.3 Bioinformatics Analysis

Human poly(A) sites with A-rich upstream sequences have a higher frequency of downstream U-rich and GU-rich elements compared to 3'end processing sites constituting A(A/U)UAAA

The above described data implies that in the context of a strong DSE, human poly(A) sites may be less dependent on the presence of a A(A/U)UAAA canonical hexamer for its function. Hence, strong DSEs may be critical for the recognition of many noncanonical poly(A) sites. If this assumption were true, a significant amount of noncanonical poly(A) sites could be expected to contain A-rich upstream sequences and they should generally have stronger DSEs (defined as increased U or GU richness) compared to canonical poly(A) sites. To test this hypothesis, a genome-wide bioinformatics analysis was conducted using over 10,000 human poly(A) sites obtained from the PolyA_DB database[352]. Since A-rich sequences in a transcript can lead to internal priming for reverse transcription, resulting in false identification of poly(A) sites[353], it was required that supporting cDNA/EST/Trace sequences for a poly(A) site contained at least 30nt As/Ts corresponding to the poly(A) tail. As can be seen in figure 24A, DSEs in poly(A) sites that constitute an A-rich upstream sequence (defined as a hexamer with ≥ 5 adenosines but excluding AAUAAA and not overlapping with A(A/U)UAAA) have a significantly higher frequency of uridines in the +1 to +40 region compared to A(A/U)UAAA poly(A) sites. A more detailed analysis comparing the frequency of 4-mers in the DSEs shows a very strong bias (*P*-value of 1.2E-17) of UUUU and a significant bias of UGUU, a sequence element present in the MC4R DSE, towards A-rich sequences (Figure 24B). No correlation was found between the appearance of A-rich noncanonical poly(A) sites and intronless genes (data not shown).

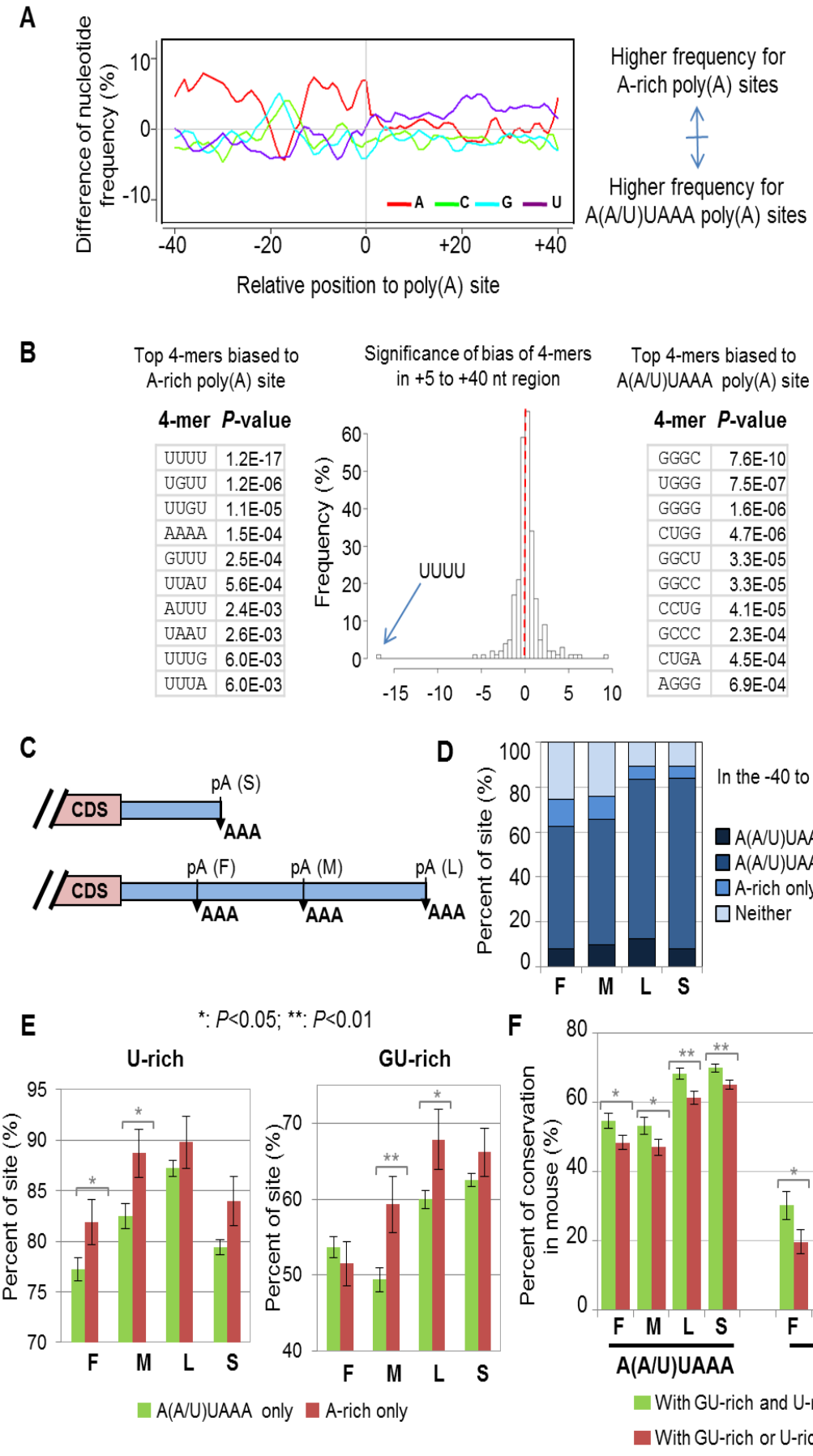

A
Difference of nucleotide frequency (%)
10
0
-10
A
C
G
U
-40
-20
0
+20
+40
Relative position to poly(A) site
Higher frequency for A-rich poly(A) sites
Higher frequency for A(A/U)UAAA poly(A) sites
B
Top 4-mers biased to A-rich poly(A) site
4-mer P-value
UUUU 1.2E-17
UGUU 1.2E-06
UUGU 1.1E-05
AAAA 1.5E-04
GUUU 2.5E-04
UUAU 5.6E-04
AUUU 2.4E-03
UAAU 2.6E-03
UUUG 6.0E-03
UUUA 6.0E-03
Significance of bias of 4-mers in +5 to +40 nt region
Frequency (%)
60
50
40
30
20
10
0
UUUU
-15
-10
-5
0
5
10
Top 4-mers biased to A(A/U)UAAA poly(A) site
4-mer P-value
GGGC 7.6E-10
UGGG 7.5E-07
GGGG 1.6E-06
CUGG 4.7E-06
GGCU 3.3E-05
GGCC 3.3E-05
CCUG 4.1E-05
GCCC 2.3E-04
CUGA 4.5E-04
AGGG 6.9E-04
C
CDS
pA (S)
AAA
CDS
pA (F)
pA (M)
pA (L)
AAA
AAA
AAA
D
Percent of site (%)
100
80
60
40
20
0
F
M
L
S
In the -40 to -10 region:
A(A/U)UAAA and A-rich
A(A/U)UAAA only
A-rich only
Neither
E
*: P<0.05; **: P<0.01
U-rich
GU-rich
Percent of site (%)
95
90
85
80
75
70
40
50
60
70
F
M
L
S
A(A/U)UAAA only
A-rich only
F
Percent of conservation in mouse (%)
80
60
40
20
0
A(A/U)UAAA
A-rich
With GU-rich and U-rich
With GU-rich or U-rich

Figure 24: Systematic analysis of poly(A) sites with A(A/U)UAAA and A-rich elements. Canonical PAS are AAUAAA or AUUAAA, and A-rich elements are hexamers containing at least 5 As excluding AAUAAA and not overlapping with A(A/U)UAAA. **(A)** Difference in nucleotide frequency surrounding poly(A) sites between A-rich poly(A) sites and A(A/U)UAAA poly(A) sites. **(B)** Significance of bias of 4-mers in the +5 to +40 nt region of A-rich and A(A/U)UAAA poly(A) sites. A Significance Score was calculated for each 4-mer based on its bias of occurrence in A-rich or A(A/U)UAAA poly(A) sites using Fisher's exact test (See Methods for detail). The Significance Score is -log(*P*-value) if the 4-mer is biased to A(A/U)UAAA poly(A) sites, or log(*P*-value) if biased to A-rich poly(A) sites. The distribution of Significance Scores are shown in a histogram. The top 10 4-mers significantly biased to A-rich poly(A) sites and to A(A/U)UAAA poly(A) sites are listed, together with their *P*-values. The most significant 4-mer, UUUU, is indicated in the histogram. **(C)** Schematics of single poly(A) sites (S); first (F), middle (M) and last (L) poly(A) sites in genes with alternative poly(A) sites located in the 3'-most exon. Poly(A) sites are indicated by arrows. CDS, coding sequence. **(D)** Percent of poly(A) sites with A(A/U)UAAA and/or A-rich elements in the -40 to -10 nt region for the 4 poly(A) site types shown in (C). **(E)** Percent of poly(A) sites with co-occurrence of A(A/U)UAAA or A-rich elements and U-rich (left) or GU-rich elements (right) for the 4 poly(A) site types. The U-rich or GU-rich sequence elements are described in Methods. The error bars are standard deviations. The differences in occurrence of U-rich or GU-rich sequence elements were evaluated by Fisher's exact test. Significant ones are indicated by one asterisk ($P<0.05$) or two asterisks ($P<0.01$). **(F)** Percent of poly(A) sites conserved in mouse with co-occurrence of A-rich elements only or A(A/U)UAAA only and downstream GU-rich and/or U-rich elements for the 4 poly(A) site types. The error bars are standard deviations.

Subsequently it was tested, if the correlation between U- and GU- richness and upstream A-stretches depends on its overall position within the pre-mRNA. For this the data was refined creating a distinction between genes that contain a single poly(A) site (S) and genes that undergo alternative cleavage and polyadenylation using multiple alternative poly(A) sites in the 3' UTR. The latter group was further separated into first (F), middle (M), and last (L) groups (Figure 24C) depending on the relative distance of the poly(A) sites to the beginning of the 3'-most exon[298]. The data shows that poly(A) sites constituting an A-rich upstream sequence represent about 5%-10% of the total poly(A) sites of each group, suggesting that upstream A-rich sequences represent a significant number of

functional noncanonical poly(A) sites. However, the correlation between A-rich sequences and an increased frequency of U and GU rich sequences appears to be more robust in genes containing alternative 3'end processing sites compared to genes with a single poly(A) site (Figure 24E: compare F, M in U-rich panel and M, L in the GU-rich panel with S). Interestingly, A-rich poly(A) sites that are flanked by two alternative processing sites, represented by group M, show a particular strong bias for both increased U and GU content in their DSEs (Figure 24E: M).

Further comparison of the conservation patterns of human A-rich and A(A/U)UAAA poly(A) sites in mouse was conducted. As shown in Figure 24F, A(A/U)UAAA poly(A) sites are more likely to be conserved than A-rich poly(A) sites for all poly(A) site types, i.e. S, F, M, and L, indicating higher evolutionary constrain on A(A/U)UAAA poly(A) sites and a general selection for A(A/U)UAAA signals. In addition, for most A-rich and A(A/U)UAAA poly(A) sites, those comprising both GU-rich and U-rich elements tend to be more conserved than those comprising either GU-rich or U-rich elements, indicating a possible evolutionary selection for DSEs.

Further analysis based on mRNA-seq data[140, 355] showed that 189 A-rich noncanonical poly(A) sites found in this data set behave identical to the 2960 canonical poly(A) sites regarding the ratios of reads positioned upstream or downstream of the cleavage site (Figure 25A). These findings provide strong evidence that the A-rich noncanonical poly(A) sites are true 3'end processing signals. Finally, the same data set also revealed that alternative A-rich noncanonical poly(A) sites are more likely to be tissue specifically regulated compared to A(A/U)UAAA alternative poly(A) sites (Figure 25B).

From this analysis it was concluded that poly(A) sites comprising upstream A-rich elements represent a significant number of noncanonical poly(A) sites and, compared to canonical poly(A) sites, generally have stronger DSEs. In addition, noncanonical A-rich poly(A) sites are more likely to be engaged in alternative polyadenylation and are more often subjected to tissue specific regulation compared to canonical poly(A) sites.

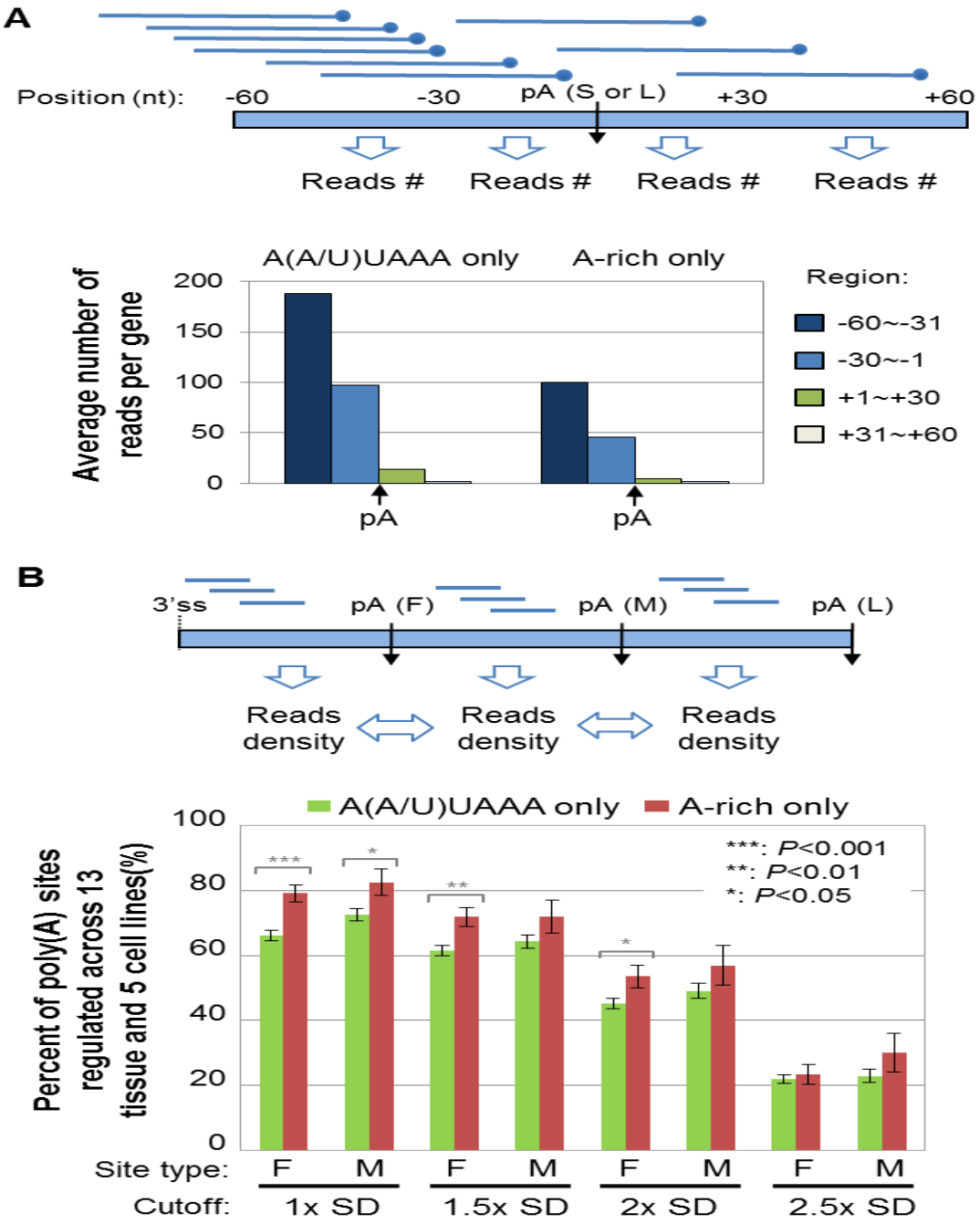


Figure 25: A-rich noncanonical poly(A) sites are true 3'end processing signals and are more likely to be subjected to alternative tissue specific cleavage and polyadenylation. **(A)** Analysis of poly(A) site usage using mRNA-seq data. Top, schematic of mapping sequencing reads to genomic regions surrounding poly(A) sites. Reads from 13 human tissues and 5 cell lines (see Methods for details) were mapped to the -60 to -1 nt upstream and +1 to +60 nt downstream regions of 3'-most poly(A) sites, i.e. S or L types. The data include 2,960 sites with A(A/U)UAAA and 184 with A-rich elements. The number of reads was based on the 3'end positions of the reads, indicated by dots in the graph. Reads were grouped into 4 regions, i.e. -60 to -31nt, -30 to -1nt, +1 to +30nt, +31 to +60nt. A drop of reads number in the downstream region indicates usage of poly(A) site. **(B)** Regulation of alternative poly(A) sites across tissues. (Top) Schematic of measuring usage of alternative poly(A) sites. Poly(A) sites are indicated by arrows. The usage of a poly(A) site is reflected by the Relative Usage of Downstream poly(A) site, or RUD score, which is the ratio of density of reads mapped to downstream region to density of reads mapped to upstream region (see Methods for details). Upstream and downstream regions correspond to the regions to the next upstream and downstream poly(A) sites, respectively, except for the first poly(A) site (F type), for which upstream region is defined by the 3'splice site (3'ss). Poly(A) sites with RUD scores above a cutoff in any one of the tissues or cell lines were considered regulated. We used 1x, 1.5x, 2x, and 2.5x standard deviation (SD) from mean as cutoffs. The percent of regulated poly(A) sites for each cut off is plotted. The error bars are SD of the percent of regulated poly(A) sites. The difference between A(A/U)UAAA and A-rich poly(A) sites with respect to percent of regulated poly(A) sites was calculated by the Fisher's exact test. Significant differences based on *P*-values are indicated by asterisks.

3.4. 3'End Formation sequence requirements in candidate genes identified through bioinformatics analysis

The bioinformatics analysis described above showed that about ⅓ of noncanonical poly(A) sites contain an A-rich sequence upstream of the cleavage site. In order to experimentally verify whether the A-rich sequences indicated by this analysis represent critical *cis*-elements for cleavage and polyadenylation, analysis of 3'end formation in the JunB and EDF1 genes, identified in our bioinformatics analysis, was conducted.

3.4.1. JUNB

The AP1 Transcription factors

Activator protein 1 (AP-1) is a transcription factor with a heterodimeric structure composed of proteins belonging to the c-Fos, c-Jun, ATF and JDP families[385]. AP-1 upregulates the transcription of genes containing the TPA DNA response element (TRE: 5'-TGAG/CTCA-3')[385]. AP-1 binds to this DNA sequence via a basic amino acid region, while the dimeric structure is formed by a leucine zipper[386]. It regulates gene expression in response to a variety of stimuli, including cytokines, growth factors, stress, and bacterial and viral infections[385]. AP-1 in turn controls a number of cellular processes including differentiation, proliferation, and apoptosis[387].

JUNB is an AP-1 transcription factor component and has been proposed to participate in the regulation of transcription from RNA polymerase II promoters, in cellular processes such as regulation of progression through cell cycle through DNA-dependent RNA polymerase II transcription. Encoded protein products are expected to have molecular functions such as RNA polymerase II transcription coactivator activity, RNA polymerase II transcription corepressor activity, protein dimerization activity and to localize in various compartments (chromatin, nucleus).

3.4.1.1 Results - JUNB 3'end formation sequence requirements

JUNB 3'end processing site: an endogenous noncanonical poly(A) site with an upstream A-rich sequence required for optimal cleavage

JUNB encodes a non-spliced transcript with a poly(A) cleavage site annotated 22 nucleotides downstream of an A-rich sequence in the 3'UTR. This annotation is verified by 177 AceView accessions.

In order to investigate the functional relevance of this endogenous A-stretch in poly(A) cleavage, 3'UTR and 3'flanking sequences including the DSE of JUNB were cloned downstream of the GFP ORF in the MC4R reporter plasmid. MC4R 3'UTR and the P1 poly(A) sequences were replaced with equivalent regions from JUNB (Figure 26A). Regions downstream of MC4R P1 including the P2 poly(A) site are retained in the constructs. Therefore, potential transcripts that are not cleaved and readthrough the JUNB poly(A) site are stabilised by cleavage and polyadenylation at the MC4R P2 site and can therefore easily be detected using the composite RNAse protection probe described in figure 23.

RP analysis was used to confirm the site of cleavage in the JUNB wild type plasmid. As can be seen in figure 26B cleavage of transfected JUNB resulted in RP products that are slightly longer than the expected lengths (123nt) which may be due to a slight shift of the cleavage site in the reporter construct. Nevertheless, in the wt JUNB clone cleavage occurs 5-10nts downstream of the annotated site 3' of the A-stretch. No endogenous JUNB mRNA was detected by RP analysis.

Subsequent RP analysis with the composite probe described previously of JUNB wild type and mutant plasmids showed that A to C mutations in the A-stretch significantly reduced cleavage efficiency at the JUNB 3'end processing site. This is manifested by the appearance of the 88 nucleotide band representing transcripts that are not cleaved at the JUNB 3'end processing site (Figure 26C: compare lanes 1, 2). In these experiments, because JUNB constructs do not contain the MC4R P1 poly(A) signals, the resulting protected readthrough band is 88 rather than 100 nucleotides long. A reduction in JUNB poly(A) site use was also observed when U to C mutations were introduced into the DSE (Figure 26C: compare lanes 1, 3). Effects of DSE mutations introduced in the JUNB reporter

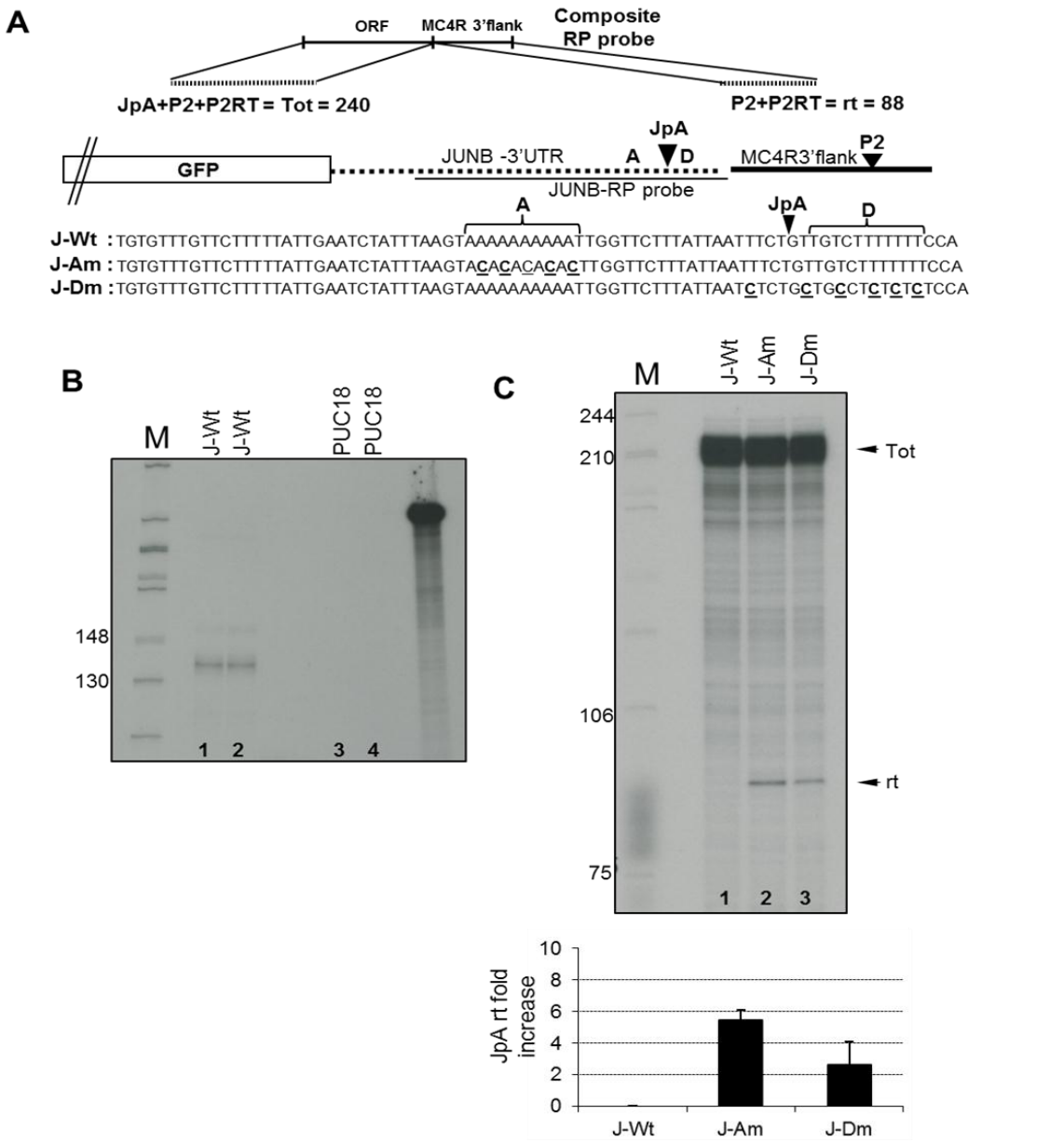


Figure 26: The JUNB pre-mRNAs requires an A-rich upstream sequence for efficient cleavage and polyadenylation. **(A)** Diagram showing the JUNB reporter gene. Origins of the sequences in the plasmid are indicated. JUNB 3'UTR and 3'flanking regions are represented by a dotted line and the graph shows how JUNB sequences are inserted into the MC4R background. The position of the JUNB A-rich region (A), DSE (D) is indicated and the poly(A) sites are represented by JpA and P2 respectively. Lengths of protected RP bands are shown above the graph. All transcripts result in a 240 nucleotide protected band (Tot) and transcripts not cleaved at JpA (rt) give an additional protected band 88 nucleotides in length. The JUNB specific probe used in B is shown as thin black line below the dotted line. The sequences surrounding the JUNB (JpA) cleavage site are shown below the graph and the nucleotide substitutions for each clone are indicated in bold and underlined letters below the wt sequences (J-Wt). (**B, C**) RNAse protection of total RNA isolated from cells transfected with JUNB wt and mutant plasmids. Quantitation is presented as fold increase of transcripts that are not cleaved at J-pA.

construct were less dramatic when compared to mutations of the A-rich sequence, a twofold effect versus a 5 fold effect, which is likely to be due to the presence of a G/U rich putative USE (Figure 26A). These results confirmed that the endogenous A-stretch found in the JUNB sequence is a critical *cis*-element for efficient 3'end processing supporting the hypothesis that A-rich elements may play a critical role in the recognition of a significant number of human noncanonical poly(A) sites.

3.4.2. Endothelial Differentiation-related Factor 1 (EDF1)

The bioinformatics analysis described above shows that about ⅓ of noncanonical poly(A) sites contain an A-rich sequence upstream of the cleavage site. Experimental analysis of the MC4R and JUNB 3'end processing sequence requirements showed that these A-rich sequences have a critical role in the definition of the 3'end processing site in both of these intronless transcripts. To test whether these A-rich sequences also represent critical *cis*-elements for cleavage and polyadenylation in the context of a spliced gene, analysis of 3'end formation was conducted in the EDF1 gene, identified in the above mentioned bioinformatics analysis.

This transcript encodes a protein that may regulate endothelial cell differentiation. It has been postulated that the protein functions as a bridging molecule that interconnects regulatory proteins and the basal transcriptional machinery, thereby modulating the transcription of genes involved in endothelial differentiation. This protein has also been found to act as a transcriptional coactivator by interconnecting the general transcription factor TATA element-binding protein (TBP) and gene-specific activators.

EDF1 has been proposed to participate in processes such as endothelial cell differentiation, multicellular organismal development, positive regulation of DNA binding, regulation of lipid metabolic process and regulation of transcription. Proteins are expected to have molecular functions (histone acetyltransferase activity, methyltransferase activity, transcription coactivator activity, transcription factor activity) and to localize in various compartments (cytoplasm, intracellular, nucleus, transcription factor TFIID complex).

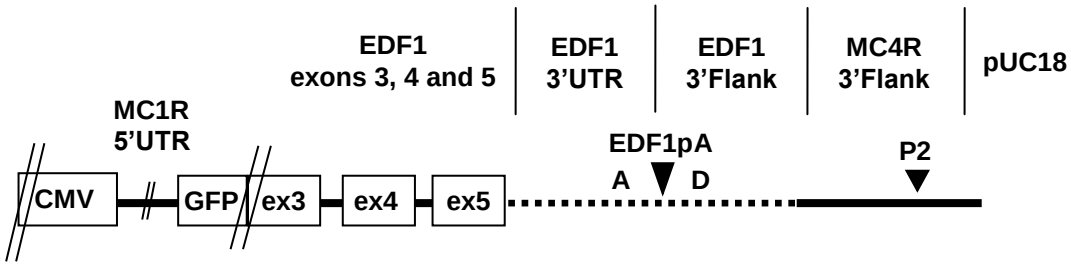


Figure 27: Schematic representation of the EDF1 Wt plasmid

3.4.2.1. Results – EDF1 3'end formation sequence requirements

An endogenous noncanonical poly(A) site with an upstream A-rich sequence element in a spliced environment

EDF1 encodes two validated alternatively spliced transcripts with a poly(A) cleavage site annotated 13-17nt downstream of an A-rich sequence within the EDF1 3'UTR. This annotation is verified by 115 AceView accessions.

To examine the role of the endogenous EDF1 A-stretch in poly(A) cleavage of a spliced gene, 129nt of EDF1 exon3, intron3, exon4, intron4, exon5, the 3'UTR and 3'flanking sequences up to 250nt downstream of the cleavage site were cloned downstream of a GFP ORF from which the stop codon was deleted in the MC4R plasmid. Thus, the MC4R 3'UTR and the P1 poly(A) sequences were replaced with the above described EDF1 regions (Figure 27 and Materials and

Methods, section 2.3.2., Figure 11 for cloning details). Note that the regions downstream of MC4R P1 including the P2 poly(A) site are retained in the construct. Therefore, potential transcripts that are not cleaved and readthrough the EDF1 poly(A) site are stabilised by cleavage and polyadenylation at the MC4R P2 site and can therefore easily be detected using either a EDF1 specific RNAse protection composite probe or the composite RNase protection probe described in the MC4R results section.

RNase protection analysis was used to confirm the site of cleavage in the EDF1 wild type plasmid using an EDF1 specific probe which contains 158nt of the GFP ORF fused to 65nt of intron4 all of exon5 (62nt), 3'UTR and 3'flanking sequences surrounding the EDF1 pA (see figure 28A). As can be seen in figure 31B the EDF1 pA is functional (Figure 28B: Lane 1, spliced EDF1 pA) generating RP products of the expected length (228nt). As can also be seen in figure 28B lane 1 there is a strong band corresponding to a non-spliced readthrough band (Figure 28B: Lane1, non-spliced rt) which is probably caused by lack of efficiency of the EDF1 3'end processing site. A strong candidate hypothesis for the explanation of this observation is that, in the context of the EDF1 reporter plasmid, the splicing reaction, despite containing both the penultimate and the terminal intron, is incomplete (absence of exon1, intron1, exon2, intron2 and part of exon3). Since interactions between components of the splicing machinery directing terminal intron removal and the cleavage and polyadenylation apparatus result in reciprocal enhancement of both processing reactions[329-331, 335] it is therefore plausible to suppose that, in the context of the EDF1 reporter plasmid, an incomplete upstream splicing reaction will cause a strong loss of efficiency of the downstream 3'end processing reaction, thus generating a strong readthrough

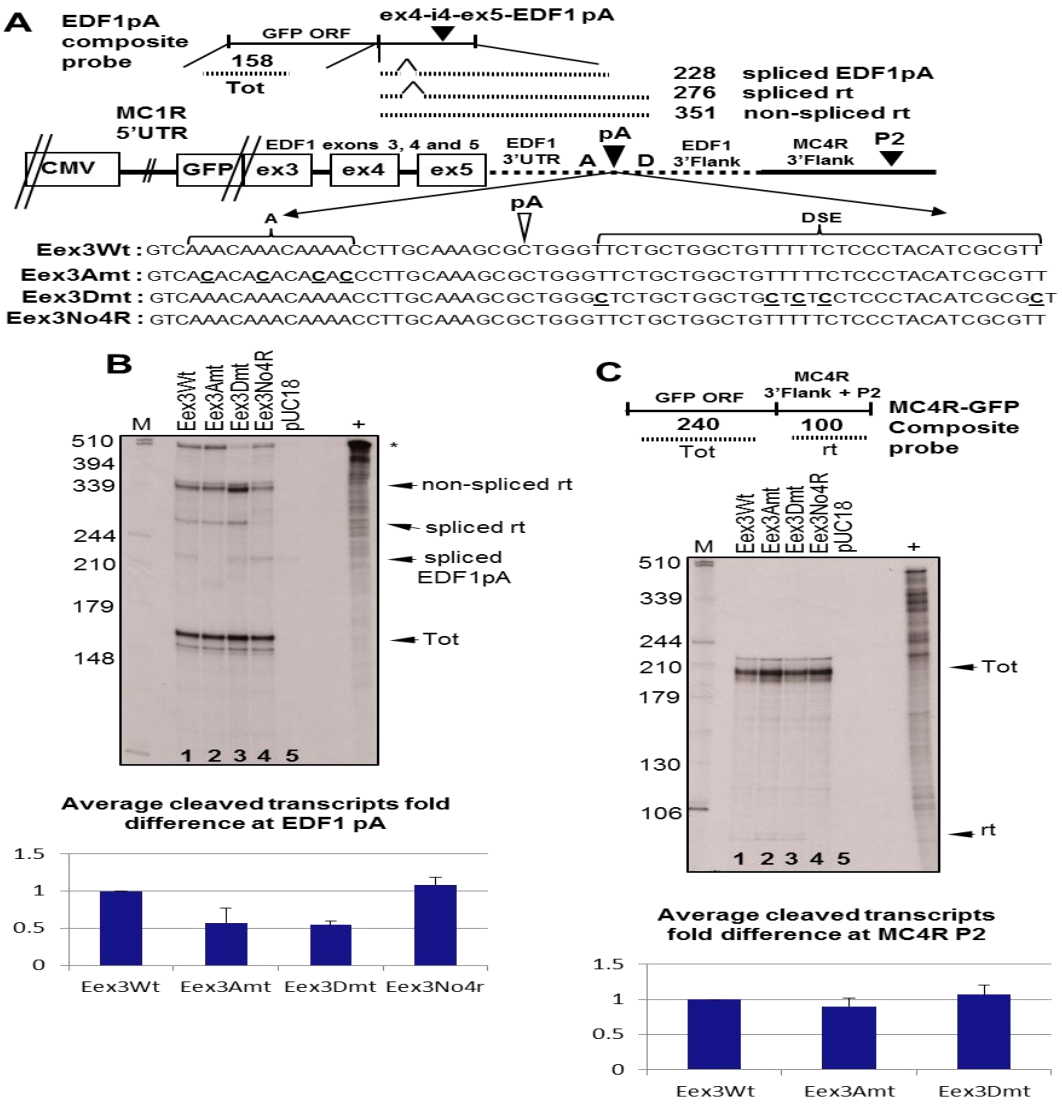


Figure 28: The intron-containing EDF1 gene requires an A-rich upstream sequence for efficient cleavage and polyadenylation. **(A)** Diagram showing the EDF1 reporter gene. Origins of the sequences in the plasmid are indicated. EDF1 exon3, intron3, exon4, intron4, exon5, 3'UTR and 3'flanking regions are represented by open boxes and a dotted line respectively. The graph shows how EDF1 sequences are inserted into the reporter plasmid backbone. The position of the EDF1 A-rich region (A) and DSE are indicated and the poly(A) sites are represented by pA and P2 respectively. Lengths of protected RP bands are shown. All RPs (n=3). All transcripts result in a 158nt protected band (Tot) and transcripts cleaved at pA (spliced EDF1 pA) give an additional protected band 228nt in length. The EDF1 specific probe used in B is shown as a black line above the diagram. The sequences surrounding the EDF1 cleavage site (pA) are shown below the graph and the nucleotide substitutions for each clone are indicated in bold and underlined letters below the wt sequences (Eex3Wt). **(B)** RNAse protection using the EDF1 specific probe of total RNA isolated from cells transfected with EDF1 wt and mutant plasmids. **(C)** RNAse protection using the EDF1 specific probe of total RNA isolated from cells transfected with EDF1 wt and mutant plasmids. Quantitation is presented as fold difference of transcripts that are not cleaved at pA normalized to the total number of transcripts (Tot).

from the EDF1 pA. Nevertheless, in the Wt EDF1 clone (Eex3Wt) and despite a less efficient 3'end processing reaction, cleavage occurs at the annotated site 3' of the endogenous A-stretch.

RP analysis of EDF1 wild type and mutant plasmids with the EDF1 composite probe showed that A to C mutations in the A-rich sequence strongly reduced cleavage at the EDF1 3'end processing site. As can be seen in figure 28A introduction of 4 A to C substitutions reduces cleavage efficiency at the EDF1 pA by ~50% compared to the Wt. This is noticeable by the loss of the 228nt band representing transcripts are cleaved at the EDF1 poly(A) site (Figure 28B: compare lanes 1, 2, spliced EDF1 pA). A similar reduction (~50%) of EDF1 poly(A) site use was also observed when U to C mutations were introduced into the DSE (Figure 28B: compare lanes 1, 3).

The results described confirmed that the endogenous A-rich sequence found upstream of the EDF1 3'end processing site is a critical *cis*-element for efficient 3'end processing also in a spliced gene environment. A-rich elements, as described for intronless genes, have also a critical role in the recognition of human noncanonical poly(A) sites in spliced gene environments.

3.5. JUND

Further to the genes identified by the bioinformatics results, a third intronless gene was chosen for analysis based on the facts that it presented two alternative 3'end formation sites with the most 5' one being defined by a non-canonical hexamer AGTAAA and the most 3' not showing a clearly identifiable upstream core sequence element.

The protein encoded by this intronless gene is a member of the JUN family, and a functional component of the AP1 transcription factor complex. It has been proposed to protect cells from p53-dependent senescence and apoptosis, to participate in pathways (FOSB gene expression and drug abuse, MAPK signalling pathway) and processes such as regulation of transcription from RNA polymerase II promoters. Encoded protein products are expected to have molecular functions such as RNA polymerase II transcription factor activity, protein dimerization activity, sequence-specific DNA binding and to localize in various chromatin and nuclear compartments.

3.5.1 Results - JUND 3'end formation sequence requirements

JUND 3'end processing is directed by an endogenous noncanonical hexamer (AGUAAA) poly(A) site

JUND encodes a non-spliced transcript with 2 annotated poly(A) cleavage sites. The upstream one is defined by a cleavage site annotated 21 nucleotides downstream of a non-canonical putative AGUAAA hexamer and is based on 8 Aceview acessions. The downstream one is annotated 1376nt downstream of the JUND stop codon, has no clearly identifiable canonical 3'end processing core elements and is supported by 53 Aceview acessions. To confirm the described annotations of the poly(A) cleavage sites, the 3'UTR and 3'flanking sequences

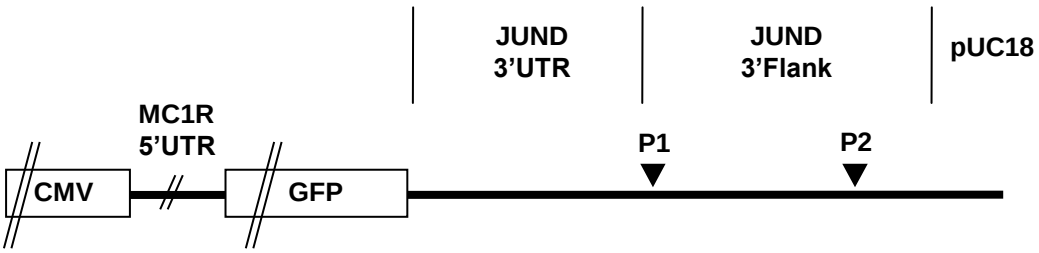


Figure 29: Schematic representation of the JUND Wt plasmid

including up to 2376nt downstream of the JUND stop codon were cloned downstream of the GFP ORF in the above described JUNB original plasmid. The JUNB 3'UTR and the A-rich poly(A) sequences were replaced with equivalent regions from JUND (Figure 30A and Materials and Methods, section 2.3.2., Figure 10 for cloning details).

In order to verify the above described annotated cleavage sites, two RP probes were used containing GFP ORF sequences fused to either the Wt sequences surrounding P1 or the Wt sequences surrounding the JUND P2 (Figure 30A). As can be seen in figure 30B, lane 1, P1 cleavage of transfected JUND is readily detectable at P1 resulting in a 190nt product which is slightly shorter than the expected length (196nt) which may be due to a slight shift of the cleavage site in the JUND Wt reporter construct. Contrastingly, as can be seen in figure 31B, lane1, at the P2 site no cleavage product can be detected at the JUND annotated P2 site which might be due to either a cell type specific 3'end formation event which cannot be reproduced by the HEK293 cell line that is used for the analysed transfections or might be due to long range primary sequence interactions that are not reproduced in our reporter constructs which harbour only 2kb of the JUND 3'UTR and 3'flanking sequences. No endogenous cleaved JUND transcripts could be detected by RP at any of the 3'end annotated sites as can be seen in figure 30B, lane 6 and in figure 31B, lane 6 - pUC18 control.

Having established the functionality of the P1 cleavage site in the JUND Wt construct subsequent RP analysis were conducted in order to study JUND P1 wild type and mutant plasmids. As can be seen in figure 30B, with the JUND P1 composite probe, T to C mutations in the T-stretchs located either ~50-80nt upstream of the putative AGUAAA (denominated as Far Upstream Sequence

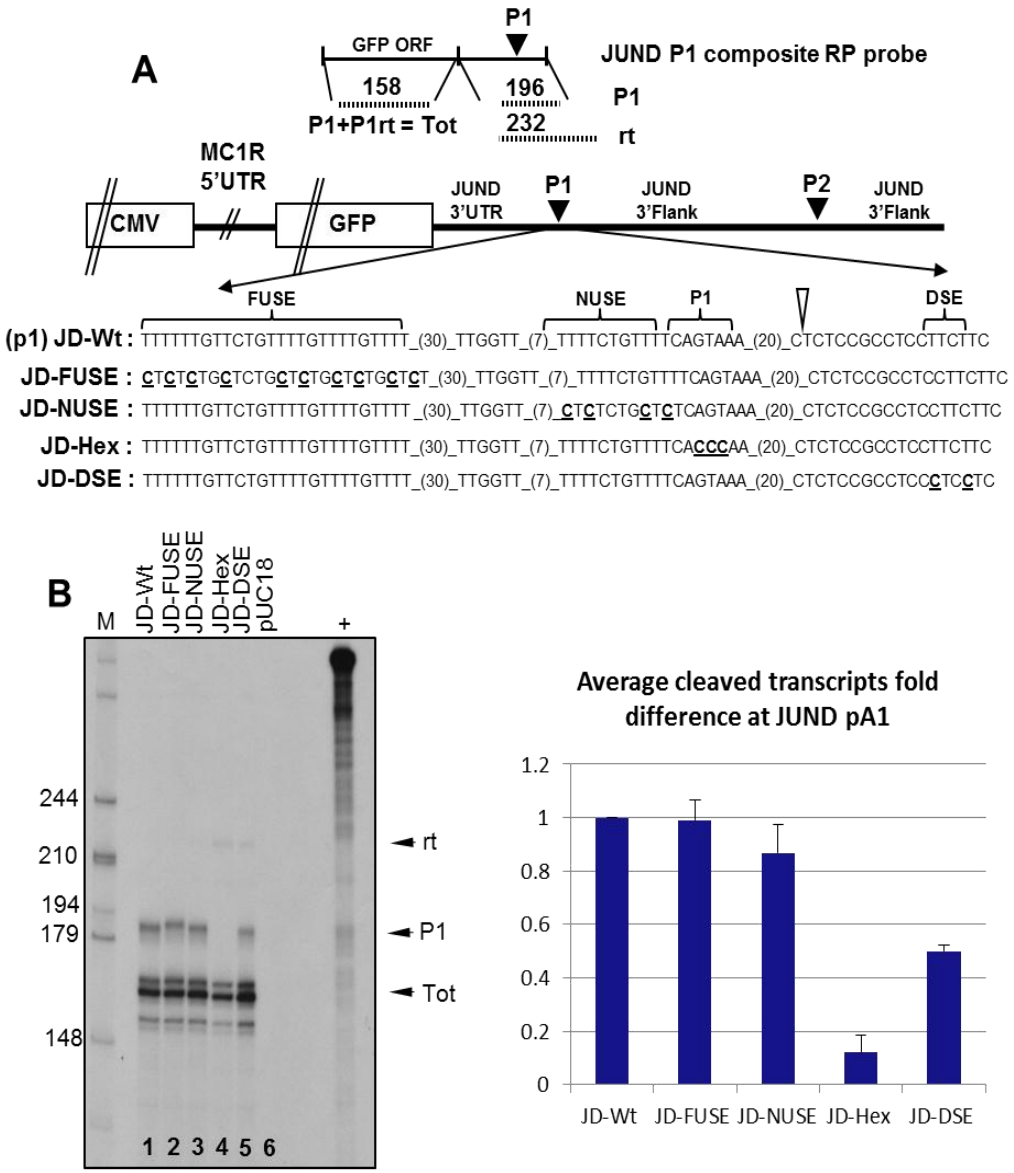
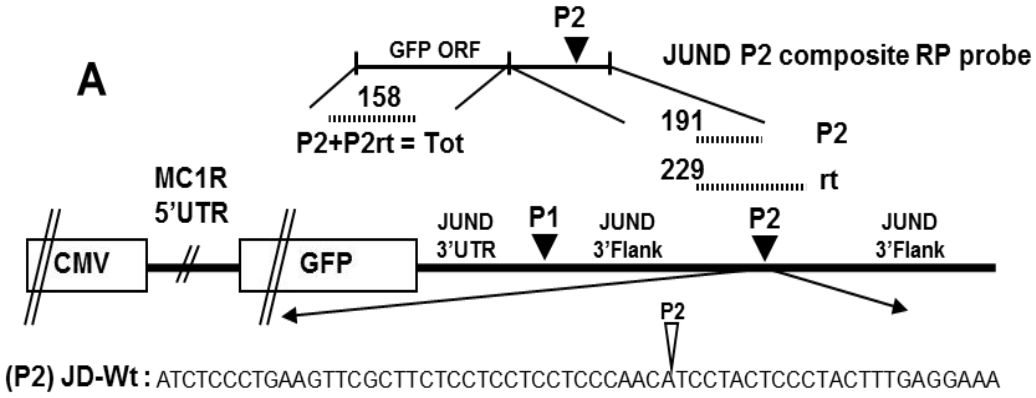


Figure 30: The JUND pre-mRNAs rely on a noncanonical AGUAAA hexamer for efficient cleavage and polyadenylation. **(A)** Diagram showing the JUND reporter gene. Origins of the sequences in the plasmid are indicated. JUND 3'UTR and 3'flanking regions are indicated and the graph shows how JUND sequences are inserted into the reporter backbone. The position of the JUND T-rich regions (FUSE, NUSE), hexamer (P1) and DSE are indicated and the annotated poly(A) sites are marked with closed triangles and P1 and P2 respectively. Lengths of protected RP bands are shown above the graph. All RPs (n=3). The JUND specific probe used is depicted above the reporter diagram. The sequences surrounding the JUND cleavage site 1 (P1) are shown below the graph and the nucleotide substitutions for each clone are indicated in bold and underlined letters below the wt sequences (JD-Wt). **(B)** RNAse protection with JUND P1 specific probe of total RNA isolated from cells transfected with JUND wt and mutant plasmids. Quantitation is presented as fold difference of transcripts that are cleaved at P1 normalized to total number of transcripts (Tot).



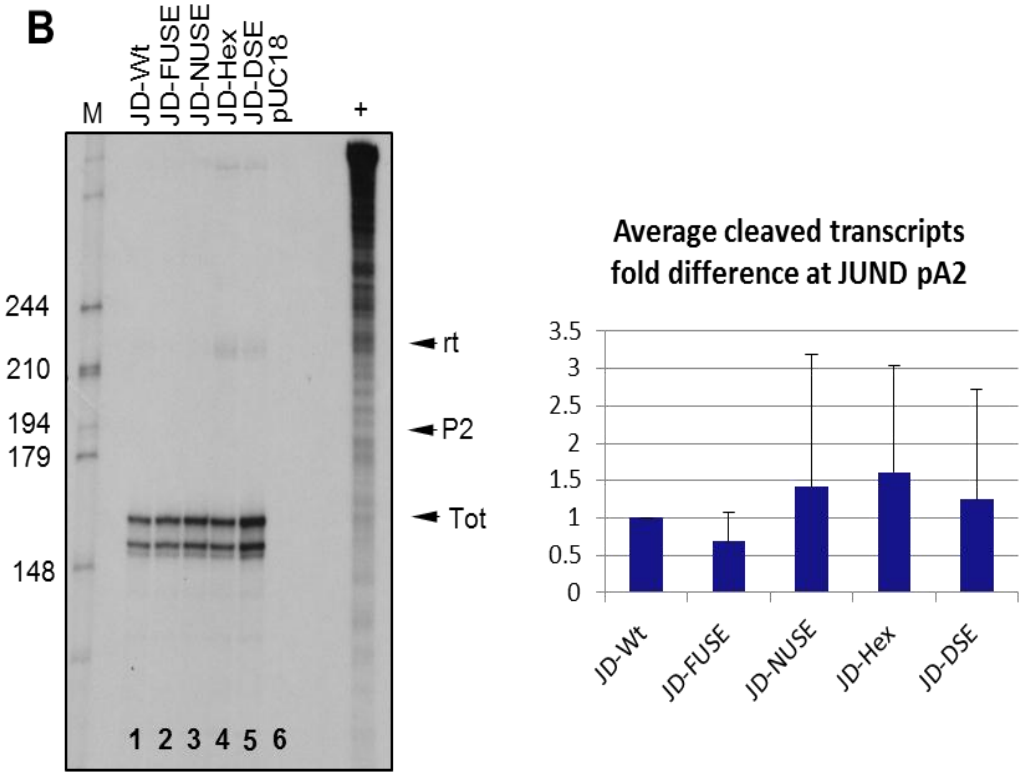


Figure 31: The annotated JUND P2 is not functional in rely on a noncanonical AGUAAA hexamer neither the JUND P1 Wt or mutant backgrounds. **(A)** Diagram showing the JUND reporter gene. Origins of the sequences in the plasmid are indicated. JUND 3'UTR and 3'flanking regions are indicated and the graph shows how JUND sequences are inserted into the reporter backbone. The position of the JUND annotated P2 is indicated and the annotated poly(A) sites is marked with closed triangles and P1 and P2 respectively. Lengths of protected RP bands are shown above the graph. All RPs (n=3). The JUND specific probe used is depicted above the reporter diagram. The sequences surrounding the JUND cleavage site 2 (P2) are shown below the graph. **(B)** RNAse protection with JUND P2 specific probe of total RNA isolated from cells transfected with JUND wt and mutant plasmids. Quantitation is presented as fold difference of transcripts that are cleaved at P2 normalized to total number of transcripts (Tot).

Element- FUSE) or ~ 10nt upstream of the AGUAAA (Near Upstream Sequence Element- NUSE) had no significant effect in the level of cleaved transcripts at the JUND P1 (Figure 30B, lanes 1-3). Introduction of three point mutations in the AGUAAA (see figure 30A for sequence) reduced cleavage by ~9 fold and introduction of two T to C substitutions in the JUND P1 DSE, minimally defined by two double uridine elements 12-15nt downstream, reduced cleavage by ~2.5 fold (Figure 30B, lanes 1, 4, 5 respectively).

All the described P1 reporter plasmids were also simultaneously analysed by RP using the JUND P2 composite probe to investigate the possibility of activation of the P2 site in the absence of a functional upstream P1 3'end processing site. The results obtained and shown in figure 31B seem to point towards a negative result but must be considered essentially inconclusive due to the very high associated experimental error which results, probably, from very low RP signals across the P2 site which can be strongly distorted even by small variations in the detected signals. The results described above indicate that despite the presence of a non-canonical hexamer AGUAAA the JUND P1 site follows a classical Hexamer-DSE dependent mechanism for the definition of the P1 3'end processing site. The annotated JUND P2 3'end processing site did not generate accurate and reliable signals from our JUND reporter plasmids.

3.6. Melanocortin 1 Receptor (MC1R)

The intronless MC1R gene encodes for a 317 amino acid seven transmembrane receptor expressed on the cell surface of melanocytes. MC1R plays a key role in the formation of skin and hair pigmentation by regulating the

relative ratio of the two major classes of melanin, eumelanin (brown/black) and pheomelanin (red/yellow), upon stimulation or antagonism by soluble ligands such as α-MSH or the agouti protein respectively. Exposure of skin to ultraviolet radiation leads to increased signal transduction through MC1R resulting in elevated eumelanin biosynthesis and a subsequent darkening of the skin[388].

The author of the present dissertation developed collaborative work in the analysis of the intronless MC1R gene 3'end formation mechanism. The MC1R gene was shown to rely on a 3'end formation mechanism which involves a canonical hexamer (AAUAAA) and the recognition of a noncanonical DSE with involvement of the heterogeneous ribonuclear protein H (hnRNPH). The bandshift experiments that showed the involvement of hnRNPH in this recognition mechanism were established by the author of the present dissertation and constitute peer reviewed published material: Dalziel, M., Nunes, N.M. & Furger, A. Two G-rich regulatory elements located adjacent to and 440 nucleotides downstream of the core poly(A) site of the intronless melanocortin receptor 1 gene are critical for efficient 3' end processing. *Mol Cell Biol* **27**, 1568-80 (2007). (see Appendix).

Chapter 4

Discussion

Discussion

Terminal intron removal and cleavage and polyadenylation have long been known to be tightly interlinked and mutually enhance each other's efficiency [327, 389]. Since removal of terminal introns in spliced genes generally results in a significant inhibition of 3'end processing, it is unclear how naturally intronless pre-mRNAs can be efficiently processed at their 3'end.

The work described in the present dissertation investigated what are the primary sequence requirements controlling 3'end formation in the single exon genes MC4R, JUNB and JUND and the multi-exon gene EDF1 pre-mRNAs (Note: MC1R single exon gene results were produced in a collaborative work and therefore are only presented and discussed as a comparative example where relevant). Contrastingly with previously analysed and described viral and human intronless gene transcripts[234, 245, 345, 346], poly(A) site recognition of MC4R is not dependent on additional auxiliary sequences located either upstream or downstream of the core poly(A) site (Figure 15 and 16). Furthermore, 3'end processing of this pre-mRNA is solely dependent on the core poly(A) sequences. Comparatively MC1R contains two 3'flank G-rich elements that are critical for efficient 3'end formation, closer to the various examples reported previously[245, 345, 346]. It should be noted that JUNB and JUND 3'UTR and flanking sequences were not tested for the presence of auxiliary sequence elements (JUND upstream T-rich elements were targeted by mutational analysis not producing any significant results (Figure 30), but a thorough deletion analysis was not conducted) and these experiments would constitute a relevant addition to the analysis initiated within the scope of the present dissertation.

Mutational analysis revealed that poly(A) cleavage at the MC4R and JUNB processing sites does not require canonical hexamers and that in the former example it is the short DSE the most critical *cis*-element (Figures 17, 22 and 23) while in the latter it is the A-rich upstream core sequence the most relevant core sequence element in the definition of efficient 3'end formation. Interestingly these data questions the current understanding of what constitutes a functional poly(A) site but also brings forth new information contributing to explain how noncanonical poly(A) sites can be recognised and efficiently processed in the absence of A(A/U)UAAA hexamers.

About 70%-80% of human core poly(A) signals are defined by the two canonical hexamer sequences AAUAAA or AUUAAA located upstream of a generally loosely defined U or G/U rich downstream sequence element [220, 221, 250, 298]. The co-transcriptional recognition of these canonical poly(A) sites has been extensively investigated. Numerous experiments have established that the hexamers are intolerant to sequence alterations in that single point mutations in the A(A/U)UAAA result in a dramatic inhibition of 3'end processing efficiency. Conversely, point mutations and even small deletions in the DSE are considered as being generally well tolerated and having only a modest impact on 3'end processing [82].

In contrast, mutations in the MC4R core AUUAAA hexamer did not significantly impair poly(A) cleavage while mutations of two or more nucleotides in the DSE sequence had a pronounced impact on 3'end processing efficiency (Figure 17). Interestingly, the 19 nucleotide MC4R DSE (CGTGTTATTCATAAG-CATT) is a good match of the consensus YGUGUUYY motif [220] and is very similar

to sequences that show strong CstF-64 subunit binding affinities [283]. The presence of this “optimal” DSE may explain why the MC4R poly(A) site tolerates inactivation of the hexamer sequence (Figure 17). Comparatively both JUNB upstream A-rich core sequence and the JUND noncanonical AGUAAA hexamer behave as predicted from the classical hexamer analysis experiments producing drastic effects on the efficiency of 3’end cleavage with the DSE mutations showing more modest results at either pA signal (Figures 26 and 30, respectively). Interestingly, the JUND DSE shows sequence similarity with the MC1R DSE with the presence of simple diuridines CstF-64 binding sites and also with the conspicuous presence of several 3’UTR and 3’flank G-rich elements. The presence of this sequence arrangement might be an indication of a 3’end formation mechanism dependent on proteins of the hnRNP family such as in the case of MC1R or, as reported recently, of the p53 pre-mRNA[252]. Such putative mechanism of JUND 3’end formation regulation will be investigated in the near future.

Yeast and plant 3’end processing signals are considered to be different of human poly(A) sites. *Cis*-elements are generally located upstream of the cleavage site and A-rich positioning elements commonly substitute for A(A/U)AAA hexamers (Figure 32). Mammalian poly(A) sites that contain the critical *cis*-elements upstream of the cleavage site have been described[243, 381, 390] but functional processing sites that constitute an A-rich sequence and are independent of auxiliary elements have so far been elusive. The results obtained within the scope of the work described in the present dissertation clearly show, that a mammalian poly(A) site with a potent DSE can function in the absence of canonical hexamers, but similar to yeast and plant poly(A) sites, requires an upstream A-rich sequence. Thus, the minimal mammalian poly(A) site may be

much more similar to both minimal plant and yeast poly(A) sites[229, 391] than previously thought.

The evidence obtained in the analysis of MC4R 3'end processing is also significant for the understanding of how noncanonical poly(A) sites are recognised. At least 20%-30% of human pre-mRNAs contain poly(A) sites that lack either AAUAAA or AUUAAA[298]. It is currently believed that auxiliary sequences located either upstream or downstream of non-canonical poly(A) sites may be able to compensate for a degenerated hexamer sequence by providing alternative binding opportunities for components of the 3'end processing machinery thus stabilizing poly(A) complex assembly (Figure 24). Indeed, analysis of the non-canonical poly(A) site located in the pre-mRNA of the polymerase γ gene (PAPOLG) has demonstrated that 3'UTR UGUAN sequence motifs were required for efficient 3'end processing *in vitro*. Interestingly, although the PAPOLG poly(A) site lacks a distinct hexamer, an adenosine rich sequence (AAAGAGAAA) located upstream of the cleavage site was essential for 3'end processing[203]. Interestingly, the bioinformatics analysis of more than 10,000 human poly(A) sites originated from the MC4R initial results showed that noncanonical A-rich 3'end processing sites can be found in a significant number of human genes (Figure 24).

Importantly, these results have led to the analysis of JUNB (an example identified in the bioinformatics screen) where an endogenous A-rich upstream core sequence proved to be a critical element in 3'end processing of such pre-mRNAs (Figure 24) further confirming the initial A-rich MC4R results and the subsequent bioinformatics transcriptome query. It is also worth noting that A-rich JUND poly(A) signal is functionally independent of the UGUAN mechanism described above

since only one of these sequence elements is found in JUND 3'UTR and this interaction is reported to need at least two of these elements to direct the binding of a dimerized CFIm[197]. Additionally, a second example, EDF1, identified in the screen was also tested to evaluate the validity of upstream A-rich core sequences in the definition of 3'end processing sites in a spliced-gene environment. The EDF1 results show that this gene is yet another example of an endogenous upstream A-rich core sequence (Figure 28). However, these results also indicate that, in a spliced-gene environment, efficient recognition of the poly(A), as shown by examples reported in the literature[327, 389], does require splicing, as the majority of detected transcripts are non-spliced and non-cleaved (Figure 28). Please note that it is not known why in the analysed reporter constructs splicing is not occurring efficiently. Additional constructs with only terminal EDF1 intron were also tested generating similar results (data not shown). Since the EDF1 DSE is not as clearly defined as the MC4R the recruitment of the 3'end processing factors may also require the cross-interactions from the splicing machinery components involved in 3'end processing. Perhaps, in the context of spliced genes, the recognition of A-rich poly(A) sites is, as in canonical poly(A) sites dependent on both the upstream splicing and the DSE.

In contrast to the PAPOLG poly(A) site, the MC4R DSE can direct efficient cleavage downstream of an A-rich sequence independent of the four UGUAN sequences in its 3'UTR (Figure 23). Therefore, strong DSE elements may be a general feature of many noncanonical poly(A) sites allowing efficient processing in the absence of distinct hexamers and auxiliary sequences. Consistent with this notion, our bioinformatics results confirm that A-rich noncanonical poly(A) sites can, not only be frequently found in human genes, but also tend to be enriched

with U and/or GU-rich sequences compared to canonical poly(A) sites (Figure 24). Since a higher content of U and/or GU nucleotides in the DSE is likely to stabilize CstF interactions it is plausible that 3'end processing of many noncanonical poly(A) sites may, as in the MC4R context, simply be mediated by strong DSEs.

A recently proposed model describing co-transcriptional poly(A) site recognition suggests that CPSF associated with the body of the RNA polymerase II (polII) captures a hexamer causing a transcriptional pause. This pausing then may allow CstF to establish contacts with both the DSE and CPSF. The formation of this early complex forces CPSF to disengage from the polymerase and the contact between the transcription machinery and the poly(A) complex is subsequently mediated by the now permitted interaction of CstF and the polII CTD [200, 392]. The DSE mediated recognition of poly(A) sites may still be compatible with this dynamic model. It is plausible that a strong DSE may allow CstF to rapidly associate with the pre-mRNA in the absence of pausing and establish contacts first with CPSF and then the CTD before the polymerase has moved too far down the template. The tight association between CstF and the DSE may be sufficient to maintain the tether between the processing site and polII CTD for long enough so that CPSF can establish a binding with a less favourable adenosine rich upstream sequence and allow the assembly of a functional 3'end processing complex.

The results presented in this dissertation suggest that at least in some mammalian poly(A) sites the functional relevance of processing site recognition may be shifted from the core upstream element to the downstream element. This further supports the growing understanding that the CstF–DSE interaction represents a critical regulatory step for poly(A) site choice [310, 311, 393-395]. Furthermore, the functional importance of the DSE mediated poly(A) site

recognition may be particularly critical for tissue specific 3'end formation. Tissue specific CstF isoforms have been found in mouse brain[396] and in testis[397-399]. Coincidently, pre-mRNAs of meiotic and post-meiotic male germ cells also appear to have a higher incidence of poly(A) sites with non-canonical hexamers[400] and the bioinformatics analysis presented indicates that noncanonical A-rich poly(A) sites are more likely to be subjected to tissue specific alternative 3' end processing (Figure 25).

Interestingly, the fact that a strong CstF-DSE interaction allows 3'end processing to occur at sites with more relaxed dependence on specific upstream sequences may further highlight the importance of the BRCA1/BARD1 mediated inactivation of CstF which is critical to prevent aberrant 3'end processing of nascent RNAs associated with stalled RNA polymerases after DNA damage[401-403].

Conclusion

Considering the results presented and discussed above, the present dissertation has elucidated that, at least for some mammalian intronless genes, efficient 3'end formation does not rely on additional sequence elements but rather on the strengthening of the core poly(A) sequences to compensate for the absence of the stimulatory effect of the splicing reaction on 3'end processing.

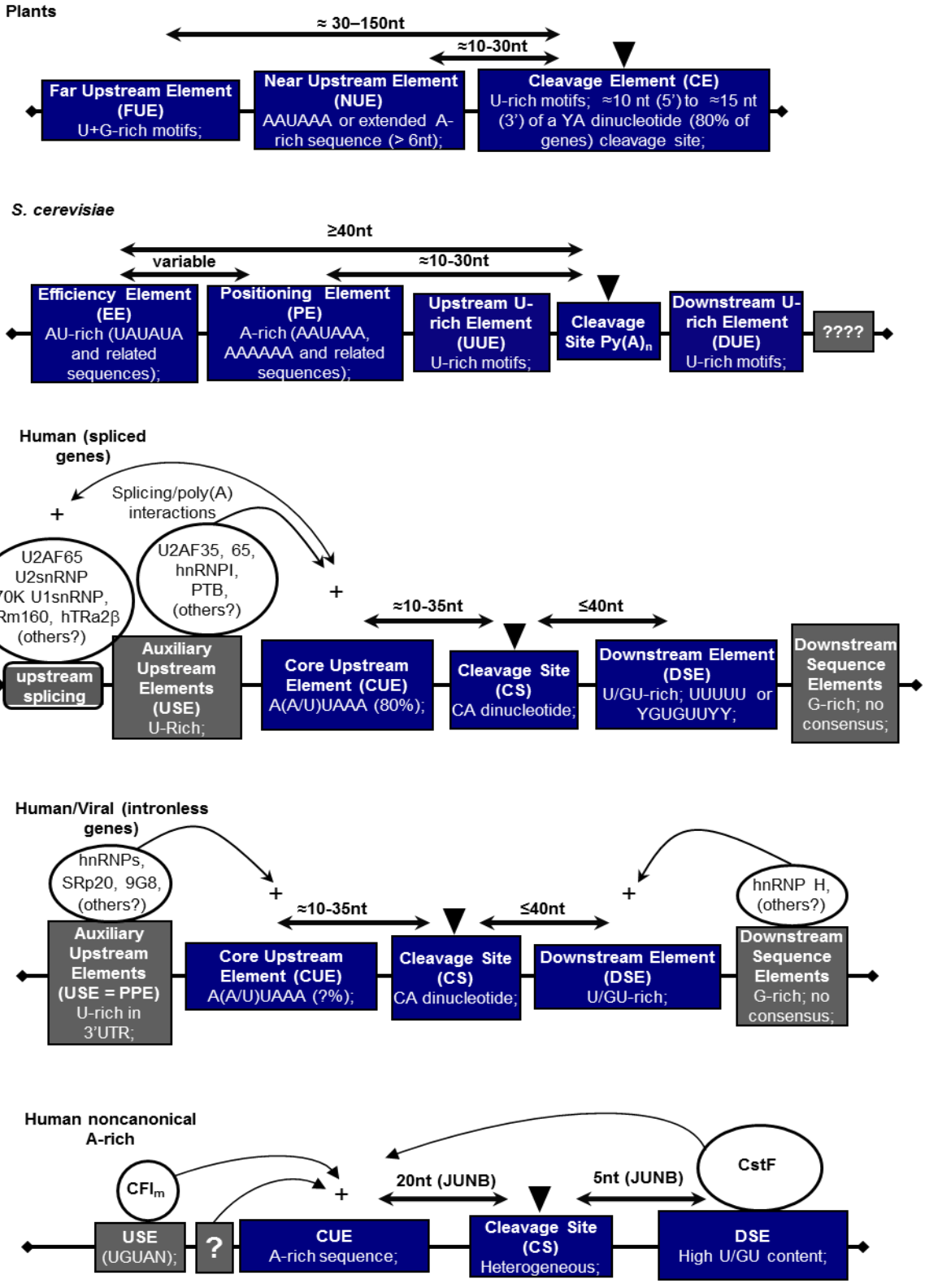


Figure 32: Comparison of yeast, plant and mammalian poly(A) sites. The sequences implicated in regulating cleavage and polyadenylation in plants, yeast (*S. cerevisiae*) and human spliced and human and viral intronless genes are represented by the grey and blue boxes. Blue boxes represent core elements (CUE, and grey boxes indicate auxiliary sequences. Black triangles indicate site of cleavage. Straight arrows and numbers above the boxes show the average distances in nucleotides between sequence motifs. Known proteins interacting with auxiliary sequences and poly(A) factors enhancing cleavage and polyadenylation are shown in the circles above the *cis*-elements. The *cis*-elements nomenclature is based on Hu et al (2005).

Chapter 5

Appendix

Appendix

Table 1: MC4R cloning primers

MC4R primer labels	Sequence (5’ - 3’)
4XBA	CGCGACTCTAGAATTAAATG
ENDS4	CCATAAGCTTTTGAAGACCCTGTAAATCCT
SPH1U	GAGCATGCTTTCTACTTCTGCA
NS2	GAAATTTCAGGCATGCTATT
FLD1	AAAGCATGCAATGCTTATGAATAACA
UTRD1	ACTCTAGATTTACAAAAAGAA
UTRD2	AAGTCTAGATTAAAAGCTTAACAATGTCT
D23F	ATTAAAAGCTTAACAATGTCTCCTTC
D23WtR	CACAGTGCCTACAACCTATAACATAG
UU2	ATGAATAGCGCGAAGGAGACATTGTT
UD1	AAGCATTGGACACTTTGCGTGCTTT
UU3	ATGAGTAGCGCGAAGGAGACATTG TT
UD2	AAGCACTGGACACTTTGCGTGCTTT
HEXU1	ACTTTTCTTTTTGTAAATCCACAGTGCCT
HEXU2	ACTTTTGTTTTTGTAAATCCACAGTGCCT
HEXD1	CCTTATTAAAAGCTTAACAATGTCTCCTT
HEXD2	CCTTATCAAAAGCTTAACAATGTCTCCTT
5WtGCF	CAGACCAGTCCTTATTAAAAGCTTAAC
5MtGCF	CACACCAGTCCTTATCAAAAGCTTAAC
Bx5GCR	TGTGTAGATCCACAGTGCCTACAACCTAT
Bx5mFmut	GACAGGAGTCCTTATCAAAAGCTTAACAA
Bx5mR	TCTGTACATCCACAGTGCCTACAACCTAT
Bx6UHF	CACACCAGTCCTTATTAAAAGCTTAAC
Bx6DHF	CAGACCAGTCCTTATCAAAAGCTTAAC
D22F	CCTTATCACAAGCTTAACAATGTCTCC
AbxU12	ACTCTTGTTTTTGTAAATCCACAGTG
AbxU22	ACTTCTGTTTTTGTAAATCCACAGTG
AbxU32	ACTTTTGTCTTTGTAAATCCACAGTG
AbxU42	ACTTTTGTTCTTGTAAATCCACAGTG
AbxU52	ACTTTTGTTTTCGTAAATCCACAGTG
AStrF	AAAAAAAAAAAAAAAAGCTTAACAATGTCTCCTT
AStrR	TTTTTTTTTTTTTTCACAGTGCCTACAACCTATA
CF1R	CTGTTGCAGAAGTAGAATATTCAGGTAGG
CF2F	CTTTCTCTTCCGTCTAGGGTACTGGTTGA
CF3R	AATAGAGAGACTGGGCATTTTTTCTC

Table 1: MC4R cloning primers (cont.)

MC4R primer labels	Sequence (5’ - 3’)
CF4F	ATTTCCAATGTCATGCTACTTTTTTG
CF5R	CCTAGAACCTATAACATAGATTCATA
CF6F	CACTGTGGATTTACAAAAAGAAAAGT
AStUGF	GTTCTAGGCACTGTGAAAAAAAAAAAA
AStUGR	CTATAACATAGATTCATATTTTATGGC

Table 2: JUNB cloning primers

JUNB primer labels	Sequence (5’ - 3’)
J3UXF	ACTCTAGATGAACGTCCCCTGCCCCTTTACGG
JFLR2	TCCCCCCGCCCTCCCAGCTTGGAAAAAAAG
4R6UF	AACAATGTCTCCTTCGCGCTACTCATAAGCACTG
JAmtR	TAAAGAACCAAGTGTGTGTGTACTTAAATAG
JDSEF	TTAATTTCTGTTGTCTTTTTTTCCAAGCTG
JAWtR	ACCAATTTTTTTTTTACTTAAATAGATTCA
JDSEmF	TCTTTATTAATCTCTGCTGCCTCTCTCTCCAAGC
SPH1U	GAGCATGCTTTCTACTTCTGCA

Table 3: JUND cloning primers

JUND primer labels	Sequence (5’ - 3’)
JDU1XbF	CGTCTAGAGTCCGCGCGCGGGGC
JD2HindR	GAAAAGCTTACACTGTAAGCCAAGCAC
JD3HindF	GTAAGCTTTTCTCTTTTGGTGGATCGG
JD3HindR	AGAAGCTTCCAGTTCTGACTTCTGACC
JDmt1TUp	AGAGCAGGACAAAAAAGGGAGGGG
JDmt1TDs	GCTCTGCTCTGCTACGAGTCCACATT
JDp1WtUp	AAAACAGAAAACCGGGCGAACC
JDp1mtUp	AGAGCAGAGAGCCGGGCGAACCAAGG
JDp1WtDs	CAGTAAAGTCTCGTTACGCCAG
JDp1mtDs	CACCCAAGTCTCGTTACGCCAGCT
JDmt2TUp	AGAGAGGGGAGGGGGGACCGGT
JDmt2TDs	GCTCTGCTCTGCTCTGCTCTGC
JDp1DmtUp1	GAGGAGGGAGGCGGAGAGCCG
JDp1DWtDs1	CCCCGCCGGGGCCTGGCGGGCT

Table 2: EDF1 cloning primers

EDF1 primer labels	Sequence (5’ - 3’)
Eex3XF	ATTCTAGAACCAAGAACACGGCCAAGC
EFLR1	CAAGGGCAGGAACTTTTCCCAGCTTCTCTT
EFLR3	CAGCATGCGGAACTTTTCCCAGCTTCTCTT
EAmutR	CGCTTTGCAAGGGTGTGTGTGTGTGACAGC
EDSEF	CTGGGTTCTGCTGGCTGTTTTTCTCCCTAC
EAWtR	CCAGCGCTTTGCAAGGTTTTGTTTGTTTGA
EDSEmF	GCTCTGCTGGCTGCTCTCCTCCCTACATCGCGCTCC

Chapter 6

Bibliography

1. Benner, S.A. Defining life. *Astrobiology* **10**, 1021-30 (2010).

2. Follmann, H. & Brownson, C. Darwin's warm little pond revisited: from molecules to the origin of life. *Naturwissenschaften* **96**, 1265-92 (2009).

3. Gayon, J. Defining life: synthesis and conclusions. *Orig Life Evol Biosph* **40**, 231-44 (2010).

4. Kutschera, U. Charles Darwin's Origin of Species, directional selection, and the evolutionary sciences today. *Naturwissenschaften* **96**, 1247-63 (2009).

5. Tsokolov, S.A. Why is the definition of life so elusive? Epistemological considerations. *Astrobiology* **9**, 401-12 (2009).

6. Kuhn, T.S. The structure of scientific revolutions (University of Chicago Press, Chicago, 1962).

7. Popper, K.R. The open society and its enemies (G. Routledge & sons, London,, 1945).

8. Canguilhem, G. Études d'histoire et de philosophie des sciences (J. Vrin, Paris,, 1968).

9. Fedorova, E. & Zink, D. Nuclear architecture and gene regulation. *Biochim Biophys Acta* **1783**, 2174-84 (2008).

10. Misteli, T. Higher-order genome organization in human disease. *Cold Spring Harb Perspect Biol* **2**, a000794 (2010).

11. Bancaud, A. et al. Molecular crowding affects diffusion and binding of nuclear proteins in heterochromatin and reveals the fractal organization of chromatin. *EMBO J* **28**, 3785-98 (2009).

12. Misteli, T. Concepts in nuclear architecture. *Bioessays* **27**, 477-87 (2005).

13. Madden, T.L. & Herzfeld, J. Crowding-induced organization of cytoskeletal elements: I. Spontaneous demixing of cytosolic proteins and model filaments to form filament bundles. *Biophys J* **65**, 1147-54 (1993).

14. Marenduzzo, D., Finan, K. & Cook, P.R. The depletion attraction: an underappreciated force driving cellular organization. *J Cell Biol* **175**, 681-6 (2006).

15. Iborra, F.J. Can visco-elastic phase separation, macromolecular crowding and colloidal physics explain nuclear organisation? *Theor Biol Med Model* **4**, 15 (2007).

16. Richter, K., Nessling, M. & Lichter, P. Macromolecular crowding and its potential impact on nuclear function. *Biochim Biophys Acta* **1783**, 2100-7 (2008).

17. Schneider, R. & Grosschedl, R. Dynamics and interplay of nuclear architecture, genome organization, and gene expression. *Genes Dev* **21**, 3027-43 (2007).

18. Woodcock, C.L. Chromatin architecture. *Curr Opin Struct Biol* **16**, 213-20 (2006).

19. Felsenfeld, G. & Groudine, M. Controlling the double helix. *Nature* **421**, 448-53 (2003).

20. Belmont, A.S. Mitotic chromosome structure and condensation. *Curr Opin Cell Biol* **18**, 632-8 (2006).

21. Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F. & Richmond, T.J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-60 (1997).

22. Richmond, T.J. & Davey, C.A. The structure of DNA in the nucleosome core. *Nature* **423**, 145-50 (2003).

23. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**, 292-301 (2001).

24. Cremer, T. et al. Chromosome territories--a functional nuclear landscape. *Curr Opin Cell Biol* **18**, 307-16 (2006).

25. Sexton, T., Schober, H., Fraser, P. & Gasser, S.M. Gene regulation through nuclear organization. *Nat Struct Mol Biol* **14**, 1049-55 (2007).

26. Branco, M.R. & Pombo, A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol* **4**, e138 (2006).

27. Gilbert, N., Gilchrist, S. & Bickmore, W.A. Chromatin organization in the mammalian nucleus. *Int Rev Cytol* **242**, 283-336 (2005).

28. Richards, E.J. & Elgin, S.C. Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. *Cell* **108**, 489-500 (2002).

29. Gilbert, N. et al. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **118**, 555-66 (2004).

30. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693-705 (2007).

31. Sadoni, N., Sullivan, K.F., Weinzierl, P., Stelzer, E.H. & Zink, D. Large-scale chromatin fibers of living cells display a discontinuous functional organization. *Chromosoma* **110**, 39-51 (2001).

32. Trojer, P. & Reinberg, D. Facultative heterochromatin: is there a distinctive molecular signature? *Mol Cell* **28**, 1-13 (2007).

33. Orphanides, G. & Reinberg, D. A unified theory of gene expression. *Cell* **108**, 439-51 (2002).

34. Paule, M.R. & White, R.J. Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Res* **28**, 1283-98 (2000).

35. Archambault, J. & Friesen, J.D. Genetics of eukaryotic RNA polymerases I, II, and III. *Microbiol Rev* **57**, 703-24 (1993).

36. Cramer, P. et al. Structure of eukaryotic RNA polymerases. *Annu Rev Biophys* **37**, 337-52 (2008).

37. Narlikar, G.J., Fan, H.Y. & Kingston, R.E. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108**, 475-87 (2002).

38. Woychik, N.A. & Hampsey, M. The RNA polymerase II machinery: structure illuminates function. *Cell* **108**, 453-63 (2002).

39. Orphanides, G., Lagrange, T. & Reinberg, D. The general transcription factors of RNA polymerase II. *Genes Dev* **10**, 2657-83 (1996).

40. Edwalds-Gilbert, G., Veraldi, K.L. & Milcarek, C. Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res* **25**, 2547-61 (1997).

41. Maniatis, T. & Reed, R. An extensive network of coupling among gene expression machines. *Nature* **416**, 499-506 (2002).

42. Hocine, S., Singer, R.H. & Grunwald, D. RNA processing and export. *Cold Spring Harb Perspect Biol* **2**, a000752 (2010).

43. Hampsey, M. Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol Mol Biol Rev* **62**, 465-503 (1998).

44. Thomas, M.C. & Chiang, C.M. The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol* **41**, 105-78 (2006).

45. Proudfoot, N.J., Furger, A. & Dye, M.J. Integrating mRNA processing with transcription. *Cell* **108**, 501-12 (2002).

46. Cook, P.R. The organization of replication and transcription. *Science* **284**, 1790-5 (1999).

47. Moore, M.J. & Proudfoot, N.J. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* **136**, 688-700 (2009).

48. Neugebauer, K.M. On the importance of being co-transcriptional. *J Cell Sci* **115**, 3865-71 (2002).

49. Martin, S. & Pombo, A. Transcription factories: quantitative studies of nanostructures in the mammalian nucleus. *Chromosome Res* **11**, 461-70 (2003).

50. Bentley, D. The mRNA assembly line: transcription and processing machines in the same factory. *Curr Opin Cell Biol* **14**, 336-42 (2002).

51. Bentley, D.L. Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr Opin Cell Biol* **17**, 251-6 (2005).

52. Jackson, D.A., Iborra, F.J., Manders, E.M. & Cook, P.R. Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei. *Mol Biol Cell* **9**, 1523-36 (1998).

53. Wetterberg, I., Zhao, J., Masich, S., Wieslander, L. & Skoglund, U. In situ transcription and splicing in the Balbiani ring 3 gene. *EMBO J* **20**, 2564-74 (2001).

54. Bird, G., Zorio, D.A. & Bentley, D.L. RNA polymerase II carboxy-terminal domain phosphorylation is required for cotranscriptional pre-mRNA splicing and 3'-end formation. *Mol Cell Biol* **24**, 8963-9 (2004).

55. Calvo, O. & Manley, J.L. Strange bedfellows: polyadenylation factors at the promoter. *Genes Dev* **17**, 1321-7 (2003).

56. Chakalova, L. & Fraser, P. Organization of transcription. *Cold Spring Harb Perspect Biol* **2**, a000729 (2010).

57. Smale, S.T. & Tjian, R. Transcription of herpes simplex virus tk sequences under the control of wild-type and mutant human RNA polymerase I promoters. *Mol Cell Biol* **5**, 352-62 (1985).

58. Lewis, E.D. & Manley, J.L. Polyadenylylation of an mRNA precursor occurs independently of transcription by RNA polymerase II in vivo. *Proc Natl Acad Sci U S A* **83**, 8555-9 (1986).

59. Sisodia, S.S., Sollner-Webb, B. & Cleveland, D.W. Specificity of RNA maturation pathways: RNAs transcribed by RNA polymerase III are not substrates for splicing or polyadenylation. *Mol Cell Biol* **7**, 3602-12 (1987).

60. Dantonel, J.C., Murthy, K.G., Manley, J.L. & Tora, L. Transcription factor TFIID recruits factor CPSF for formation of 3' end of mRNA. *Nature* **389**, 399-402 (1997).

61. Calvo, O. & Manley, J.L. Evolutionarily conserved interaction between CstF-64 and PC4 links transcription, polyadenylation, and termination. *Mol Cell* **7**, 1013-23 (2001).

62. Allison, L.A., Wong, J.K., Fitzpatrick, V.D., Moyle, M. & Ingles, C.J. The C-terminal domain of the largest subunit of RNA polymerase II of Saccharomyces cerevisiae, Drosophila melanogaster, and mammals: a conserved structure with an essential function. *Mol Cell Biol* **8**, 321-9 (1988).

63. Barron-Casella, E. & Corden, J.L. Conservation of the mammalian RNA polymerase II largest-subunit C-terminal domain. *J Mol Evol* **35**, 405-10 (1992).

64. Proudfoot, N. New perspectives on connecting messenger RNA 3' end formation to transcription. *Curr Opin Cell Biol* **16**, 272-8 (2004).

65. Palancade, B. & Bensaude, O. Investigating RNA polymerase II carboxyl-terminal domain (CTD) phosphorylation. *Eur J Biochem* **270**, 3859-70 (2003).

66. Meinhart, A., Kamenski, T., Hoeppner, S., Baumli, S. & Cramer, P. A structural perspective of CTD function. *Genes Dev* **19**, 1401-15 (2005).

67. Phatnani, H.P. & Greenleaf, A.L. Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev* **20**, 2922-36 (2006).

68. de Almeida, S.F. & Carmo-Fonseca, M. The CTD role in cotranscriptional RNA processing and surveillance. *FEBS Lett* **582**, 1971-6 (2008).

69. Mayer, A. et al. Uniform transitions of the general RNA polymerase II transcription complex. *Nat Struct Mol Biol* **17**, 1272-8 (2010).

70. Buratowski, S. Progression through the RNA polymerase II CTD cycle. *Mol Cell* **36**, 541-6 (2009).

71. Mosley, A.L. et al. Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation. *Mol Cell* **34**, 168-78 (2009).

72. Tietjen, J.R. et al. Chemical-genomic dissection of the CTD code. *Nat Struct Mol Biol* **17**, 1154-61 (2010).

73. Kim, H. et al. Gene-specific RNA polymerase II phosphorylation and the CTD code. *Nat Struct Mol Biol* **17**, 1279-86 (2010).

74. Egloff, S., Zaborowska, J., Laitem, C., Kiss, T. & Murphy, S. Ser7 Phosphorylation of the CTD Recruits the RPAP2 Ser5 Phosphatase to snRNA Genes. *Mol Cell* (2011).

75. Egloff, S. et al. Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. *Science* **318**, 1777-9 (2007).

76. Cho, H. et al. A protein phosphatase functions to recycle RNA polymerase II. *Genes Dev* **13**, 1540-52 (1999).

77. Lunde, B.M. et al. Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain. *Nat Struct Mol Biol* **17**, 1195-201 (2010).

78. Hirose, Y. & Manley, J.L. RNA polymerase II is an essential mRNA polyadenylation factor. *Nature* **395**, 93-6 (1998).

79. Hirose, Y., Tacke, R. & Manley, J.L. Phosphorylated RNA polymerase II stimulates pre-mRNA splicing. *Genes Dev* **13**, 1234-9 (1999).

80. Zorio, D.A. & Bentley, D.L. The link between mRNA processing and transcription: communication works both ways. *Exp Cell Res* **296**, 91-7 (2004).

81. McCracken, S. et al. The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* **385**, 357-61 (1997).

82. Zhao, J., Hyman, L. & Moore, C. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* **63**, 405-45 (1999).

83. Salditt-Georgieff, M., Harpold, M., Chen-Kiang, S. & Darnell, J.E., Jr. The addition of 5' cap structures occurs early in hnRNA synthesis and prematurely terminated molecules are capped. *Cell* **19**, 69-78 (1980).

84. Coppola, J.A., Field, A.S. & Luse, D.S. Promoter-proximal pausing by RNA polymerase II in vitro: transcripts shorter than 20 nucleotides are not capped. *Proc Natl Acad Sci U S A* **80**, 1251-5 (1983).

85. Shatkin, A.J. Capping of eucaryotic mRNAs. *Cell* **9**, 645-53 (1976).

86. Schoenberg, D.R. & Maquat, L.E. Re-capping the message. *Trends Biochem Sci* **34**, 435-42 (2009).

87. Jimeno-Gonzalez, S., Haaning, L.L., Malagon, F. & Jensen, T.H. The yeast 5'-3' exonuclease Rat1p functions during transcription elongation by RNA polymerase II. *Mol Cell* **37**, 580-7 (2010).

88. Sonenberg, N. & Hinnebusch, A.G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136**, 731-45 (2009).

89. Furuichi, Y. & Shatkin, A.J. Viral and cellular mRNA capping: past and prospects. *Adv Virus Res* **55**, 135-84 (2000).

90. McCracken, S. et al. 5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Genes Dev* **11**, 3306-18 (1997).

91. Shatkin, A.J. & Manley, J.L. The ends of the affair: capping and polyadenylation. *Nat Struct Biol* **7**, 838-42 (2000).

92. Shuman, S. & Schwer, B. RNA capping enzyme and DNA ligase: a superfamily of covalent nucleotidyl transferases. *Mol Microbiol* **17**, 405-10 (1995).

93. Lehninger, A.L., Nelson, D.L. & Cox, M.M. Lehninger principles of biochemistry (W.H. Freeman, New York, 2008).

94. Howe, K.J. RNA polymerase II conducts a symphony of pre-mRNA processing activities. *Biochim Biophys Acta* **1577**, 308-24 (2002).

95. Izaurralde, E. et al. A nuclear cap binding protein complex involved in pre-mRNA splicing. *Cell* **78**, 657-68 (1994).

96. Visa, N., Izaurralde, E., Ferreira, J., Daneholt, B. & Mattaj, I.W. A nuclear cap-binding complex binds Balbiani ring pre-mRNA cotranscriptionally and accompanies the ribonucleoprotein particle during nuclear export. *J Cell Biol* **133**, 5-14 (1996).

97. Muhlrad, D., Decker, C.J. & Parker, R. Deadenylation of the unstable mRNA encoded by the yeast MFA2 gene leads to decapping followed by 5'-->3' digestion of the transcript. *Genes Dev* **8**, 855-66 (1994).

98. Shuman, S. Structure, mechanism, and evolution of the mRNA capping apparatus. *Prog Nucleic Acid Res Mol Biol* **66**, 1-40 (2001).

99. Lewis, J.D. & Izaurralde, E. The role of the cap structure in RNA processing and nuclear export. *Eur J Biochem* **247**, 461-9 (1997).

100. Raczynska, K.D. et al. Involvement of the nuclear cap-binding protein complex in alternative splicing in Arabidopsis thaliana. *Nucleic Acids Res* **38**, 265-78 (2010).

101. Kim, S. et al. Two cap-binding proteins CBP20 and CBP80 are involved in processing primary MicroRNAs. *Plant Cell Physiol* **49**, 1634-44 (2008).

102. Lenasi, T., Peterlin, B.M. & Barboric, M. Cap-binding protein complex links pre-mRNA capping to transcription elongation and alternative splicing through positive transcription elongation factor b (P-TEFb). *J Biol Chem* **286**, 22758-68 (2011).

103. Sachs, A.B., Sarnow, P. & Hentze, M.W. Starting at the beginning, middle, and end: translation initiation in eukaryotes. *Cell* **89**, 831-8 (1997).

104. Gunnery, S. & Mathews, M.B. Functional mRNA can be generated by RNA polymerase III. *Mol Cell Biol* **15**, 3597-607 (1995).

105. Lo, H.J., Huang, H.K. & Donahue, T.F. RNA polymerase I-promoted HIS4 expression yields uncapped, polyadenylated mRNA that is unstable and inefficiently translated in Saccharomyces cerevisiae. *Mol Cell Biol* **18**, 665-75 (1998).

106. Yue, Z. et al. Mammalian capping enzyme complements mutant Saccharomyces cerevisiae lacking mRNA guanylyltransferase and selectively binds the elongating form of RNA polymerase II. *Proc Natl Acad Sci U S A* **94**, 12898-903 (1997).

107. Cho, E.J., Rodriguez, C.R., Takagi, T. & Buratowski, S. Allosteric interactions between capping enzyme subunits and the RNA polymerase II carboxy-terminal domain. *Genes Dev* **12**, 3482-7 (1998).

108. Chiba, K., Yamamoto, J., Yamaguchi, Y. & Handa, H. Promoter-proximal pausing and its release: molecular mechanisms and physiological functions. *Exp Cell Res* **316**, 2723-30 (2010).

109. Ho, C.K. & Shuman, S. Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme. *Mol Cell* **3**, 405-11 (1999).

110. Komarnitsky, P., Cho, E.J. & Buratowski, S. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev* **14**, 2452-60 (2000).

111. Schroeder, S.C., Schwer, B., Shuman, S. & Bentley, D. Dynamic association of capping enzymes with transcribing RNA polymerase II. *Genes Dev* **14**, 2435-40 (2000).

112. Mandal, S.S. et al. Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II. *Proc Natl Acad Sci U S A* **101**, 7572-7 (2004).

113. Myers, L.C., Lacomis, L., Erdjument-Bromage, H. & Tempst, P. The yeast capping enzyme represses RNA polymerase II transcription. *Mol Cell* **10**, 883-94 (2002).

114. Schroeder, S.C., Zorio, D.A., Schwer, B., Shuman, S. & Bentley, D. A function of yeast mRNA cap methyltransferase, Abd1, in transcription by RNA polymerase II. *Mol Cell* **13**, 377-87 (2004).

115. Kim, M., Ahn, S.H., Krogan, N.J., Greenblatt, J.F. & Buratowski, S. Transitions in RNA polymerase II elongation complexes at the 3' ends of genes. *EMBO J* **23**, 354-64 (2004).

116. Viladevall, L. et al. TFIIH and P-TEFb coordinate transcription with capping enzyme recruitment at specific genes in fission yeast. *Mol Cell* **33**, 738-51 (2009).

117. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

118. Kramer, A. The structure and function of proteins involved in mammalian pre-mRNA splicing. *Annu Rev Biochem* **65**, 367-409 (1996).

119. Wahl, M.C., Will, C.L. & Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**, 701-18 (2009).

120. Chen, Y.I. et al. Proteomic analysis of in vivo-assembled pre-mRNA splicing complexes expands the catalog of participating factors. *Nucleic Acids Res* **35**, 3928-44 (2007).

121. Jurica, M.S. & Moore, M.J. Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell* **12**, 5-14 (2003).

122. Blencowe, B.J. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci* **25**, 106-10 (2000).

123. Liu, H.X., Chew, S.L., Cartegni, L., Zhang, M.Q. & Krainer, A.R. Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol Cell Biol* **20**, 1063-71 (2000).

124. Busch, A. & Hertel, K.J. Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdiscip Rev RNA* **3**, 1-12 (2012).

125. Tarn, W.Y. & Steitz, J.A. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell* **84**, 801-11 (1996).

126. Soller, M. Pre-messenger RNA processing and its regulation: a genomic perspective. *Cell Mol Life Sci* **63**, 796-819 (2006).

127. Moore, M.J. & Sharp, P.A. Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing. *Nature* **365**, 364-8 (1993).

128. Zhou, Z., Licklider, L.J., Gygi, S.P. & Reed, R. Comprehensive proteomic analysis of the human spliceosome. *Nature* **419**, 182-5 (2002).

129. Rappsilber, J., Ryder, U., Lamond, A.I. & Mann, M. Large-scale proteomic analysis of the human spliceosome. *Genome Res* **12**, 1231-45 (2002).

130. Sun, J.S. & Manley, J.L. A novel U2-U6 snRNA structure is necessary for mammalian mRNA splicing. *Genes Dev* **9**, 843-54 (1995).

131. Madhani, H.D. & Guthrie, C. A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome. *Cell* **71**, 803-17 (1992).

132. Perales, R. & Bentley, D. "Cotranscriptionality": the transcription elongation complex as a nexus for nuclear transactions. *Mol Cell* **36**, 178-91 (2009).

133. Berget, S.M. Exon recognition in vertebrate splicing. *J Biol Chem* **270**, 2411-4 (1995).

134. Azubel, M., Wolf, S.G., Sperling, J. & Sperling, R. Three-dimensional structure of the native spliceosome by cryo-electron microscopy. *Mol Cell* **15**, 833-9 (2004).

135. Chusainow, J. et al. FRET analyses of the U2AF complex localize the U2AF35/U2AF65 interaction in vivo and reveal a novel self-interaction of U2AF35. *RNA* **11**, 1201-14 (2005).

136. Rino, J., Desterro, J.M., Pacheco, T.R., Gadella, T.W., Jr. & Carmo-Fonseca, M. Splicing factors SF1 and U2AF associate in extraspliceosomal complexes. *Mol Cell Biol* **28**, 3045-57 (2008).

137. Ellis, J.D., Lleres, D., Denegri, M., Lamond, A.I. & Caceres, J.F. Spatial mapping of splicing factor complexes involved in exon and intron definition. *J Cell Biol* **181**, 921-34 (2008).

138. Rino, J. & Carmo-Fonseca, M. The spliceosome: a self-organized macromolecular machine in the nucleus? *Trends Cell Biol* **19**, 375-84 (2009).

139. Nilsen, T.W. & Graveley, B.R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457-63 (2010).

140. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413-5 (2008).

141. Wang, E.T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-6 (2008).

142. Richard, P. & Manley, J.L. Transcription termination by nuclear RNA polymerases. *Genes Dev* **23**, 1247-69 (2009).

143. Buratowski, S. Connections between mRNA 3' end processing and transcription termination. *Curr Opin Cell Biol* **17**, 257-61 (2005).

144. Han, J., Xiong, J., Wang, D. & Fu, X.D. Pre-mRNA splicing: where and when in the nucleus. *Trends Cell Biol* **21**, 336-43 (2011).

145. Bauren, G. & Wieslander, L. Splicing of Balbiani ring 1 gene pre-mRNA occurs simultaneously with transcription. *Cell* **76**, 183-92 (1994).

146. Pandya-Jones, A. & Black, D.L. Co-transcriptional splicing of constitutive and alternative exons. *RNA* **15**, 1896-908 (2009).

147. Brody, Y. & Shav-Tal, Y. Transcription and splicing: When the twain meet. *Transcription* **2**, 216-20 (2011).

148. Denis, M.M. et al. Escaping the nuclear confines: signal-dependent pre-mRNA splicing in anucleate platelets. *Cell* **122**, 379-91 (2005).

149. Glanzer, J. et al. RNA splicing capability of live neuronal dendrites. *Proc Natl Acad Sci U S A* **102**, 16859-64 (2005).

150. Bell, T.J. et al. Intron retention facilitates splice variant diversity in calcium-activated big potassium channel populations. *Proc Natl Acad Sci U S A* **107**, 21152-7 (2010).

151. Brody, Y. et al. The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. *PLoS Biol* **9**, e1000573 (2011).

152. Beyer, A.L. & Osheim, Y.N. Visualization of RNA transcription and processing. *Semin Cell Biol* **2**, 131-40 (1991).

153. Beyer, A.L. & Osheim, Y.N. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev* **2**, 754-65 (1988).

154. Osheim, Y.N. & Beyer, A.L. EM analysis of Drosophila chorion genes: amplification, transcription termination and RNA splicing. *Electron Microsc Rev* **4**, 111-28 (1991).

155. Zhang, G., Taneja, K.L., Singer, R.H. & Green, M.R. Localization of pre-mRNA splicing in mammalian nuclei. *Nature* **372**, 809-12 (1994).

156. Singh, J. & Padgett, R.A. Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol* **16**, 1128-33 (2009).

157. Gornemann, J. et al. Cotranscriptional spliceosome assembly and splicing are independent of the Prp40p WW domain. *RNA* **17**, 2119-29 (2011).

158. Gornemann, J., Kotovic, K.M., Hujer, K. & Neugebauer, K.M. Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. *Mol Cell* **19**, 53-63 (2005).

159. Lacadie, S.A. & Rosbash, M. Cotranscriptional spliceosome assembly dynamics and the role of U1 snRNA:5'ss base pairing in yeast. *Mol Cell* **19**, 65-75 (2005).

160. Tardiff, D.F., Lacadie, S.A. & Rosbash, M. A genome-wide analysis indicates that yeast pre-mRNA splicing is predominantly posttranscriptional. *Mol Cell* **24**, 917-29 (2006).

161. Listerman, I., Sapra, A.K. & Neugebauer, K.M. Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. *Nat Struct Mol Biol* **13**, 815-22 (2006).

162. Carrillo Oesterreich, F., Preibisch, S. & Neugebauer, K.M. Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Mol Cell* **40**, 571-81 (2010).

163. Oesterreich, F.C., Bieberstein, N. & Neugebauer, K.M. Pause locally, splice globally. *Trends Cell Biol* **21**, 328-35 (2011).

164. Alexander, R.D., Innocente, S.A., Barrass, J.D. & Beggs, J.D. Splicing-dependent RNA polymerase pausing in yeast. *Mol Cell* **40**, 582-93 (2010).

165. de la Mata, M. et al. A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* **12**, 525-32 (2003).

166. Cramer, P. et al. Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer. *Mol Cell* **4**, 251-8 (1999).

167. Batsche, E., Yaniv, M. & Muchardt, C. The human SWI/SNF subunit Brm is a regulator of alternative splicing. *Nat Struct Mol Biol* **13**, 22-9 (2006).

168. Luco, R.F. et al. Regulation of alternative splicing by histone modifications. *Science* **327**, 996-1000 (2010).

169. Munoz, M.J. et al. DNA damage regulates alternative splicing through inhibition of RNA polymerase II elongation. *Cell* **137**, 708-20 (2009).

170. Damgaard, C.K. et al. A 5' splice site enhances the recruitment of basal transcription initiation factors in vivo. *Mol Cell* **29**, 271-8 (2008).

171. Lin, S., Coutinho-Mansfield, G., Wang, D., Pandit, S. & Fu, X.D. The splicing factor SC35 has an active role in transcriptional elongation. *Nat Struct Mol Biol* **15**, 819-26 (2008).

172. Fong, Y.W. & Zhou, Q. Stimulatory effect of splicing factors on transcriptional elongation. *Nature* **414**, 929-33 (2001).

173. Zeng, C. & Berget, S.M. Participation of the C-terminal domain of RNA polymerase II in exon definition during pre-mRNA splicing. *Mol Cell Biol* **20**, 8290-301 (2000).

174. Misteli, T. & Spector, D.L. RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo. *Mol Cell* **3**, 697-705 (1999).

175. Fong, N. & Bentley, D.L. Capping, splicing, and 3' processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD. *Genes Dev* **15**, 1783-95 (2001).

176. Vincent, M. et al. The nuclear matrix protein p255 is a highly phosphorylated form of RNA polymerase II largest subunit which associates with spliceosomes. *Nucleic Acids Res* **24**, 4649-52 (1996).

177. Mortillaro, M.J. et al. A hyperphosphorylated form of the large subunit of RNA polymerase II is associated with splicing complexes and the nuclear matrix. *Proc Natl Acad Sci U S A* **93**, 8253-7 (1996).

178. Yuryev, A. et al. The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins. *Proc Natl Acad Sci U S A* **93**, 6975-80 (1996).

179. Kim, E., Du, L., Bregman, D.B. & Warren, S.L. Splicing factors associate with hyperphosphorylated RNA polymerase II in the absence of pre-mRNA. *J Cell Biol* **136**, 19-28 (1997).

180. Das, R. et al. SR proteins function in coupling RNAP II transcription to pre-mRNA splicing. *Mol Cell* **26**, 867-81 (2007).

181. Sapra, A.K. et al. SR protein family members display diverse activities in the formation of nascent and mature mRNPs in vivo. *Mol Cell* **34**, 179-90 (2009).

182. David, C.J., Boyne, A.R., Millhouse, S.R. & Manley, J.L. The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex. *Genes Dev* **25**, 972-83 (2011).

183. Natalizio, B.J., Robson-Dixon, N.D. & Garcia-Blanco, M.A. The Carboxyl-terminal Domain of RNA Polymerase II Is Not Sufficient to Enhance the Efficiency of Pre-mRNA Capping or Splicing in the Context of a Different Polymerase. *J Biol Chem* **284**, 8692-702 (2009).

184. Du, L. & Warren, S.L. A functional interaction between the carboxy-terminal domain of RNA polymerase II and pre-mRNA splicing. *J Cell Biol* **136**, 5-18 (1997).

185. de la Mata, M. & Kornblihtt, A.R. RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. *Nat Struct Mol Biol* **13**, 973-80 (2006).

186. Munoz, M.J., de la Mata, M. & Kornblihtt, A.R. The carboxy terminal domain of RNA polymerase II and alternative splicing. *Trends Biochem Sci* **35**, 497-504 (2010).

187. Brinster, R.L., Allen, J.M., Behringer, R.R., Gelinas, R.E. & Palmiter, R.D. Introns increase transcriptional efficiency in transgenic mice. *Proc Natl Acad Sci U S A* **85**, 836-40 (1988).

188. Furger, A., O'Sullivan, J.M., Binnie, A., Lee, B.A. & Proudfoot, N.J. Promoter proximal splice sites enhance transcription. *Genes Dev* **16**, 2792-9 (2002).

189. Cramer, P., Pesce, C.G., Baralle, F.E. & Kornblihtt, A.R. Functional association between promoter structure and transcript alternative splicing. *Proc Natl Acad Sci U S A* **94**, 11456-60 (1997).

190. Ge, H., Si, Y. & Wolffe, A.P. A novel transcriptional coactivator, p52, functionally interacts with the essential splicing factor ASF/SF2. *Mol Cell* **2**, 751-9 (1998).

191. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol* **16**, 990-5 (2009).

192. Tilgner, H. et al. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16**, 996-1001 (2009).

193. Dye, M.J., Gromak, N. & Proudfoot, N.J. Exon tethering in transcription by RNA polymerase II. *Mol Cell* **21**, 849-59 (2006).

194. Fong, N., Ohman, M. & Bentley, D.L. Fast ribozyme cleavage releases transcripts from RNA polymerase II and aborts co-transcriptional pre-mRNA processing. *Nat Struct Mol Biol* **16**, 916-22 (2009).

195. Huang, Y. & Carmichael, G.G. Role of polyadenylation in nucleocytoplasmic transport of mRNA. *Mol Cell Biol* **16**, 1534-42 (1996).

196. Mapendano, C.K., Lykke-Andersen, S., Kjems, J., Bertrand, E. & Jensen, T.H. Crosstalk between mRNA 3' end processing and transcription initiation. *Mol Cell* **40**, 410-22 (2010).

197. Mandel, C.R., Bai, Y. & Tong, L. Protein factors in pre-mRNA 3'-end processing. *Cell Mol Life Sci* **65**, 1099-122 (2008).

198. Colgan, D.F. & Manley, J.L. Mechanism and regulation of mRNA polyadenylation. *Genes Dev* **11**, 2755-66 (1997).

199. Millevoi, S. & Vagner, S. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic Acids Res* **38**, 2757-74 (2010).

200. Nag, A., Narsinh, K. & Martinson, H.G. The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. *Nat Struct Mol Biol* **14**, 662-9 (2007).

201. Glover-Cutter, K., Kim, S., Espinosa, J. & Bentley, D.L. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nat Struct Mol Biol* **15**, 71-8 (2008).

202. Ryan, K., Murthy, K.G., Kaneko, S. & Manley, J.L. Requirements of the RNA polymerase II C-terminal domain for reconstituting pre-mRNA 3' cleavage. *Mol Cell Biol* **22**, 1684-92 (2002).

203. Venkataraman, K., Brown, K.M. & Gilmartin, G.M. Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. *Genes Dev* **19**, 1315-27 (2005).

204. Licatalosi, D.D. et al. Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II. *Mol Cell* **9**, 1101-11 (2002).

205. Birse, C.E., Minvielle-Sebastia, L., Lee, B.A., Keller, W. & Proudfoot, N.J. Coupling termination of transcription to messenger RNA maturation in yeast. *Science* **280**, 298-301 (1998).

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

207. Wang, Y., Fairley, J.A. & Roberts, S.G. Phosphorylation of TFIIB links transcription initiation and termination. *Curr Biol* **20**, 548-53 (2010).

208. Nagaike, T. et al. Transcriptional activators enhance polyadenylation of mRNA precursors. *Mol Cell* **41**, 409-18 (2011).

209. Krishnamurthy, S., He, X., Reyes-Reyes, M., Moore, C. & Hampsey, M. Ssu72 Is an RNA polymerase II CTD phosphatase. *Mol Cell* **14**, 387-94 (2004).

210. O'Sullivan, J.M. et al. Gene loops juxtapose promoters and terminators in yeast. *Nat Genet* **36**, 1014-8 (2004).

211. Hilleren, P.J. & Parker, R. Cytoplasmic degradation of splice-defective pre-mRNAs and intermediates. *Mol Cell* **12**, 1453-65 (2003).

212. Rigo, F. & Martinson, H.G. Polyadenylation releases mRNA from RNA polymerase II in a process that is licensed by splicing. *RNA* **15**, 823-36 (2009).

213. Nag, A., Narsinh, K., Kazerouninia, A. & Martinson, H.G. The conserved AAUAAA hexamer of the poly(A) signal can act alone to trigger a stable decrease in RNA polymerase II transcription velocity. *RNA* **12**, 1534-44 (2006).

214. Whitelaw, E. & Proudfoot, N. Alpha-thalassaemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to 3' end processing in the human alpha 2 globin gene. *EMBO J* **5**, 2915-22 (1986).

215. Russo, P. & Sherman, F. Transcription terminates near the poly(A) site in the CYC1 gene of the yeast Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **86**, 8348-52 (1989).

216. Birse, C.E., Lee, B.A., Hansen, K. & Proudfoot, N.J. Transcriptional termination signals for RNA polymerase II in fission yeast. *EMBO J* **16**, 3633-43 (1997).

217. Osheim, Y.N., Proudfoot, N.J. & Beyer, A.L. EM visualization of transcription by RNA polymerase II: downstream termination requires a poly(A) signal but not transcript cleavage. *Mol Cell* **3**, 379-87 (1999).

218. Gromak, N., West, S. & Proudfoot, N.J. Pause sites promote transcriptional termination of mammalian RNA polymerase II. *Mol Cell Biol* **26**, 3986-96 (2006).

219. Kim, M. et al. The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature* **432**, 517-22 (2004).

220. Zarudnaya, M.I., Kolomiets, I.M., Potyahaylo, A.L. & Hovorun, D.M. Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. *Nucleic Acids Res* **31**, 1375-86 (2003).

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

222. Manley, J.L. Polyadenylation of mRNA precursors. *Biochim Biophys Acta* **950**, 1-12 (1988).

223. Proudfoot, N.J. & Brownlee, G.G. 3' non-coding region sequences in eukaryotic messenger RNA. *Nature* **263**, 211-4 (1976).

224. Proudfoot, N. Poly(A) signals. *Cell* **64**, 671-4 (1991).

225. Tian, B. & Graber, J.H. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip Rev RNA* (2011).

226. Wickens, M. & Stephenson, P. Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3' end formation. *Science* **226**, 1045-51 (1984).

227. Sheets, M.D., Ogg, S.C. & Wickens, M.P. Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res* **18**, 5799-805 (1990).

228. Wilusz, J., Pettine, S.M. & Shenk, T. Functional analysis of point mutations in the AAUAAA motif of the SV40 late polyadenylation signal. *Nucleic Acids Res* **17**, 3899-908 (1989).

229. Hunt, A.G. Messenger RNA 3' end formation in plants. *Curr Top Microbiol Immunol* **326**, 151-77 (2008).

230. Chou, Z.F., Chen, F. & Wilusz, J. Sequence and position requirements for uridylate-rich downstream elements of polyadenylation signals. *Nucleic Acids Res* **22**, 2525-31 (1994).

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

232. McLauchlan, J., Gaffney, D., Whitton, J.L. & Clements, J.B. The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini. *Nucleic Acids Res* **13**, 1347-68 (1985).

233. McDevitt, M.A., Hart, R.P., Wong, W.W. & Nevins, J.R. Sequences capable of restoring poly(A) site function define two distinct downstream elements. *EMBO J* **5**, 2907-13 (1986).

234. Dalziel, M., Nunes, N.M. & Furger, A. Two G-rich regulatory elements located adjacent to and 440 nucleotides downstream of the core poly(A) site of the intronless melanocortin receptor 1 gene are critical for efficient 3' end processing. *Mol Cell Biol* **27**, 1568-80 (2007).

235. Chen, F., MacDonald, C.C. & Wilusz, J. Cleavage site determinants in the mammalian polyadenylation signal. *Nucleic Acids Res* **23**, 2614-20 (1995).

236. Shi, Y. et al. Molecular architecture of the human pre-mRNA 3' processing complex. *Mol Cell* **33**, 365-76 (2009).

237. Qiu, J., Nayak, R. & Pintel, D.J. 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. *J Virol* **78**, 83-93 (2004).

238. Valsamakis, A., Schek, N. & Alwine, J.C. Elements upstream of the AAUAAA within the human immunodeficiency virus polyadenylation signal are required for efficient polyadenylation in vitro. *Mol Cell Biol* **12**, 3699-705 (1992).

239. Valsamakis, A., Zeichner, S., Carswell, S. & Alwine, J.C. The human immunodeficiency virus type 1 polyadenylylation signal: a 3' long terminal repeat element upstream of the AAUAAA necessary for efficient polyadenylylation. *Proc Natl Acad Sci U S A* **88**, 2108-12 (1991).

240. Zhao, X. et al. A 57-nucleotide upstream early polyadenylation element in human papillomavirus type 16 interacts with hFip1, CstF-64, hnRNP C1/C2, and polypyrimidine tract binding protein. *J Virol* **79**, 4270-88 (2005).

241. Moreira, A. et al. The upstream sequence element of the C2 complement poly(A) signal activates mRNA 3' end formation by two distinct mechanisms. *Genes Dev* **12**, 2522-34 (1998).

242. Brackenridge, S., Ashe, H.L., Giacca, M. & Proudfoot, N.J. Transcription and polyadenylation in a short human intergenic region. *Nucleic Acids Res* **25**, 2326-36 (1997).

243. Brackenridge, S. & Proudfoot, N.J. Recruitment of a basal polyadenylation factor by the upstream sequence element of the human lamin B2 polyadenylation signal. *Mol Cell Biol* **20**, 2660-9 (2000).

244. Hall-Pogar, T., Zhang, H., Tian, B. & Lutz, C.S. Alternative polyadenylation of cyclooxygenase-2. *Nucleic Acids Res* **33**, 2565-79 (2005).

245. Huang, Y. & Carmichael, G.G. The mouse histone H2a gene contains a small element that facilitates cytoplasmic accumulation of intronless gene transcripts and of unspliced HIV-1-related mRNAs. *Proc Natl Acad Sci U S A* **94**, 10104-9 (1997).

246. Bagga, P.S., Ford, L.P., Chen, F. & Wilusz, J. The G-rich auxiliary downstream element has distinct sequence and position requirements and mediates efficient 3' end pre-mRNA processing through a trans-acting factor. *Nucleic Acids Res* **23**, 1625-31 (1995).

247. Chen, F. & Wilusz, J. Auxiliary downstream elements are required for efficient polyadenylation of mammalian pre-mRNAs. *Nucleic Acids Res* **26**, 2891-8 (1998).

248. Arhin, G.K., Boots, M., Bagga, P.S., Milcarek, C. & Wilusz, J. Downstream sequence elements with different affinities for the hnRNP H/H' protein influence the processing efficiency of mammalian polyadenylation signals. *Nucleic Acids Res* **30**, 1842-50 (2002).

249. Caputi, M. & Zahler, A.M. Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family. *J Biol Chem* **276**, 43850-9 (2001).

250. Hu, J., Lutz, C.S., Wilusz, J. & Tian, B. Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA* **11**, 1485-93 (2005).

251. Oberg, D., Fay, J., Lambkin, H. & Schwartz, S. A downstream polyadenylation element in human papillomavirus type 16 L2 encodes multiple GGG motifs and interacts with hnRNP H. *J Virol* **79**, 9254-69 (2005).

252. Decorsiere, A., Cayrel, A., Vagner, S. & Millevoi, S. 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 Dev* **25**, 220-5 (2011).

253. Yonaha, M. & Proudfoot, N.J. Transcriptional termination and coupled polyadenylation in vitro. *EMBO J* **19**, 3770-7 (2000).

254. Yonaha, M. & Proudfoot, N.J. Specific transcriptional pausing activates polyadenylation in a coupled in vitro system. *Mol Cell* **3**, 593-600 (1999).

255. Lou, H., Helfman, D.M., Gagel, R.F. & Berget, S.M. Polypyrimidine tract-binding protein positively regulates inclusion of an alternative 3'-terminal exon. *Mol Cell Biol* **19**, 78-85 (1999).

256. Furth, P.A., Choe, W.T., Rex, J.H., Byrne, J.C. & Baker, C.C. Sequences homologous to 5' splice sites are required for the inhibitory activity of papillomavirus late 3' untranslated regions. *Mol Cell Biol* **14**, 5278-89 (1994).

257. Ashe, M.P., Griffin, P., James, W. & Proudfoot, N.J. Poly(A) site selection in the HIV-1 provirus: inhibition of promoter-proximal polyadenylation by the downstream major splice donor site. *Genes Dev* **9**, 3008-25 (1995).

258. Kaneko, S. & Manley, J.L. The mammalian RNA polymerase II C-terminal domain interacts with RNA to suppress transcription-coupled 3' end formation. *Mol Cell* **20**, 91-103 (2005).

259. Ahmed, Y.F., Gilmartin, G.M., Hanly, S.M., Nevins, J.R. & Greene, W.C. The HTLV-I Rex response element mediates a novel form of mRNA polyadenylation. *Cell* **64**, 727-37 (1991).

260. Gilmartin, G.M., Fleming, E.S. & Oetjen, J. Activation of HIV-1 pre-mRNA 3' processing in vitro requires both an upstream element and TAR. *EMBO J* **11**, 4419-28 (1992).

261. Klasens, B.I., Das, A.T. & Berkhout, B. Inhibition of polyadenylation by stable RNA secondary structure. *Nucleic Acids Res* **26**, 1870-6 (1998).

262. Das, A.T., Klaver, B. & Berkhout, B. A hairpin structure in the R region of the human immunodeficiency virus type 1 RNA genome is instrumental in polyadenylation site selection. *J Virol* **73**, 81-91 (1999).

263. Klasens, B.I., Thiesen, M., Virtanen, A. & Berkhout, B. The ability of the HIV-1 AAUAAA signal to bind polyadenylation factors is controlled by local RNA structure. *Nucleic Acids Res* **27**, 446-54 (1999).

264. Hans, H. & Alwine, J.C. Functionally significant secondary structure of the simian virus 40 late polyadenylation signal. *Mol Cell Biol* **20**, 2926-32 (2000).

265. Wu, C. & Alwine, J.C. Secondary structure as a functional feature in the downstream region of mammalian polyadenylation signals. *Mol Cell Biol* **24**, 2789-96 (2004).

266. Phillips, C., Kyriakopoulou, C.B. & Virtanen, A. Identification of a stem-loop structure important for polyadenylation at the murine IgM secretory poly(A) site. *Nucleic Acids Res* **27**, 429-38 (1999).

267. Graveley, B.R., Fleming, E.S. & Gilmartin, G.M. RNA structure is a critical determinant of poly(A) site recognition by cleavage and polyadenylation specificity factor. *Mol Cell Biol* **16**, 4942-51 (1996).

268. Graveley, B.R., Fleming, E.S. & Gilmartin, G.M. Restoration of both structure and function to a defective poly(A) site by in vitro selection. *J Biol Chem* **271**, 33654-63 (1996).

269. Shi, Y. et al. Molecular Architecture of the Human Pre-mRNA 3' Processing Complex. *Molecular Cell* **33**, 365 (2009).

270. Kaufmann, I., Martin, G., Friedlein, A., Langen, H. & Keller, W. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *EMBO J* **23**, 616-26 (2004).

271. Murthy, K.G. & Manley, J.L. The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation. *Genes Dev* **9**, 2672-83 (1995).

272. Callebaut, I., Moshous, D., Mornon, J.P. & de Villartay, J.P. Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family. *Nucleic Acids Res* **30**, 3592-601 (2002).

273. Ryan, K., Calvo, O. & Manley, J.L. Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. *RNA* **10**, 565-73 (2004).

274. Mandel, C.R. et al. Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature* **444**, 953-6 (2006).

275. Dominski, Z., Yang, X.C. & Marzluff, W.F. The polyadenylation factor CPSF-73 is involved in histone-pre-mRNA processing. *Cell* **123**, 37-48 (2005).

276. Gilmartin, G.M. & Nevins, J.R. Molecular analyses of two poly(A) site-processing factors that determine the recognition and efficiency of cleavage of the pre-mRNA. *Mol Cell Biol* **11**, 2432-8 (1991).

277. Takagaki, Y., Manley, J.L., MacDonald, C.C., Wilusz, J. & Shenk, T. A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs. *Genes Dev* **4**, 2112-20 (1990).

278. MacDonald, C.C., Wilusz, J. & Shenk, T. The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. *Mol Cell Biol* **14**, 6647-54 (1994).

279. Takagaki, Y., Ryner, L.C. & Manley, J.L. Four factors are required for 3'-end cleavage of pre-mRNAs. *Genes Dev* **3**, 1711-24 (1989).

280. Gilmartin, G.M. & Nevins, J.R. An ordered pathway of assembly of components required for polyadenylation site recognition and processing. *Genes Dev* **3**, 2180-90 (1989).

281. Takagaki, Y. & Manley, J.L. A polyadenylation factor subunit is the human homologue of the Drosophila suppressor of forked protein. *Nature* **372**, 471-4 (1994).

282. Takagaki, Y. & Manley, J.L. RNA recognition by the human polyadenylation factor CstF. *Mol Cell Biol* **17**, 3907-14 (1997).

283. Perez Canadillas, J.M. & Varani, G. Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. *EMBO J* **22**, 2821-30 (2003).

284. Ryan, K. Pre-mRNA 3' cleavage is reversibly inhibited in vitro by cleavage factor dephosphorylation. *RNA Biol* **4**, 26-33 (2007).

285. Ruegsegger, U., Blank, D. & Keller, W. Human pre-mRNA cleavage factor Im is related to spliceosomal SR proteins and can be reconstituted in vitro from recombinant subunits. *Mol Cell* **1**, 243-53 (1998).

286. Ruegsegger, U., Beyer, K. & Keller, W. Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors. *J Biol Chem* **271**, 6107-13 (1996).

287. Brown, K.M. & Gilmartin, G.M. A mechanism for the regulation of pre-mRNA 3' processing by human cleavage factor Im. *Mol Cell* **12**, 1467-76 (2003).

288. de Vries, H. et al. Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. *EMBO J* **19**, 5895-904 (2000).

289. Wahle, E., Martin, G., Schiltz, E. & Keller, W. Isolation and expression of cDNA clones encoding mammalian poly(A) polymerase. *EMBO J* **10**, 4251-7 (1991).

290. Wahle, E. Purification and characterization of a mammalian polyadenylate polymerase involved in the 3' end processing of messenger RNA precursors. *J Biol Chem* **266**, 3131-9 (1991).

291. Raabe, T., Bollum, F.J. & Manley, J.L. Primary structure and expression of bovine poly(A) polymerase. *Nature* **353**, 229-34 (1991).

292. Bienroth, S., Keller, W. & Wahle, E. Assembly of a processive messenger RNA polyadenylation complex. *EMBO J* **12**, 585-94 (1993).

293. Kuhn, U. et al. Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. *J Biol Chem* **284**, 22803-14 (2009).

294. Wahle, E. & Ruegsegger, U. 3'-End processing of pre-mRNA in eukaryotes. *FEMS Microbiol Rev* **23**, 277-95 (1999).

295. Takagaki, Y. & Manley, J.L. Complex protein interactions within the human polyadenylation machinery identify a novel component. *Mol Cell Biol* **20**, 1515-25 (2000).

296. Xing, H., Mayhew, C.N., Cullen, K.E., Park-Sarge, O.K. & Sarge, K.D. HSF1 modulation of Hsp70 mRNA polyadenylation via interaction with symplekin. *J Biol Chem* **279**, 10551-5 (2004).

297. Wahle, E. Poly(A) tail length control is caused by termination of processive synthesis. *J Biol Chem* **270**, 2800-8 (1995).

298. Tian, B., Hu, J., Zhang, H. & Lutz, C.S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* **33**, 201-12 (2005).

299. Fu, Y. et al. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res* **21**, 741-7 (2011).

300. Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-9 (2008).

301. Di Giammartino, D.C., Nishida, K. & Manley, J.L. Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43**, 853-66 (2011).

302. Ara, T., Lopez, F., Ritchie, W., Benech, P. & Gautheret, D. Conservation of alternative polyadenylation patterns in mammalian genes. *BMC Genomics* **7**, 189 (2006).

303. Nunes, N.M., Li, W., Tian, B. & Furger, A. A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence. *EMBO J* **29**, 1523-36 (2010).

304. Zhang, H., Lee, J.Y. & Tian, B. Biased alternative polyadenylation in human tissues. *Genome Biol* **6**, R100 (2005).

305. Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A. & Burge, C.B. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* **320**, 1643-7 (2008).

306. Mayr, C. & Bartel, D.P. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**, 673-84 (2009).

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

308. Ji, Z., Lee, J.Y., Pan, Z., Jiang, B. & Tian, B. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc Natl Acad Sci U S A* **106**, 7028-33 (2009).

309. Shepard, P.J. et al. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**, 761-72 (2011).

310. Takagaki, Y. & Manley, J.L. Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation. *Mol Cell* **2**, 761-71 (1998).

311. Takagaki, Y., Seipelt, R.L., Peterson, M.L. & Manley, J.L. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell* **87**, 941-52 (1996).

312. Mavrich, T.N. et al. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* **18**, 1073-83 (2008).

313. Shivaswamy, S. et al. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol* **6**, e65 (2008).

314. Spies, N., Nielsen, C.B., Padgett, R.A. & Burge, C.B. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell* **36**, 245-54 (2009).

315. Barreau, C., Paillard, L. & Osborne, H.B. AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Res* **33**, 7138-50 (2005).

316. Fabian, M.R., Sonenberg, N. & Filipowicz, W. Regulation of mRNA translation and stability by microRNAs. *Annu Rev Biochem* **79**, 351-79 (2010).

317. Edery, I. & Sonenberg, N. Cap-dependent RNA splicing in a HeLa nuclear extract. *Proc Natl Acad Sci U S A* **82**, 7590-4 (1985).

318. Lewis, J.D., Izaurralde, E., Jarmolowski, A., McGuigan, C. & Mattaj, I.W. A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5' splice site. *Genes Dev* **10**, 1683-98 (1996).

319. Ohno, M., Sakamoto, H. & Shimura, Y. Preferential excision of the 5' proximal intron from mRNA precursors with two introns as mediated by the cap structure. *Proc Natl Acad Sci U S A* **84**, 5187-91 (1987).

320. Shibagaki, Y., Itoh, N., Yamada, H., Nagata, S. & Mizumoto, K. mRNA capping enzyme. Isolation and characterization of the gene encoding mRNA guanylytransferase subunit from Saccharomyces cerevisiae. *J Biol Chem* **267**, 9521-8 (1992).

321. Hossain, M.A., Claggett, J.M., Nguyen, T. & Johnson, T.L. The cap binding complex influences H2B ubiquitination by facilitating splicing of the SUS1 pre-mRNA. *RNA* **15**, 1515-27 (2009).

322. Kuhn, J.M., Hugouvieux, V. & Schroeder, J.I. mRNA cap binding proteins: effects on abscisic acid signal transduction, mRNA processing, and microarray analyses. *Curr Top Microbiol Immunol* **326**, 139-50 (2008).

323. Hart, R.P., McDevitt, M.A. & Nevins, J.R. Poly(A) site cleavage in a HeLa nuclear extract is dependent on downstream sequences. *Cell* **43**, 677-83 (1985).

324. Cooke, C. & Alwine, J.C. The cap and the 3' splice site similarly affect polyadenylation efficiency. *Mol Cell Biol* **16**, 2579-84 (1996).

325. Le Hir, H., Nott, A. & Moore, M.J. How introns influence and enhance eukaryotic gene expression. *Trends Biochem Sci* **28**, 215-20 (2003).

326. Lu, S. & Cullen, B.R. Analysis of the stimulatory effect of splicing on mRNA production and utilization in mammalian cells. *RNA* **9**, 618-30 (2003).

327. Niwa, M., Rose, S.D. & Berget, S.M. In vitro polyadenylation is stimulated by the presence of an upstream intron. *Genes Dev* **4**, 1552-9 (1990).

328. Niwa, M., MacDonald, C.C. & Berget, S.M. Are vertebrate exons scanned during splice-site selection? *Nature* **360**, 277-80 (1992).

329. Millevoi, S. et al. A novel function for the U2AF 65 splicing factor in promoting pre-mRNA 3'-end processing. *EMBO Rep* **3**, 869-74 (2002).

330. Vagner, S., Vagner, C. & Mattaj, I.W. The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF 65 to couple 3'-end processing and splicing. *Genes Dev* **14**, 403-13 (2000).

331. Millevoi, S. et al. An interaction between U2AF 65 and CF I(m) links the splicing and 3' end processing machineries. *EMBO J* **25**, 4854-64 (2006).

332. McCracken, S., Lambermon, M. & Blencowe, B.J. SRm160 splicing coactivator promotes transcript 3'-end cleavage. *Mol Cell Biol* **22**, 148-60 (2002).

333. McCracken, S., Longman, D., Johnstone, I.L., Caceres, J.F. & Blencowe, B.J. An evolutionarily conserved role for SRm160 in 3'-end processing that functions independently of exon junction complex formation. *J Biol Chem* **278**, 44153-60 (2003).

334. Lutz, C.S. et al. Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency in vitro. *Genes Dev* **10**, 325-37 (1996).

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

336. Gunderson, S.I., Polycarpou-Schwarz, M. & Mattaj, I.W. U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70K and poly(A) polymerase. *Mol Cell* **1**, 255-64 (1998).

337. Ashe, M.P., Furger, A. & Proudfoot, N.J. Stem-loop 1 of the U1 snRNP plays a critical role in the suppression of HIV-1 polyadenylation. *RNA* **6**, 170-7 (2000).

338. Kaida, D. et al. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664-8 (2010).

339. Box, J.A., Bunch, J.T., Tang, W. & Baumann, P. Spliceosomal cleavage generates the 3' end of telomerase RNA. *Nature* **456**, 910-4 (2008).

340. Gentles, A.J. & Karlin, S. Why are human G-protein-coupled receptors predominantly intronless? *Trends Genet* **15**, 47-9 (1999).

341. Guang, S., Felthauser, A.M. & Mertz, J.E. Binding of hnRNP L to the pre-mRNA processing enhancer of the herpes simplex virus thymidine kinase gene enhances both polyadenylation and nucleocytoplasmic export of intronless mRNAs. *Mol Cell Biol* **25**, 6303-13 (2005).

342. Huang, Z.M. & Yen, T.S. Role of the hepatitis B virus posttranscriptional regulatory element in export of intronless transcripts. *Mol Cell Biol* **15**, 3864-9 (1995).

343. Zang, W.Q. & Yen, T.S. Distinct export pathway utilized by the hepatitis B virus posttranscriptional regulatory element. *Virology* **259**, 299-304 (1999).

344. Donello, J.E., Loeb, J.E. & Hope, T.J. Woodchuck hepatitis virus contains a tripartite posttranscriptional regulatory element. *J Virol* **72**, 5085-92 (1998).

345. Conrad, N.K. & Steitz, J.A. A Kaposi's sarcoma virus RNA element that increases the nuclear abundance of intronless transcripts. *EMBO J* **24**, 1831-41 (2005).

346. Guang, S. & Mertz, J.E. Pre-mRNA processing enhancer (PPE) elements from intronless genes play additional roles in mRNA biogenesis than do ones from intron-containing genes. *Nucleic Acids Res* **33**, 2215-26 (2005).

347. Huang, Y. & Steitz, J.A. Splicing factors SRp20 and 9G8 promote the nucleocytoplasmic export of mRNA. *Mol Cell* **7**, 899-905 (2001).

348. Huang, Y., Wimler, K.M. & Carmichael, G.G. Intronless mRNA transport elements may affect multiple steps of pre-mRNA processing. *EMBO J* **18**, 1642-52 (1999).

349. Liu, X. & Mertz, J.E. HnRNP L binds a cis-acting RNA sequence element that enables intron-dependent gene expression. *Genes Dev* **9**, 1766-80 (1995).

350. Blanchette, M., Labourier, E., Green, R.E., Brenner, S.E. & Rio, D.C. Genome-wide analysis reveals an unexpected function for the Drosophila splicing factor U2AF50 in the nuclear export of intronless mRNAs. *Mol Cell* **14**, 775-86 (2004).

351. Marchuk, D., Drumm, M., Saulino, A. & Collins, F.S. Construction of T-vectors, a rapid and general system for direct cloning of unmodified PCR products. *Nucleic Acids Res* **19**, 1154 (1991).

352. Lee, J.Y., Yeh, I., Park, J.Y. & Tian, B. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res* **35**, D165-8 (2007).

353. Lee, J.Y., Park, J.Y. & Tian, B. Identification of mRNA polyadenylation sites in genomes using cDNA sequences, expressed sequence tags, and Trace. *Methods in molecular biology (Clifton, N.J.)* **419**, 23-37 (2008).

354. Lee, J.Y., Ji, Z. & Tian, B. Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. *Nucleic Acids Res* **36**, 5581-90 (2008).

355. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).

356. Proudfoot, N.J. Ending the message: poly(A) signals then and now. *Genes Dev* **25**, 1770-82 (2011).

357. Minneman, K.P. Splice variants of G protein-coupled receptors. *Mol Interv* **1**, 108-16 (2001).

358. Kroeze, W.K., Sheffler, D.J. & Roth, B.L. G-protein-coupled receptors at a glance. *J Cell Sci* **116**, 4867-9 (2003).

359. Fredriksson, R., Lagerstrom, M.C., Lundin, L.G. & Schioth, H.B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol* **63**, 1256-72 (2003).

360. Mountjoy, K.G., Robbins, L.S., Mortrud, M.T. & Cone, R.D. The cloning of a family of genes that encode the melanocortin receptors. *Science* **257**, 1248-51 (1992).

361. Chhajlani, V. & Wikberg, J.E. Molecular cloning and expression of the human melanocyte stimulating hormone receptor cDNA. *FEBS Lett* **309**, 417-20 (1992).

362. Gantz, I. et al. Molecular cloning, expression, and gene localization of a fourth melanocortin receptor. *J Biol Chem* **268**, 15174-9 (1993).

363. Gantz, I. et al. Molecular cloning of a novel melanocortin receptor. *J Biol Chem* **268**, 8246-50 (1993).

364. Gantz, I. et al. Molecular cloning, expression, and characterization of a fifth melanocortin receptor. *Biochem Biophys Res Commun* **200**, 1214-20 (1994).

365. Mountjoy, K.G., Mortrud, M.T., Low, M.J., Simerly, R.B. & Cone, R.D. Localization of the melanocortin-4 receptor (MC4-R) in neuroendocrine and autonomic control circuits in the brain. *Mol Endocrinol* **8**, 1298-308 (1994).

366. Roselli-Rehfuss, L. et al. Identification of a receptor for gamma melanotropin and other proopiomelanocortin peptides in the hypothalamus and limbic system. *Proc Natl Acad Sci U S A* **90**, 8856-60 (1993).

367. Irani, B.G. & Haskell-Luevano, C. Feeding effects of melanocortin ligands--a historical perspective. *Peptides* **26**, 1788-99 (2005).

368. Todorovic, A. & Haskell-Luevano, C. A review of melanocortin receptor small molecule ligands. *Peptides* **26**, 2026-36 (2005).

369. Sundaramurthy, D., Campbell, D.A., Leek, J.P., Markham, A.F. & Pieri, L.F. Assignment of the melanocortin 4 receptor (MC4R) gene to human chromosome band 18q22 by in situ hybridisation and radiation hybrid mapping. *Cytogenet Cell Genet* **82**, 97-8 (1998).

370. Liu, H. et al. Transgenic mice expressing green fluorescent protein under the control of the melanocortin-4 receptor promoter. *J Neurosci* **23**, 7143-54 (2003).

371. Kishi, T. et al. Expression of melanocortin 4 receptor mRNA in the central nervous system of the rat. *J Comp Neurol* **457**, 213-35 (2003).

372. Daniel, P.B. et al. 1 kb of 5' flanking sequence from mouse MC4R gene is sufficient for tissue specific expression in a transgenic mouse. *Mol Cell Endocrinol* **239**, 63-71 (2005).

373. Cone, R.D. The Central Melanocortin System and Energy Homeostasis. *Trends Endocrinol Metab* **10**, 211-216 (1999).

374. Cone, R.D. Anatomy and regulation of the central melanocortin system. *Nat Neurosci* **8**, 571-8 (2005).

375. Coll, A.P., Farooqi, I.S., Challis, B.G., Yeo, G.S. & O'Rahilly, S. Proopiomelanocortin and energy balance: insights from human and murine genetics. *J Clin Endocrinol Metab* **89**, 2557-62 (2004).

376. Huszar, D. et al. Targeted disruption of the melanocortin-4 receptor results in obesity in mice. *Cell* **88**, 131-41 (1997).

377. Ste Marie, L., Miura, G.I., Marsh, D.J., Yagaloff, K. & Palmiter, R.D. A metabolic defect promotes obesity in mice lacking melanocortin-4 receptors. *Proc Natl Acad Sci U S A* **97**, 12339-44 (2000).

378. Vaisse, C., Clement, K., Guy-Grand, B. & Froguel, P. A frameshift mutation in human MC4R is associated with a dominant form of obesity. *Nat Genet* **20**, 113-4 (1998).

379. Yeo, G.S. et al. A frameshift mutation in MC4R associated with dominantly inherited human obesity. *Nat Genet* **20**, 111-2 (1998).

380. Butler, A.A. et al. Melanocortin-4 receptor is required for acute homeostatic responses to increased dietary fat. *Nat Neurosci* **4**, 605-11 (2001).

381. Natalizio, B.J., Muniz, L.C., Arhin, G.K., Wilusz, J. & Lutz, C.S. Upstream elements present in the 3'-untranslated region of collagen genes influence the processing efficiency of overlapping polyadenylation signals. *J Biol Chem* **277**, 42733-40 (2002).

382. Wilusz, J. & Shenk, T. A 64 kd nuclear protein binds to RNA segments that include the AAUAAA polyadenylation motif. *Cell* **52**, 221-8 (1988).

383. Anand, S., Batista, F.D., Tkach, T., Efremov, D.G. & Burrone, O.R. Multiple transcripts of the murine immunoglobulin epsilon membrane locus are generated by alternative splicing and differential usage of two polyadenylation sites. *Mol Immunol* **34**, 175-83 (1997).

384. Conway, L. & Wickens, M. Analysis of mRNA 3' end formation by modification interference: the only modifications which prevent processing lie in AAUAAA and the poly(A) site. *EMBO J* **6**, 4177-84 (1987).

385. Hess, J., Angel, P. & Schorpp-Kistner, M. AP-1 subunits: quarrel and harmony among siblings. *J Cell Sci* **117**, 5965-73 (2004).

386. Glover, J.N. & Harrison, S.C. Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature* **373**, 257-61 (1995).

387. Ameyar, M., Wisniewska, M. & Weitzman, J.B. A role for AP-1 in apoptosis: the case for and against. *Biochimie* **85**, 747-52 (2003).

388. Rees, J.L. The genetics of sun sensitivity in humans. *Am J Hum Genet* **75**, 739-51 (2004).

389. Niwa, M. & Berget, S.M. Mutation of the AAUAAA polyadenylation signal depresses in vitro splicing of proximal but not distal introns. *Genes Dev* **5**, 2086-95 (1991).

390. Moreira, A., Wollerton, M., Monks, J. & Proudfoot, N.J. Upstream sequence elements enhance poly(A) site efficiency of the C2 complement gene and are phylogenetically conserved. *Embo J* **14**, 3809-19 (1995).

391. Guo, Z. & Sherman, F. 3'-end-forming signals of yeast mRNA. *Trends Biochem Sci* **21**, 477-81 (1996).

392. Rigo, F., Kazerouninia, A., Nag, A. & Martinson, H.G. The RNA tether from the poly(A) signal to the polymerase mediates coupling of transcription to cleavage and polyadenylation. *Mol Cell* **20**, 733-45 (2005).

393. Phillips, C., Pachikara, N. & Gunderson, S.I. U1A inhibits cleavage at the immunoglobulin M heavy-chain secretory poly(A) site by binding between the two downstream GU-rich regions. *Mol Cell Biol* **24**, 6162-71 (2004).

394. Shell, S.A., Hesse, C., Morris, S.M., Jr. & Milcarek, C. Elevated levels of the 64-kDa cleavage stimulatory factor (CstF-64) in lipopolysaccharide-stimulated macrophages influence gene expression and induce alternative poly(A) site selection. *J Biol Chem* **280**, 39950-61 (2005).

395. Veraldi, K.L. et al. hnRNP F influences binding of a 64-kilodalton subunit of cleavage stimulation factor to mRNA precursors in mouse B cells. *Mol Cell Biol* **21**, 1228-38 (2001).

396. Shankarling, G.S., Coates, P.W., Dass, B. & Macdonald, C.C. A family of splice variants of CstF-64 expressed in vertebrate nervous systems. *BMC Mol Biol* **10**, 22 (2009).

397. Dass, B. et al. Loss of polyadenylation protein tauCstF-64 causes spermatogenic defects and male infertility. *Proc Natl Acad Sci U S A* **104**, 20374-9 (2007).

398. Wallace, A.M. et al. Two distinct forms of the 64,000 Mr protein of the cleavage stimulation factor are expressed in mouse male germ cells. *Proc Natl Acad Sci U S A* **96**, 6763-8 (1999).

399. McMahon, K.W., Hirsch, B.A. & MacDonald, C.C. Differences in polyadenylation site choice between somatic and male germ cells. *BMC Mol Biol* **7**, 35 (2006).

400. Liu, D. et al. Systematic variation in mRNA 3'-processing signals during mouse spermatogenesis. *Nucl. Acids Res.* **35**, 234-246 (2007).

401. Kim, H.S. et al. DNA damage-induced BARD1 phosphorylation is critical for the inhibition of messenger RNA processing by BRCA1/BARD1 complex. *Cancer Res* **66**, 4561-5 (2006).

402. Kleiman, F.E. & Manley, J.L. The BARD1-CstF-50 interaction links mRNA 3' end formation to DNA damage and tumor suppression. *Cell* **104**, 743-53 (2001).

403. Mirkin, N. et al. The 3' processing factor CstF functions in the DNA repair response. *Nucleic Acids Res* **36**, 1792-804 (2008).