The EMBO Journal (2010) 29, 1523–1536 | © 2010 European Molecular Biology Organization | Some Rights Reserved 0261-4189/10
www.embojournal.org

THE EMBO JOURNAL

EMBO open

A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence

Nuno Miguel Nunes[1], Wencheng Li[2], Bin Tian[2] and André Furger[1,*]

[1]Laboratory of Genes and Development, Department of Biochemistry, University of Oxford, Oxford, UK and [2]Department of Biochemistry and Molecular Biology, UMDNJ-New Jersey Medical School, Newark, NJ, USA

We have analysed the sequences required for cleavage and polyadenylation in the intronless melanocortin 4 receptor (MC4R) pre-mRNA. Unlike other intronless genes, 3′end processing of the MC4R primary transcript is independent of any auxiliary sequence elements and only requires the core poly(A) sequences. Mutation of the AUUAAA hexamer had little effect on MC4R 3′end processing but small changes in the short DSE severely reduced cleavage efficiency. The MC4R poly(A) site requires only the DSE and an A-rich upstream sequence to direct efficient cleavage and polyadenylation. Our observation may be highly relevant for the understanding of how human noncanonical poly(A) sites are recognised. This is supported by a genome-wide analysis of over 10 000 poly(A) sites where we show that many human noncanonical poly(A) signals contain A-rich upstream sequences and tend to have a higher frequency of U and GU nucleotides in their DSE compared with canonical poly(A) signals. The importance of A-rich elements for noncanonical poly(A) site recognition was confirmed by mutational analysis of the human *JUNB* gene, which contains an A-rich noncanonical poly(A) signal.

The EMBO Journal (2010) **29,** 1523–1536. doi:10.1038/emboj.2010.42; Published online 25 March 2010
Subject Categories: RNA
Keywords: 3′end processing; intronless genes; melanocortin 4 receptor; noncanonical poly(A) sites

Introduction

3′end formation is a fundamental processing step for the maturation of mRNAs in eukaryotes. All protein encoding primary transcripts are cleaved at their 3′end and, with the exception of replication-dependent metazoan histone genes, are subsequently subjected to polyadenylation resulting in mature transcripts with characteristic poly(A) tails. (Proudfoot *et al*, 2002; Soller, 2006).

*Corresponding author. Laboratory of Genes and Development, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK. Tel.: +44 0 1865 613 261; Fax: +44 0 1865 613 276; E-mail: andre.furger@bioch.ox.ac.uk

Received: 5 October 2009; accepted: 3 March 2010; published online: 25 March 2010

Cleavage and polyadenylation define critical biochemical events in eukaryotic gene expression promoting transcription termination, terminal intron removal, nuclear–cytoplasmic transport, translation initiation and stability of the mRNA (Proudfoot *et al*, 2002; Proudfoot, 2004). It is therefore not surprising that 3′end formation represents a crucial regulatory step in eukaryotic gene expression and improper 3′end processing of pre-mRNAs is associated with a number of human diseases (Danckwardt *et al*, 2008).

The co-transcriptional cleavage and polyadenylation reaction is directed by a large multi-protein complex, which in humans can constitute more than 80 proteins (Shi *et al*, 2009). The key subunits of this large complex are evolutionarily highly conserved and in mammals are represented by four multi-protein components, the cleavage and polyadenylation specificity factor (CPSF), the cleavage stimulation factor (CstF), cleavage factors I and II (CFI_m, $CFII_m$) and the single polypeptide poly(A) polymerase (PAP). In mammals, assembly of the 3′end processing complex is initiated by the cooperative interaction of CPSF and CstF with specific core sequences on the pre-mRNA (Zhao *et al*, 1999; Proudfoot *et al*, 2002). This core poly(A) sequence is a bipartite sequence element, which in humans, in 70–80%, is defined by a conserved canonical AAUAAA or AUUAAA upstream hexamer motif, recognised by CPSF, and a somewhat less defined downstream U- or GU-rich region (DSE) contacted by CstF (Zarudnaya *et al*, 2003; Tian *et al*, 2005). *In vitro*, these initial contacts are critical for the recruitment of additional key factors including CFI_m, $CFII_m$ and PAP. The efficiency of assembly and subsequent processing is strongly influenced by both the core sequence composition and the spatial arrangement (Zhao *et al*, 1999).

In addition, in many mammalian genes, further auxiliary sequence elements that influence complex assembly can be found both upstream (USEs) and downstream of the core sequences. Furthermore, extensive protein–protein 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 (Vagner *et al*, 2000; Millevoi *et al*, 2002, 2006; Kyburz *et al*, 2006).

The cleavage and polyadenylation machineries of plants, yeast and mammals share many homologies (Hunt, 2008) but the sequence elements directing the assembly of the processing complexes differ in parts. Although the canonical A(A/U)UAAA hexamer sequences can be found upstream of the cleavage sites in both yeast and plants, there appears to be fewer constraints regarding their sequence composition. 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 (Zhao *et al*, 1999; Hunt, 2008). This is in stark

contrast to mammalian hexamer sequences that are generally highly intolerant to sequence alterations (Wickens and Stephenson, 1984; Sheets *et al*, 1990).

Interestingly, 20–30% of human genes do not contain canonical hexamers and 3′end processing is directed by noncanonical sequences (Zarudnaya *et al*, 2003; Tian *et al*, 2005). In addition, noncanonical poly(A) sites appear to be more frequent in genes that undergo alternative cleavage and polyadenylation (Tian *et al*, 2005 and this study) suggesting that they may have a critical role in this process. As about half of all human protein encoding genes undergo alternative cleavage and polyadenylation, it is somewhat surprising that we currently only have a poor understanding how noncanonical poly(A) sites are recognised and regulated.

It has been shown that at least in some cases the interaction of additional core factors, such as CFI_m with UGUAN motifs located in the 3′UTR, can promote cleavage and polyadenylation at such 3′end processing sites (Venkataraman *et al*, 2005). Interestingly, the sequence composition of these noncanonical poly(A) sites was proposed to be much more similar to those described for plants and yeast (Venkataraman *et al*, 2005).

Similar to noncanonical poly(A) sites, 3′end processing of mammalian and viral intronless primary transcripts also appears to require additional auxiliary *cis*-elements, perhaps compensating for the lack of enhancement by the splicing factors observed in spliced genes. Several so-called pre-mRNA processing enhancer elements (PPE) located upstream of the core poly(A) sequences have been described for many viral and mammalian non-spliced genes (Huang and Carmichael, 1997; Conrad and Steitz, 2005; Guang and Mertz, 2005).

In this work we analysed how 3′end processing is regulated in the human melanocortin 4 receptor gene (*MC4R*). MC4R is a 333 amino acid 7 transmembrane domain protein encoded by a single exon gene. The *MC4R* gene is expressed in multiple sites in the brain (Liu *et al*, 2003) and mutations in this gene have been associated with obesity (Huszar *et al*, 1997).

Our results show that 3′end processing in the intronless MC4R primary transcript, unlike other intronless genes, does not require any auxiliary sequence elements located either in the 3′UTR or in the 3′flanking regions. Optimal cleavage at the MC4R poly(A) site relies solely on the core poly(A) sequences. The relatively short MC4R DSE is shown to be the most critical element and is able to direct efficiently 3′end cleavage independent of the upstream AUUAAA hexamer. We show that the MC4R DSE, similar to both yeast and plant poly(A) sites, only requires an A-rich sequence upstream of the cleavage site for optimal 3′end processing. Our bioinformatics analysis highlights the significance of this finding by showing that many noncanonical poly(A) sites contain A-rich upstream sequences. Importantly, this analysis further shows that such A-rich 3′end processing sites correlate with an increased frequency of U/GU sequences in their DSE compared with poly(A) sites that contain the canonical A(A/U)UAAA hexamers. These observations could be significant for the understanding of how noncanonical poly(A) sites are recognised by the 3′end processing machinery. Indeed, we show that the A-rich sequence in the JUNB pre-mRNA is critical for cleavage and polyadenylation at its noncanonical poly(A) site. Thus, we propose that many noncanonical poly(A) sites may, similar to MC4R, predominantly rely on a potent DSE mediating a tight interaction with CstF. This interaction may in turn be critical to facilitate recruitment of CPSF to less favourable A-rich upstream sequences and so afford efficient 3′end processing of pre-mRNAs in the absence of canonical hexamers.

Results

MC4R pre-mRNA 3′end formation does not require additional sequence elements located in the 3′flank or 3′UTR

As terminal intron removal has long been known to contribute significantly to the efficiency of cleavage and polyadenylation, we addressed how efficient 3′end processing is achieved in naturally intronless genes. For this, we analysed 3′end formation in the human *MC4R* gene. The analysis of viral and some human intronless genes suggested that these pre-mRNAs generally rely on auxiliary sequences located in the 3′UTR or 3′flanking regions to direct efficient 3′end processing (Huang and Carmichael, 1997; Conrad and Steitz, 2005; Guang and Mertz, 2005; Dalziel *et al*, 2007). To identify auxiliary *cis*-elements required for cleavage and polyadenylation of the MC4R pre-mRNA, we used several reporter plasmids. We first cloned 1.3 kb of sequences located downstream of the MC4R stop codon including the 3′UTR, and 1044 nucleotides of 3′flanking sequence into a cytomegalovirus (CMV) promoter-driven reporter gene that contained 5′untranslated region (5′UTR) sequences from the *MC1R* gene and the green fluorescence protein (GFP) open reading frame (ORF) (Figure 1A).

This original MC4R reporter plasmid was transfected into HEK 293 cells and total RNA was subsequently isolated and analysed by RNAse protection (RP) to map the MC4R 3′end processing site. This was necessary because the poly(A) site of MC4R is not annotated and sequence comparison shows at least two potential 3′end processing sites located in the first 400 nucleotides downstream of the MC4R stop codon (Figure 1A: P1, P2). Our analysis mapped the poly(A) cleavage site to a position 292 nucleotides downstream of the MC4R ORF (Figure 1A: P1). This initial mapping of the processing site was subsequently confirmed by 3′RACE (data not shown). It is worth noting that we also confirmed that a hexamer overlapping the endogenous MC4R stop codon and the first nucleotides within the 3′UTR is not functional (Supplementary Figure 1: A and B). The above described 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 not processed at all (rt) were detected (Figure 1B: lane 2: P2, rt). To verify the presence of potential auxiliary sequence elements in the 3′flank we constructed and analysed three additional plasmids with gradually shorter 3′flanking sequences compared with the full-length clone F (Figure 1A: FΔ1, FΔ2, FΔ3). For the RP analysis, we used an antisense riboprobe complementary to sequences overlapping P1 and P2, which results in protected bands of the same length for all clones representing transcripts cleaved at P1. Note that because of the deletions, transcripts from the FΔ2 and FΔ3 plasmids, which are not processed at P1, would result in protected readthrough bands of different lengths compared with transcripts originating from the F and FΔ1 plasmids (rt(F, FΔ1), rt(FΔ2) and

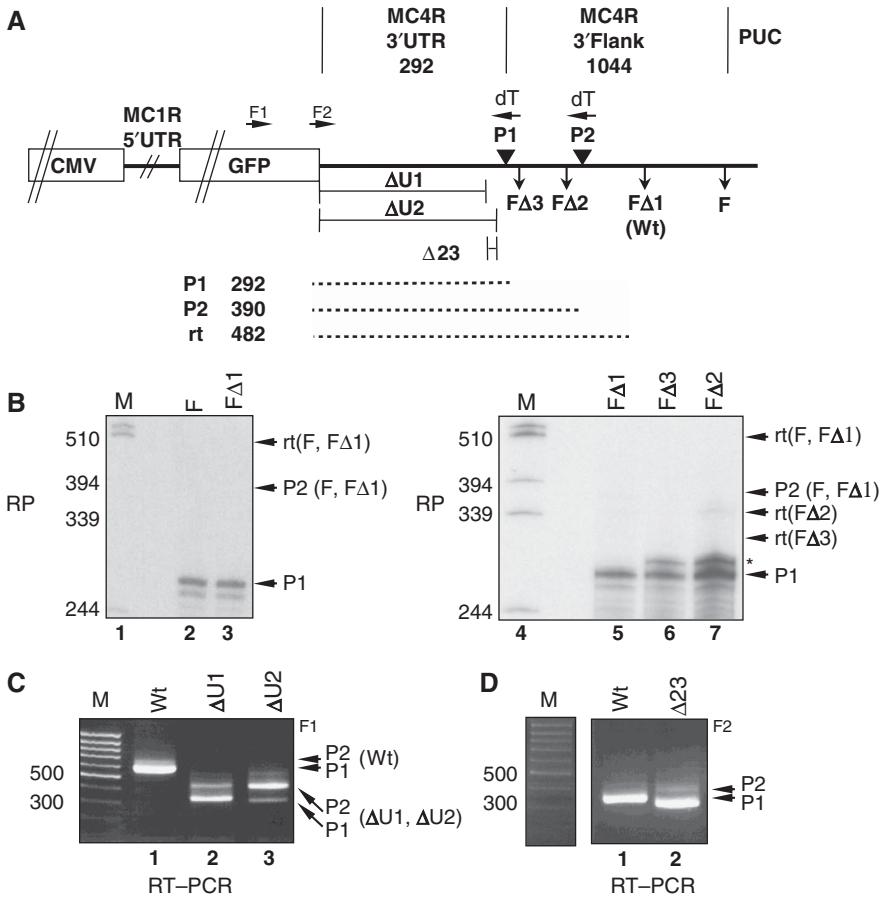


Figure 1 The MC4R poly(A) site does not require auxiliary sequence elements. (**A**) Diagram depicting the MC4R reporter genes: F and deletion of 3′flanking sequences (FΔ1, FΔ2 and FΔ3). Vertical arrows indicate end of the deletion clones relative to Wt sequence. The borders between ORF, 3′UTR, 3′flank and vector backbone are indicated by thin straight vertical lines. Promoter (CMV) and the 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 regions deleted in clones ΔU1, ΔU2 and Δ23 are indicated below the graph. RP fragments uncleaved (rt), cleaved at P1 (P1) or cleaved at P2 (P2) are shown as dotted lines and the expected lengths are indicated. The positions of the F1 and F2 forward primers used in the RT–PCR analysis shown in (**C**) and (**D**), respectively, are shown above the diagram. (**B**) RP 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(FΔ2), rt(FΔ3). Alternative cleavage site used at P1 observed with plasmids FΔ2 and FΔ3 is indicated by (*). FΔ1 is subsequently referred to as wild-type (Wt) (**C**, **D**) RT–PCR analysis of constructs containing deletions in the 3′UTR. RT–PCR products corresponding to mRNAs cleaved at either P1 or P2 are indicated for Wt and UTR deletion clones. Size markers are indicated.

rt(FΔ3)). As can be seen in Figure 1B, 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 because no bands can be seen corresponding to transcripts that failed to cleave at P1 (Figure 1B: compare lanes 2 and 3 and lanes 5–7, rt(F, FΔ1), rt(FΔ3), rt(FΔ2) and P2, respectively). However, large deletions of the 3′flank resulted in the appearance of an additional less intense protected band that is likely to be caused by a shift in the site of cleavage in some of the FΔ3 and FΔ2 transcripts (Figure 1B: lanes 6 and 7; *). For all further experiments, we used FΔ1 as the wild-type reference and thus FΔ1 is subsequently referred to as Wt.

We next addressed whether sequences located in the 3′UTR are required for efficient 3′end processing of the MC4R pre-mRNA. To that end, we constructed a plasmid that had almost all 3′UTR sequences removed, retaining only the last 23 nucleotides immediately upstream of the P1 AUUAAA hexamer and a second plasmid that had the AUUAAA directly fused to the GFP stop codon (Figure 1A: ΔU1, ΔU2, respectively). As can be seen in Figure 1C, oligo-dT primed RT–PCR analysis of total RNA isolated from transfected cells showed that ΔU1 only had a marginal effect on P1 usage. In contrast, the deletion of the entire 3′UTR (ΔU2) appeared to dramatically shift the preferred cleavage site from P1 to P2. This initial result suggested that, there is either a 23 nucleotide long enhancer element located immediately upstream of the P1 hexamer or that locating P1 close to the GFP ORF or 5′UTR sequences somehow reduces processing efficiency at the P1 poly(A) site. To clarify this, a third construct was built that retained all but the last 23 nucleotides of the 3′UTR (Figure 1A: Δ23). As can be seen in Figure 1D, deletion of these nucleotides did not result in a significant shift from P1 into P2 usage (compare lanes 1 and 2). From this analysis we concluded that no sequences in the 3′UTR or 3′flank are required to direct efficient cleavage at the MC4R poly(A) site. This analysis also showed that P2 is an additional functional poly(A) site and that transcripts not

cleaved at the P1 poly(A) site are subsequently efficiently cleaved at P2. The switch of cleavage at P1 to cleavage at P2 was thus used to measure effects on poly(A) cleavage at P1.

Analysis of the MC4R P1 core poly(A) signal

The above described analysis suggested that 3′end processing of the MC4R pre-mRNA is directed by nucleotides including and surrounding the core poly(A) sequences. To address the functional relevance of the core sequences, we first focused on the hexamer motifs. The P1 poly(A) site in MC4R contains two potential hexamers, AAGAAA and AUUAAA (see H1 and H2 in Figure 2A). Although the presence of a guanosine at position 3 in hexamers is normally considered as a potent inactivating mutation (Wilusz and Shenk, 1988; Natalizio *et al*, 2002), it has been shown to be active in at least one gene (Anand *et al*, 1997). Thus, reporter constructs were created with mutations destroying either (H1h2, h1H2,) or both (h1h2) hexamers (Figure 2A). Cells were transfected with these plasmids and total RNA was subsequently analysed by RP and RT–PCR. Although mutations of hexamer sequences have long been known to severely impair 3′end processing (Conway and Wickens, 1987; Wilusz *et al*, 1989; Sheets *et al*, 1990), mutating the MC4R hexamer(s) surprisingly had little effect on cleavage at P1 when analysed by RP and RT–PCR (Figure 2B: compare lanes 1–4 in RP and RT–PCR panels). As the hexamer mutations produced unexpected results, we next created two constructs where two or four uridines in the DSE were substituted by cytidines (Figure 2A: d2, d4 respectively). In contrast to mutations of the hexamers, DSEs are generally described as being relatively tolerant to sequence changes (Zhao *et al*, 1999).

Unlike the hexamer mutations described above, substitution of two uridines resulted in a more than 7-fold increase in P2 usage and introduction of two further substitutions caused a significant 11-fold increase in transcripts cleaved at P2 (Figure 2C). Combining the four U to C substitutions in the DSE with hexamer mutations resulted in a five- to seven-fold increase in P2 usage compared with the hexamer mutations alone, again highlighting the importance of the DSE for the recognition of the MC4R P1 poly(A) site (Figure 2D). From these experiments, we 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.

Mutations in an A-rich upstream sequence and the AUUAAA are necessary to inactivate the MC4R P1 poly(A) site

To clarify why the MC4R P1 poly(A) site can tolerate mutations in the AUUAAA hexamer, a series of additional plasmids were constructed. Several mutations were introduced into the previously mentioned 23 nucleotides long upstream sequence (Figure 3A: Wt underlined sequence). Interestingly, this 23 nucleotide long sequence is very A-rich and thus resembles somewhat the PE and NUE elements found in yeast and plant 3′end processing sites. Extensive substitutions of adenosines in this sequence with cytidines had no effect on P1 usage as can be seen in the oligo-dT primed RT–PCR analysis presented in Figure 3B (compare lanes 1–3). However, 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 3B: lanes 4 and 5). Cleavage at P1 could not be rescued in a construct containing hexamer mutations and where the adenosines in the A-rich motif were substituted by guanosines rather than cytidines (Figure 3C). This suggests that the MC4R poly(A) site contains two potential CPSF-binding sites: an upstream A-rich motif and the AUUAAA. Cleavage and polyadenylation appear to be equally well directed by both of these sequences, which implies that an A-rich motif can functionally substitute for the hexamer and direct cleavage and polyadenylation.

Interestingly, a possible molecular mechanism explaining how A-rich noncanonical poly(A) sites can be recognised by the poly(A) machinery has recently been presented. The PAPOLG pre-mRNA lacks a canonical hexamer but instead contains a critical A-rich sequence upstream of the cleavage site. The recognition and function of this poly(A) site was shown to be dependent on UGUAN motifs located in the 3′UTR and the A-rich upstream sequence. The UGUAN motifs were shown 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 (Venkataraman *et al*, 2005).

Therefore, similar to PAPOLG, the presence of four UGUAN sequences in the MC4R 3′UTR could explain why the mutation of the AUUAAA hexamer was tolerated in our earlier experiments. To clarify this issue, G to C substitutions were introduced into the four UGUAN elements present in the MC4R 3′UTR in the wild-type plasmid and in the h1h2 construct (Figure 3A: UCUAN-Wt, UCUAN-h1h2). It is important to note that the UCUAN sequence significantly reduces interaction with CFI_m *in vitro* (Venkataraman *et al*, 2005). Total RNA isolated from transiently transfected cells was analysed by oligo-dT primed RT–PCR. These results show that mutations of the UGUAN elements present in the MC4R 3′UTR had no significant effects on MC4R 3′end processing efficiency at P1 either in the presence or in the absence of a defined hexamer (Figure 3D: compare lanes 1–4: P1, P2).

From these observations, we concluded that the UGUAN elements present in the MC4R 3′UTR do not have 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.

A potent DSE downstream of a stretch of adenosines is sufficient to direct 3′end cleavage

To further analyse the role of individual adenosines in the A-rich upstream region, we constructed additional plasmids that contained single A to G changes in this area in the double hexamer mutant (h1h2*) background (Figure 4A: h1h2*-a1–8). For this RP analysis we used a ‘composite’ antisense riboprobe that can be used for all mutant constructs because it does not contain the 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 200 nt of 3′flanking sequences (Figure 4A). The 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 4A).

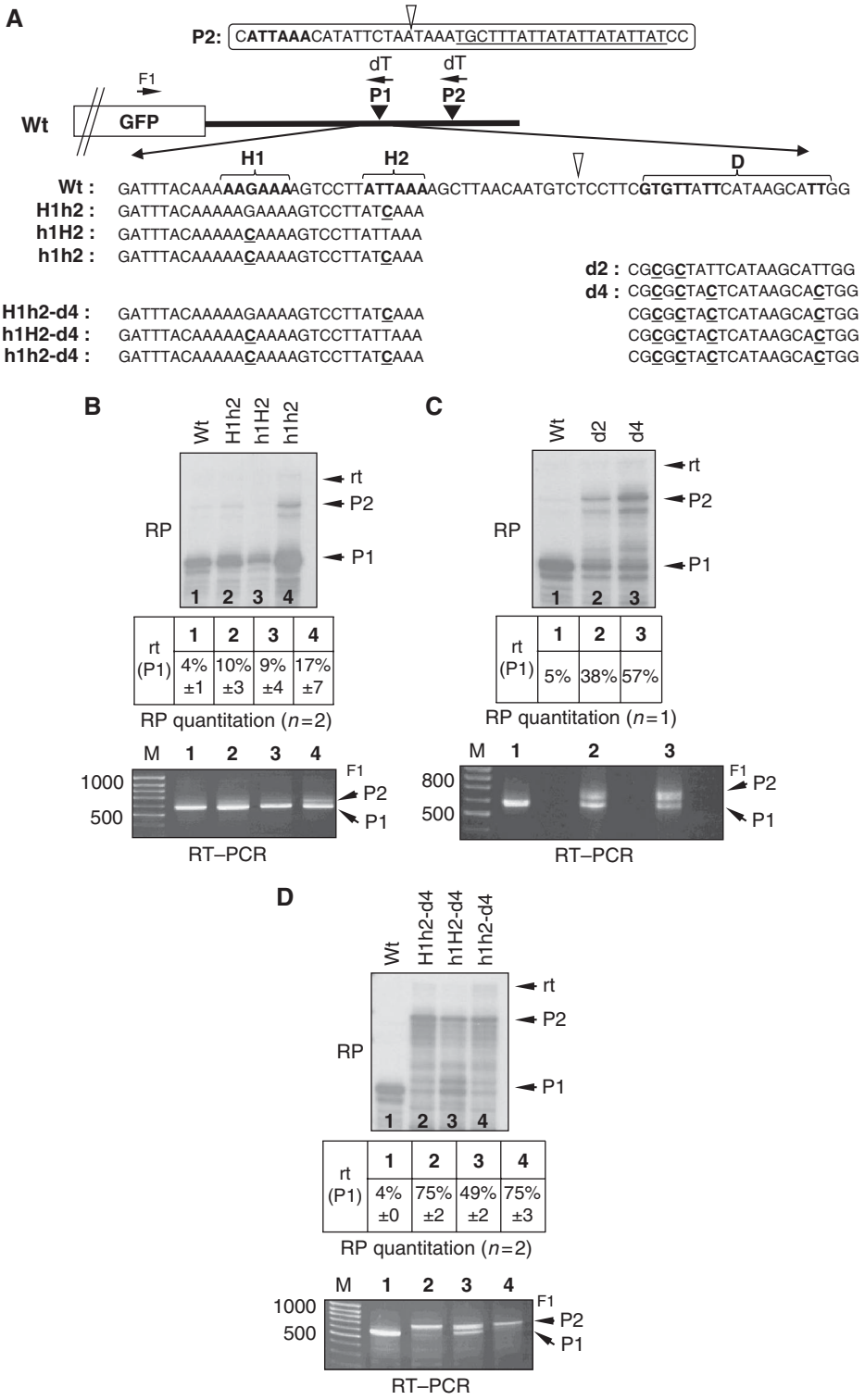


Figure 2 Mutations of the core poly(A) sequences have unexpected effects on cleavage efficiency. (**A**) Diagram depicting the *MC4R* reporter gene where both potential poly(A) sites are indicated by filled triangles. Potential priming of oligo-dT reverse primers (dT) at P1 and P2 and forward primer (F1) is depicted above the diagram. The sequences surrounding P2 are indicated in the box above the graph. The hexamer is in bold, nucleotides of the DSE are underlined and the site of cleavage is indicated by the open triangle. Below the diagram is the Wt sequence surrounding P1, the two hexamers are in bold and indicated by (H1) and (H2), respectively. Capital H represents clones with wild-type hexamer sequences, small letters h (h1 and/or h2) represent mutated hexamers. Open triangle marks the site of cleavage at P1 and the DSE (D) is indicated in bold letter. The mutated nucleotides for each clone are depicted in bold and underlined below the Wt sequence. (**B**–**D**) 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. Quantitations of independent RPs as an average percentage of total P1 readthrough transcripts are shown below.

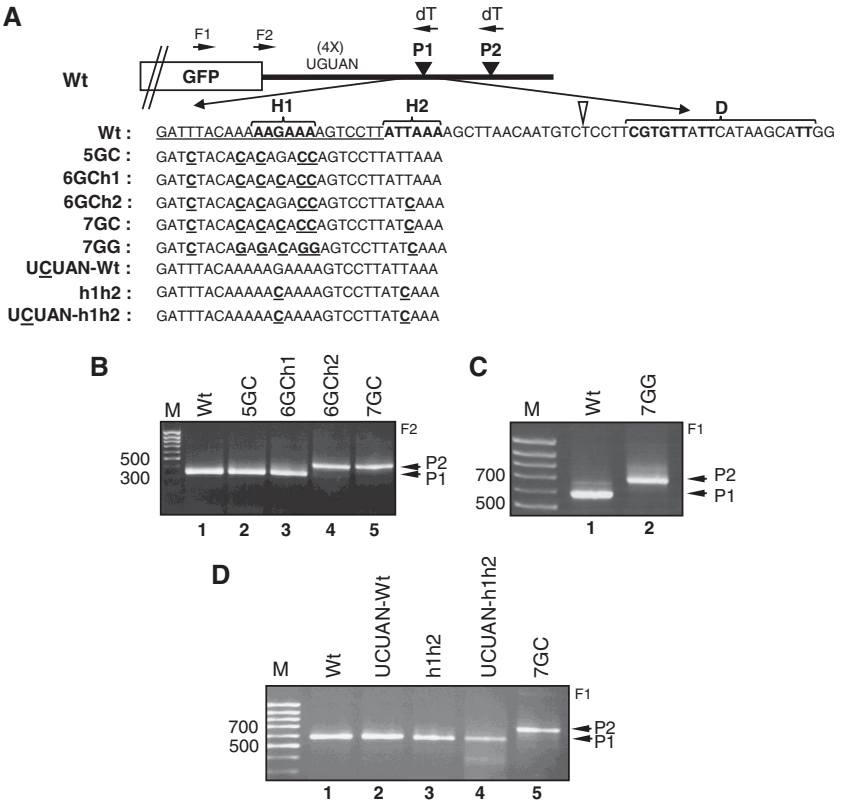


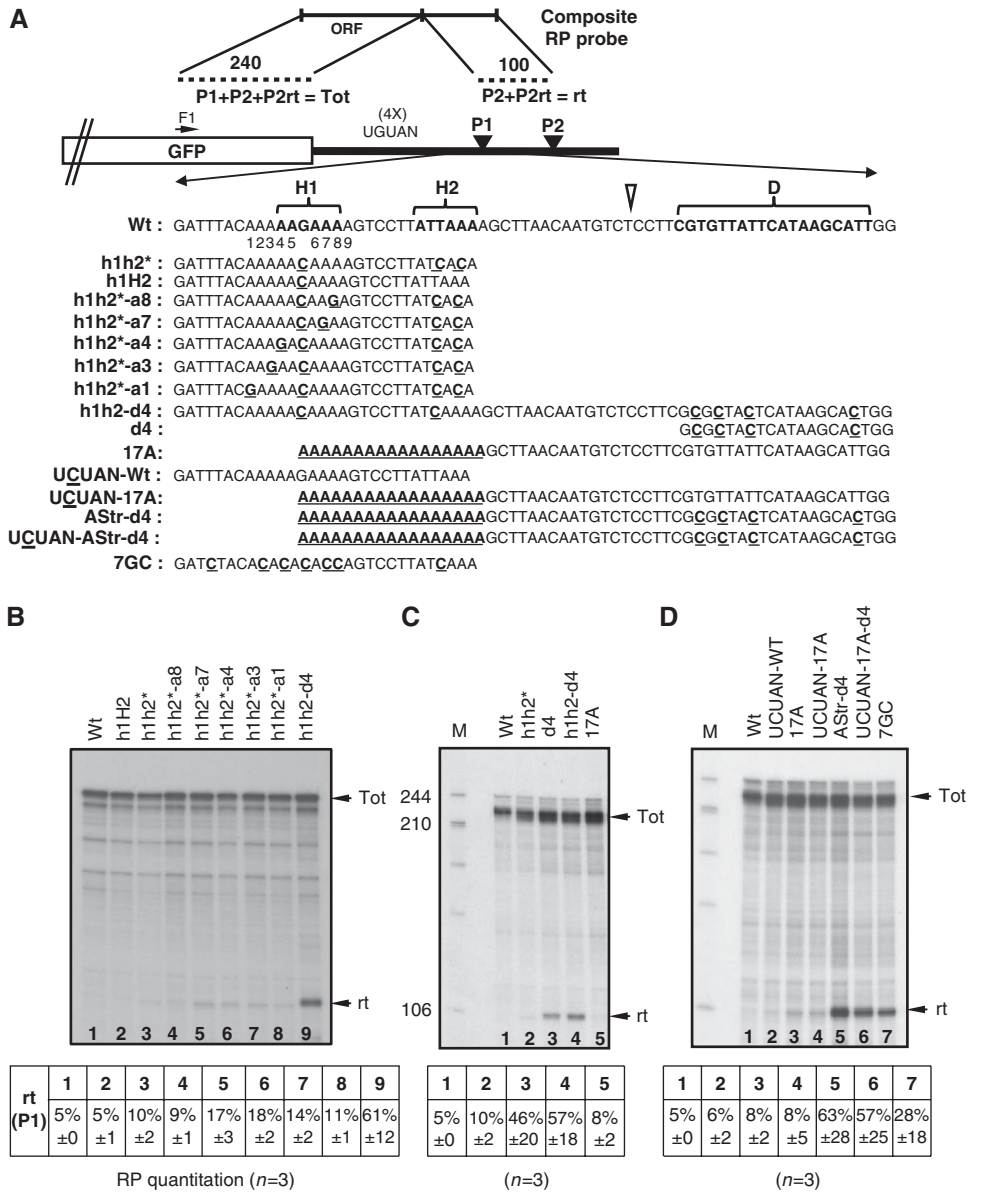
Figure 3 Mutations in the A-rich sequence and the hexamer are required to inactivate MC4R P1. (**A**) The diagram of the *MC4R* reporter gene and the Wt sequence surrounding the P1 poly(A) site is shown, underlined letters represent the 23 nucleotides long upstream sequence. The open triangle marks the site of cleavage at P1 and the DSE (D) 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. 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**–**D**) 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 is indicated on top right of gels.

The use of this composite probe confirmed the earlier results, that mutations of upstream hexamers in the MC4R poly(A) site had little effect on P1 usage (Figure 4B: compare lanes 1–3; Figure 4C: lanes 1 and 2). In contrast, mutations affecting the hexamers and/or the DSE reduced P1 usage dramatically, which is shown by the appearance of a second protected band representing transcripts that are not cleaved at P1 (Figure 4B: compare lanes 1–8 with lane 9; Figure 4C: lanes 3 and 4). The overall effects of mutating individual adenosines in the h1h2* background compared with h1h2-d4 clone were modest (Figure 4B: lanes 4–9). Interestingly, mutations of adenosines, which disrupt the stretches of five or four consecutive adenosines surrounding H1 (Figure 4A), consistently affected cleavage at P1 more severely than changing adenosines at the periphery of either stretch (Figure 4B: compare rt in lanes 4 and 8 with 5–7). These results suggested 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 test this hypothesis, we created a plasmid in which both the A-rich region and the AUUAAA were replaced with a stretch of 17 adenosines. From this parental 17A construct two more plasmids 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 (Supplementary Figure 1C). The use of the composite probe for the analysis of the 17A plasmids was essential as this probe avoids an experimental artefact that was 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 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, as miss-priming at the 17 nt A-stretch by oligo-dT (17T) readily occurred and resulted in a slightly faster migrating band (Supplementary Figure 1D: lane 2). The composite probe in contrast avoids these problems and allows accurate quantification of effects on cleavage at P1.

The analysis of total RNA isolated from cells transfected with the plasmid containing the 17 adenosine substitution showed that an uninterrupted stretch of adenosines is able to direct cleavage at P1 to levels that 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 4C: compare ‘rt’ in lanes 1 and 5 with lanes 3 and 4). When the A-stretch was substituted with a C-stretch no cleavage is seen at P1 and all detectable transcripts represent mRNAs that are cleaved and polyadenylated at P2 (Supplementary Figure 1D: lanes 1 and 3).

To verify that 3′end processing in the plasmid containing the A-stretch is not dependent on UGUAN sequences found in the 3′UTR, total RNA was analysed from cells transfected



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

RP quantitation (*n*=3)

1	2	3	4	5
5% ±0	10% ±2	46% ±20	57% ±18	8% ±2

(*n*=3)

1	2	3	4	5	6	7
5% ±0	6% ±2	8% ±2	8% ±5	63% ±28	57% ±25	28% ±18

(*n*=3)

Figure 4 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 outline of the composite RP probe is depicted above and the protected fragments are shown as dotted lines. All transcripts result in a 240 nt protected band Tot (P1 + P2 + P2rt) and transcripts not cleaved at P1 (P2 + P2rt) give an additional protected band rt (100 nt). 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**–**D**) RPs 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 three independent RPs for each clone for each gel is given below the gels. Average percentage of the total transcripts that are not cleaved at P1 is set to 5% for the wild type as determined in Figure 2.

with respective constructs where the UGUAN motifs were changed to UCUAN (Figure 4A). This analysis showed that the 17 adenosines positioned upstream of the MC4R DSE can direct efficient cleavage independently of the presence of upstream UGUAN sequences (Figure 4D: compare rt lanes 1–4). In contrast, A-stretch constructs that contained mutations in the DSE resulted in a dramatic increase in transcripts that were not cleaved at P1 (Figure 4D: compare rt lanes 5–7). 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 2C and D. From these results we concluded that, in the context of the MC4R poly(A) signal, a strong DSE and a consecutive stretch of adenosines are sufficient to direct efficient 3′end cleavage.

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

The above described data imply that in the context of a strong DSE, human poly(A) sites may be less dependent on the presence of an 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 with canonical poly(A) sites. To test this hypothesis, we conducted a genome-wide bioinformatics analysis using over 10 000 human poly(A) sites obtained from the PolyA_DB database (Lee *et al*, 2007). As A-rich sequences in a transcript can lead to internal priming for reverse transcription, resulting in false identification of poly(A) sites (Lee *et al*, 2008b), we required that supporting cDNA/EST/Trace sequences for a poly(A) site contained at least 30 nt As/Ts corresponding to the poly(A) tail. As can be seen in Figure 5A, DSEs in poly(A) sites that constitute an A-rich upstream sequence (defined as a hexamer with $\geqslant 5$ 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 with 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 5B). We have found no correlation between the appearance of A-rich noncanonical poly(A) sites and intronless genes (data not shown).

We next tested whether the correlation between U richness and GU richness and upstream A-stretches depends on its overall position within the pre-mRNA. For this we refined the data and distinguished 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 5C) depending on the relative distance of the poly(A) sites to the beginning of the 3′-most exon (Tian *et al*, 2005). Our data show 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 appear to be more robust in genes containing alternative 3′end processing sites compared with genes with a single poly(A) site (Figure 5E: 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 5E: M).

We further compared the conservation patterns of human A-rich and A(A/U)UAAA poly(A) sites in mouse. As shown in Figure 5F, 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, that is 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.

We also analysed mRNA-seq data (Pan *et al*, 2008; Wang *et al*, 2008) and show that 189 A-rich noncanonical 3′-most poly(A) sites found in this data set behave identical to the 2960 canonical 3′-most poly(A) sites regarding the ratios of reads positioned upstream or downstream of the cleavage site (Supplementary Figure 3A). 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 showed that alternative A-rich noncanonical poly(A) sites are more likely to be tissue specifically regulated compared with A(A/U)UAAA alternative poly(A) sites (Supplementary Figure 3B).

From this analysis we concluded that poly(A) sites comprising upstream A-rich elements represent a significant number of noncanonical poly(A) sites and, compared with 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 with canonical poly(A) sites.

JUNB contains an endogenous noncanonical poly(A) site with an upstream A-rich sequence required for optimal cleavage

The above described bioinformatics analysis showed that about one-third of noncanonical poly(A) sites contain an A-rich sequence upstream of the cleavage site. To experimentally verify whether these A-rich sequences represent critical *cis*-elements for cleavage and polyadenylation, we analysed 3′end formation in the *JunB* gene, identified in our bioinformatics analysis (Supplementary data 2).

JUNB encodes a non-spliced transcript with a poly(A) cleavage site annotated 20 nucleotides downstream of an A-rich sequence in the 3′UTR. This annotation is verified by 177 AceView accessions.

To confirm the role of this endogenous A-stretch in poly(A) cleavage, we cloned the 3′UTR and 3′flanking sequences including the DSE of JUNB downstream of the GFP ORF in the MC4R plasmid. Thus, the MC4R 3′UTR and the P1 poly(A) sequences were replaced with equivalent regions from JUNB (Figure 6A). Note that the 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 easily be detected using the composite RP probe described in Figure 4.

First, we used RP to confirm the site of cleavage in our JUNB wild-type plasmid. As can be seen in Figure 6B, cleavage of transfected JUNB resulted in RP products that are slightly longer than the expected lengths (123 nt), which may be due to a slight shift of the cleavage site in our reporter construct. Nevertheless, in the Wt JUNB clone cleavage occurs 5–10 nts downstream of the annotated site 3′ of the A-stretch. We were unable to detect any endogenous JUNB mRNA by RP.

Importantly, the subsequent RP analysis of JUNB wild type and mutant plasmids with the composite probe showed that A to C mutations in the A-stretch reduced cleavage 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 poly(A) site (Figure 6C: compare lanes 1 and 2). Note that the readthrough band in these experiments is 88 rather than 100 nucleotides because JUNB constructs do not contain the MC4R P1 poly(A) signals resulting in a shorter protected readthrough band. A reduc-

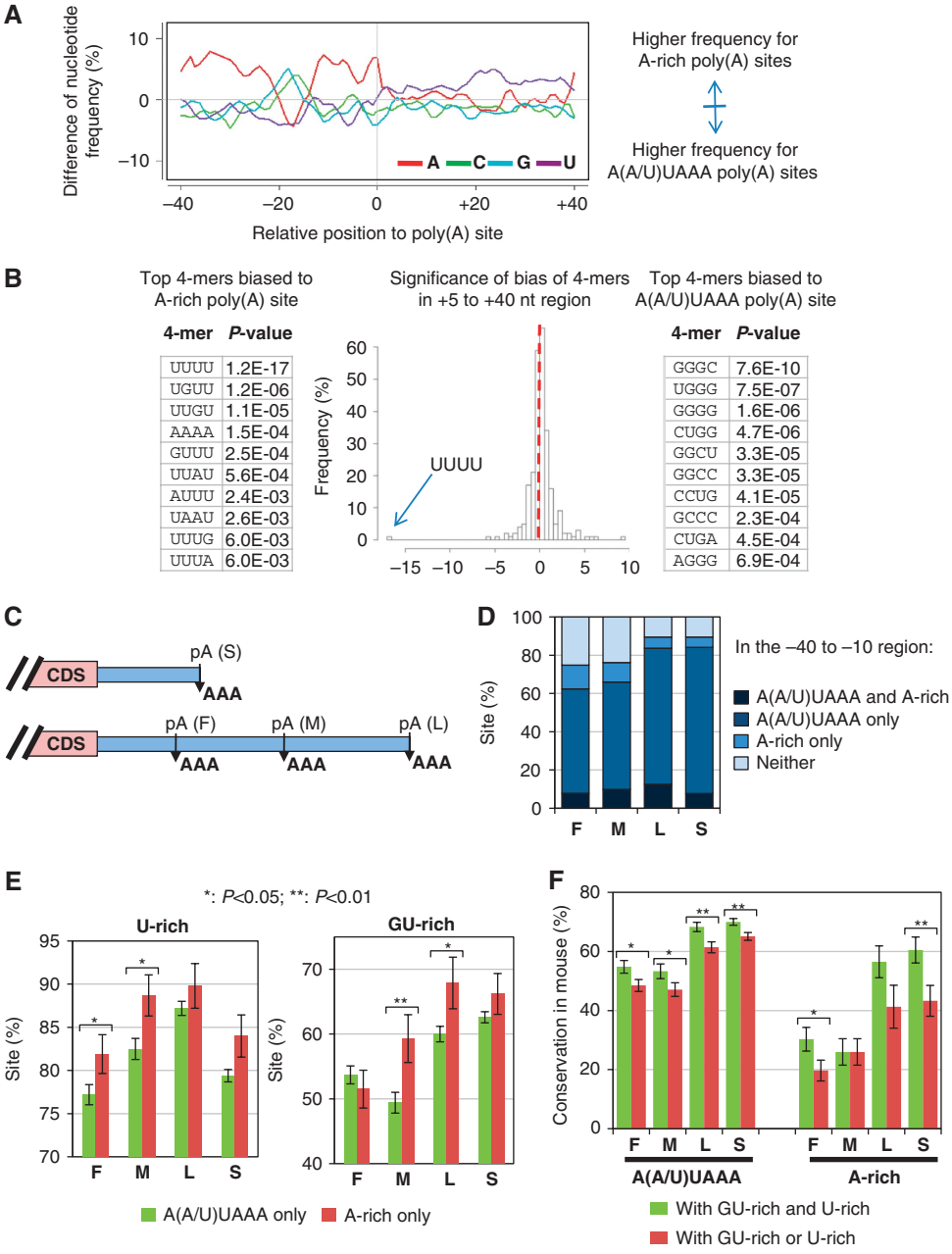


Figure 5 Systematic analysis of poly(A) sites with A(A/U)UAAA and A-rich elements. A-rich elements are hexamers containing at least five As excluding AAUAAA and not overlapping with A(A/U)UAAA. (**A**) Difference in nucleotide frequency surrounding poly(A) sites with A-rich elements vs A(A/U)UAAA poly(A) 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 Materials and 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 is 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 four poly(A) site types. The U-rich or GU-rich sequence elements are described in 'Materials and 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 four poly(A) site types. The error bars are standard deviations.

tion in JUNB poly(A) site use was also observed when U to C mutations were introduced into the DSE (Figure 6C: compare lanes 1 and 3). The effects of the DSE mutations introduced into the JUNB were less dramatic compared to mutations of the A-rich sequence, which is likely to be due to the presence of a G/U-rich putative USE (Figure 6A). Furthermore, transient transfections of the plasmids used had no effects on 3′end processing of endogenous genes (Supplementary

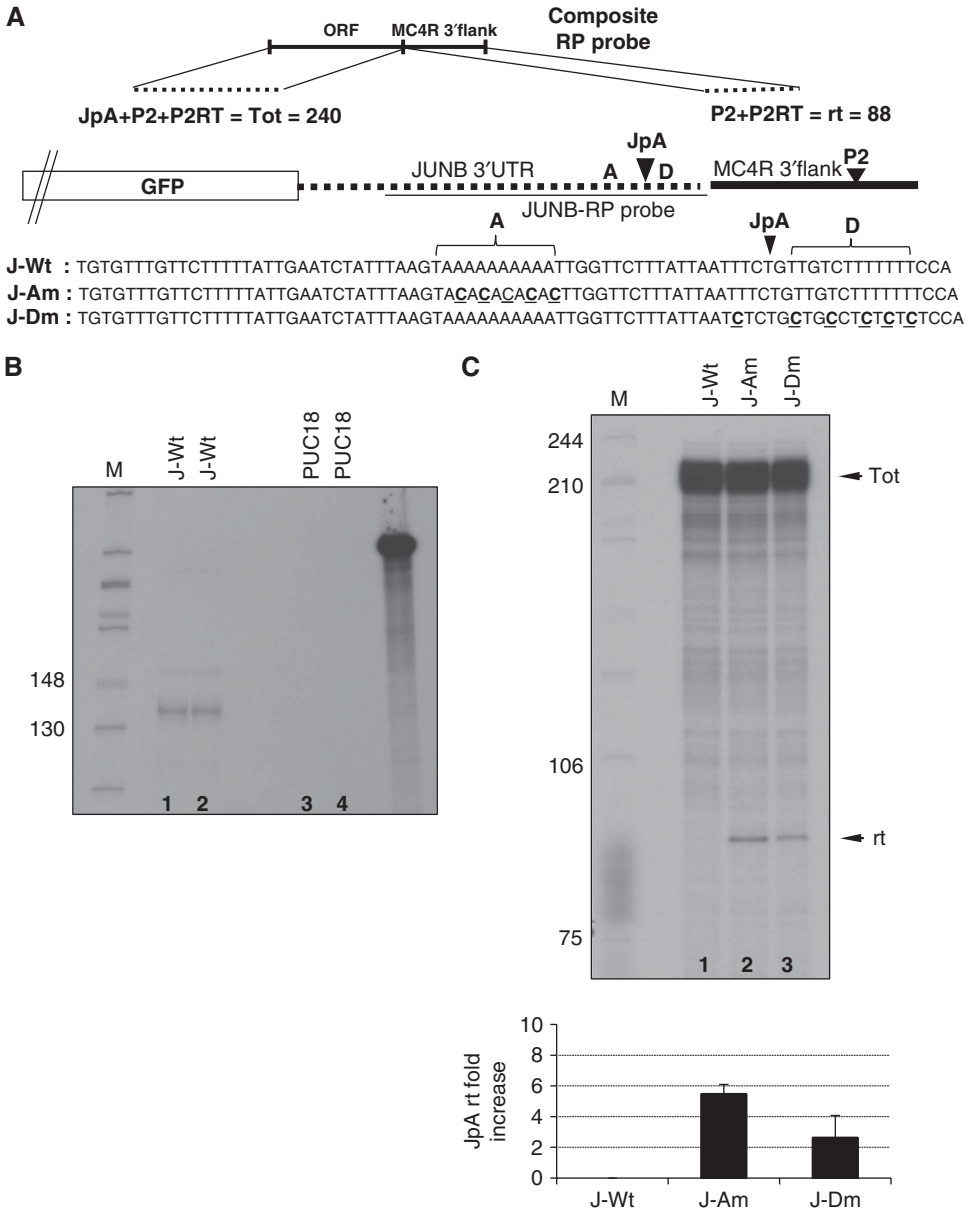


Figure 6 The JUNB pre-mRNAs require an A-rich upstream sequence for efficient cleavage and polyadenylation. (**A**) Diagram showing the *JUNB* reporter gene. Origins of the sequences in the plasmid are indicated. JUNB 3′UTR and 3′flanking regions are represented by a dotted line and the graph shows how 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**) RP of total RNA isolated from cells transfected with JUNB Wt and mutant plasmids. Quantitation ($n = 3$) is presented as fold increase of transcripts that are not cleaved at JpA.

Figure 4). 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 suggestion that A-rich elements may have an important function in the recognition of a significant number of human noncanonical poly(A) sites.

Discussion

As the 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. To address this issue, we have analysed how 3′end processing is controlled in the single exon MC4R pre-mRNA. Unlike in the so far analysed viral and human intronless genes (Huang and Carmichael, 1997; Conrad and Steitz, 2005; Guang and Mertz, 2005; Dalziel *et al*, 2007), we found that MC4R poly(A) site recognition is not dependent on additional auxiliary sequences located either upstream or downstream of the core poly(A) site (Figure 1); 3′end processing of the MC4R pre-mRNA is solely dependent on the core poly(A) sequences. Mutational analysis further showed that poly(A) cleavage at the MC4R processing site does not require a

canonical hexamer and that its short DSE is the most critical *cis*-element. Our data not only question our current understanding of what constitutes a functional poly(A) site but may also help 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 (Beaudoing *et al*, 2000; Zarudnaya *et al*, 2003; Hu *et al*, 2005; Tian *et al*, 2005). The co-transcriptional recognition of these canonical poly(A) sites has been extensively investigated. Countless 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 (Zhao *et al*, 1999).

In contrast to these observations, mutating the equivalent sequences in the MC4R poly(A) site had the opposite effects. Although mutations in the AUUAAA hexamer sequence did not significantly impair poly(A) cleavage, mutations of two or more nucleotides in the DSE had a greater impact on 3′end processing efficiency (Figure 2). Interestingly, the 19 nucleotide long MC4R DSE (CGTGTTATTCATAAGCATT) is a good match of the consensus YGUGUUYY motif (Zarudnaya *et al*, 2003) and is very similar to sequences that show strong CstF-64 subunit-binding affinities (Perez Canadillas and Varani, 2003). The presence of this 'optimal' DSE may explain why the MC4R poly(A) site tolerates inactivation of the hexamer sequence (Figure 2).

Yeast and plant 3′end processing signals have long been considered to be different to human poly(A) sites. In particular, as they appear to generally have all their *cis*-elements located upstream of the cleavage site and instead of A(A/U)UAAA hexamers, often have A-rich positioning elements (Supplementary Figure 5). Mammalian poly(A) sites that contain the critical *cis*-elements upstream of the cleavage site have been described (Moreira *et al*, 1995; Brackenridge and Proudfoot, 2000; Natalizio *et al*, 2002) but functional processing sites that constitute an A-rich sequence and are independent of auxiliary elements have so far been more elusive. Our results presented in this study 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 (Guo and Sherman, 1996; Hunt, 2008) than previously thought.

The observation made in the analysis of MC4R 3′end processing may also be 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 (Tian *et al*, 2005). However, the recognition of these noncanonical poly(A) sites is to date only poorly understood. It is currently believed that auxiliary sequences located either upstream or downstream of noncanonical poly(A) sites may be able to compensate for a degenerated hexamer sequence. Such sequences may serve to stabilise poly(A) complex assembly by providing alternative binding opportunities for components of the 3′end processing machinery (Supplementary Figure 5). A recent analysis of the noncanonical poly(A) site located in the pre-mRNA of the poly(A) polymerase γ gene (PAPOLG) has demonstrated that UGUAN sequence motifs in the 3′UTR 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 critical for 3′end processing (Venkataraman *et al*, 2005). Interestingly, our analysis of more than 10 000 human poly(A) sites showed that non-canonical A-rich 3′end processing sites can be found in a significant number of human genes (Figure 5; Supplementary Figure 2). Importantly, when we analysed an example (JUNB) that has been identified in our screen, we found evidence that the A-rich upstream sequence is indeed a critical element in 3′end processing of such pre-mRNAs.

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 4). 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 with canonical poly(A) sites (Figure 5). As a higher content of U and/or GU nucleotides in the DSE is likely to stabilise 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 the transcription apparatus to pause. This pausing may then allow CstF to establish contacts with both the DSE and CPSF. The formation of this early complex then 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 (Rigo *et al*, 2005; Nag *et al*, 2007). The DSE-mediated recognition of noncanonical poly(A) sites may still be compatible with this dynamic model. It is plausible that a strong DSE allows CstF to rapidly associate with the pre-mRNA in the absence of pausing and then 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 enable the assembly of a functional 3′end processing complex. Alternatively, it could be that A-stretches themselves have an intrinsic ability to pause elongating polII and so facilitate assembly of the 3′ processing machinery.

Our analysis suggests that at least in some mammalian poly(A) sites the functional importance of processing site recognition may be shifted from the core upstream element to the downstream element. This supports the growing under-

standing that the CstF–DSE interaction represents a critical regulatory step for poly(A) site choice (Takagaki *et al*, 1996; Takagaki and Manley, 1998; Veraldi *et al*, 2001; Phillips *et al*, 2004; Shell *et al*, 2005). 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 (Shankarling *et al*, 2009) and in testis (Wallace *et al*, 1999; McMahon *et al*, 2006; Dass *et al*, 2007). Coincidently, pre-mRNAs of meiotic and post-meiotic male germ cells also appear to have a higher incidence of poly(A) sites with noncanonical hexamers (Liu *et al*, 2007) and our bioinformatics analysis indicates that noncanonical A-rich poly(A) sites are more likely to be subjected to tissue-specific alternative 3′ end processing (Supplementary Figure 3).

Finally, 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 (Kleiman and Manley, 2001; Kim *et al*, 2006; Mirkin *et al*, 2008).

Materials and methods

Plasmids

The plasmids used are based on the melanocortin 1 receptor (MC1R) reporter plasmid described earlier (Dalziel *et al*, 2007). The original MC4R reporter was constructed by retaining the CMV promoter, the MC1R 5′UTR and the ORF of the GFP of the MC1R reporter, but replacing the 3′UTR and flanking regions with 1336 base pairs (bp) of MC4R genomic sequence including the 3′UTR and 3′flanking regions.

The MC4R 3′UTR and flanking regions of variable length were amplified from HeLa cells genomic DNA and inserted into the above plasmid using *Xba*I-*Sph*I, *Xba*I-*Hind*III or *Hind*III-*Hind*III restriction sites depending on the specific constructs. Mutations affecting the core hexamers, DSE, UGUAN sequences and sequences upstream of the hexamer were created by PCR amplification with specific primers.

The 3′UTR and 3′flanking sequences of the *JUNB* gene were amplified by PCR, fused to MC4R sequences downstream of P1 replacing the MC4R 3′UTR and P1 poly(A) sequences. Mutation of the A-rich sequences and the DSE were introduced by PCR.

Sequences of all primers used are available on request.

RNA analysis

RNA isolations and RP analysis were performed as described earlier (Dalziel *et al*, 2007). Quantitations were carried out using a Fuji phosphor imager.

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 an oligo-dT (oligo-dT17) reverse primer. cDNAs were amplified by PCR and analysed on agarose gels.

Bioinformatics analysis

Poly(A) sites. The poly(A) sites in PolyA_DB2 (Lee *et al*, 2007) were used for this study. To minimise the issue of internal priming (Lee *et al*, 2008b), 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 described earlier in Tian *et al* (2005). The resulting numbers of human poly(A) sites studied were 2361, 1722, 2347 and 4204 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 analysed as described earlier (Lee *et al*, 2008a).

Cis-element analysis. A-rich elements were hexamers with at least five 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 (Hu *et al*, 2005), 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.

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 Wang *et al* (2008) and Pan *et al* (2008). The combined set includes 13 human tissue samples, that are 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 five mammary epithelial or breast cancer cell line samples, that are 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 Supplementary Figure 3. 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 (Lee *et al*, 2007). 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.

Supplementary data

Supplementary data are available at *The EMBO Journal* Online (http://www.embojournal.org).

Acknowledgements

We thank all the members of the Furger, Murphy and Proudfoot laboratories for valuable discussions. We thank Simon Haenni, Carol Lutz and Nick Proudfoot for the critical reading of the paper. We are grateful to the Wellcome Trust (081083/Z/06/Z) and the Portuguese FCT for funding AF and NMN, respectively. BT and WL are funded by the NIH, United States (R01 GM084089).

Conflict of interest

The authors declare that they have no conflict of interest.

References

Anand S, Batista FD, Tkach T, Efremov DG, Burrone OR (1997) Multiple transcripts of the murine immunoglobulin epsilon membrane locus are generated by alternative splicing and differential usage of two polyadenylation sites. *Mol Immunol* **34:** 175–183

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D (2000) Patterns of variant polyadenylation signal usage in human genes. *Genome Res* **10:** 1001–1010

Brackenridge S, Proudfoot NJ (2000) Recruitment of a basal polyadenylation factor by the upstream sequence element of the human lamin B2 polyadenylation signal. *Mol Cell Biol* **20:** 2660–2669

Conrad NK, Steitz JA (2005) A Kaposi's sarcoma virus RNA element that increases the nuclear abundance of intronless transcripts. *EMBO J* **24:** 1831–1841

Conway L, Wickens M (1987) 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:** 4177–4184

Dalziel M, Nunes NM, Furger A (2007) Two G-rich regulatory elements located adjacent to and 440 nucleotides downstream of the core poly(A) site of the intronless melanocortin receptor 1 gene are critical for efficient 3′ end processing. *Mol Cell Biol* **27:** 1568–1580

Danckwardt S, Hentze MW, Kulozik AE (2008) 3′ end mRNA processing: molecular mechanisms and implications for health and disease. *EMBO J* **27:** 482–498

Dass B, Tardif S, Park JY, Tian B, Weitlauf HM, Hess RA, Carnes K, Griswold MD, Small CL, Macdonald CC (2007) Loss of polyadenylation protein tauCstF-64 causes spermatogenic defects and male infertility. *Proc Natl Acad Sci USA* **104:** 20374–20379

Guang S, Mertz JE (2005) 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:** 2215–2226

Guo Z, Sherman F (1996) 3′-end-forming signals of yeast mRNA. *Trends Biochem Sci* **21:** 477–481

Hu J, Lutz CS, Wilusz J, Tian B (2005) Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA* **11:** 1485–1493

Huang Y, Carmichael GG (1997) The mouse histone H2a gene contains a small element that facilitates cytoplasmic accumulation of intronless gene transcripts and of unspliced HIV-1-related mRNAs. *Proc Natl Acad Sci USA* **94:** 10104–10109

Hunt AG (2008) Messenger RNA 3′ end formation in plants. *Curr Top Microbiol Immunol* **326:** 151–177

Huszar D, Lynch CA, Fairchild-Huntress V, Dunmore JH, Fang Q, Berkemeier LR, Gu W, Kesterson RA, Boston BA, Cone RD, Smith FJ, Campfield LA, Burn P, Lee F (1997) Targeted disruption of the melanocortin-4 receptor results in obesity in mice. *Cell* **88:** 131–141

Kim HS, Li H, Cevher M, Parmelee A, Fonseca D, Kleiman FE, Lee SB (2006) DNA damage-induced BARD1 phosphorylation is critical for the inhibition of messenger RNA processing by BRCA1/BARD1 complex. *Cancer Res* **66:** 4561–4565

Kleiman FE, Manley JL (2001) The BARD1-CstF-50 interaction links mRNA 3′ end formation to DNA damage and tumor suppression. *Cell* **104:** 743–753

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

Lee JY, Ji Z, Tian B (2008a) Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3′-end of genes. *Nucleic Acids Res* **36:** 5581–5590

Lee JY, Park JY, Tian B (2008b) Identification of mRNA polyadenylation sites in genomes using cDNA sequences, expressed sequence tags, and Trace. *Methods Mol Biol* **419:** 23–37

Lee JY, Yeh I, Park JY, Tian B (2007) PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res* **35** (database issue)**:** D165–D168

Liu D, Brockman JM, Dass B, Hutchins LN, Singh P, McCarrey JR, MacDonald CC, Graber JH (2007) Systematic variation in mRNA 3′-processing signals during mouse spermatogenesis. *Nucleic Acids Res* **35:** 234–246

Liu H, Kishi T, Roseberry AG, Cai X, Lee CE, Montez JM, Friedman JM, Elmquist JK (2003) Transgenic mice expressing green fluorescent protein under the control of the melanocortin-4 receptor promoter. *J Neurosci* **23:** 7143–7154

McMahon KW, Hirsch BA, MacDonald CC (2006) Differences in polyadenylation site choice between somatic and male germ cells. *BMC Mol Biol* **7:** 35

Millevoi S, Geraghty F, Idowu B, Tam JL, Antoniou M, Vagner S (2002) A novel function for the U2AF 65 splicing factor in promoting pre-mRNA 3′-end processing. *EMBO Rep* **3:** 869–874

Millevoi S, Loulergue C, Dettwiler S, Karaa SZ, Keller W, Antoniou M, Vagner S (2006) An interaction between U2AF 65 and CF I(m) links the splicing and 3′ end processing machineries. *EMBO J* **25:** 4854–4864

Mirkin N, Fonseca D, Mohammed S, Cevher MA, Manley JL, Kleiman FE (2008) The 3′ processing factor CstF functions in the DNA repair response. *Nucleic Acids Res* **36:** 1792–1804

Moreira A, Wollerton M, Monks J, Proudfoot NJ (1995) Upstream sequence elements enhance poly(A) site efficiency of the C2 complement gene and are phylogenetically conserved. *EMBO J* **14:** 3809–3819

Nag A, Narsinh K, Martinson HG (2007) The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. *Nat Struct Mol Biol* **14:** 662–669

Natalizio BJ, Muniz LC, Arhin GK, Wilusz J, Lutz CS (2002) Upstream elements present in the 3′-untranslated region of collagen genes influence the processing efficiency of overlapping polyadenylation signals. *J Biol Chem* **277:** 42733–42740

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40:** 1413–1415

Perez Canadillas JM, Varani G (2003) Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. *EMBO J* **22:** 2821–2830

Phillips C, Pachikara N, Gunderson SI (2004) U1A inhibits cleavage at the immunoglobulin M heavy-chain secretory poly(A) site by binding between the two downstream GU-rich regions. *Mol Cell Biol* **24:** 6162–6171

Proudfoot N (2004) New perspectives on connecting messenger RNA 3′ end formation to transcription. *Curr Opin Cell Biol* **16:** 272–278

Proudfoot NJ, Furger A, Dye MJ (2002) Integrating mRNA processing with transcription. *Cell* **108:** 501–512

Rigo F, Kazerouninia A, Nag A, Martinson HG (2005) The RNA tether from the poly(A) signal to the polymerase mediates coupling of transcription to cleavage and polyadenylation. *Mol Cell* **20:** 733–745

Shankarling GS, Coates PW, Dass B, Macdonald CC (2009) A family of splice variants of CstF-64 expressed in vertebrate nervous systems. *BMC Mol Biol* **10:** 22

Sheets MD, Ogg SC, Wickens MP (1990) Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation *in vitro*. *Nucleic Acids Res* **18:** 5799–5805

Shell SA, Hesse C, Morris Jr SM, Milcarek C (2005) Elevated levels of the 64-kDa cleavage stimulatory factor (CstF-64) in lipopolysaccharide-stimulated macrophages influence gene expression and induce alternative poly(A) site selection. *J Biol Chem* **280:** 39950–39961

Shi Y, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, Yates Iii JR, Frank J, Manley JL (2009) Molecular architecture of the human Pre-mRNA 3′ processing complex. *Mol Cell* **33:** 365

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

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

Takagaki Y, Seipelt RL, Peterson ML, Manley JL (1996) The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell* **87:** 941–952

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

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

Venkataraman K, Brown KM, Gilmartin GM (2005) Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. *Genes Dev* **19:** 1315–1327

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

Wallace AM, Dass B, Ravnik SE, Tonk V, Jenkins NA, Gilbert DJ, Copeland NG, MacDonald CC (1999) 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 USA* **96:** 6763–6768

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476

Wickens M, Stephenson P (1984) Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3′ end formation. *Science* **226:** 1045–1051

Wilusz J, Pettine SM, Shenk T (1989) Functional analysis of point mutations in the AAUAAA motif of the SV40 late polyadenylation signal. *Nucleic Acids Res* **17:** 3899–3908

Wilusz J, Shenk T (1988) A 64 kd nuclear protein binds to RNA segments that include the AAUAAA polyadenylation motif. *Cell* **52:** 221–228

Zarudnaya MI, Kolomiets IM, Potyahaylo AL, Hovorun DM (2003) Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. *Nucleic Acids Res* **31:** 1375–1386

Zhao J, Hyman L, Moore C (1999) Formation of mRNA 3′ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* **63:** 405–445

