PolyA signals located near 5’ of genes are silenced by a general mechanism that prevents premature 3’ end processing

Jiannan Guo

Supervised by Dr Saverio Brogna

A thesis submitted to
The University of Birmingham
For the degree of
DOCTOR OF PHILOSOPHY

School of Biosciences
College of Life and Environmental Sciences
The University of Birmingham
September 2010

Abstract

PolyA signals located at the 3’ end of eukaryotic genes drive the cleavage and polyadenylation reaction to the nascent pre-mRNA. Although these sequences are expected only at the 3’ end of genes, we found that strong polyA signals are also present within the 5’ untranslated regions (UTRs) of many *Drosophila melanogaster* mRNAs. Although the polyA signals in 5’ UTRs show little activity of triggering 3’ end processing in the endogenous transcripts, they are very active when placed at the 3’ end of reporter genes. We further investigated these unexpected observations and discovered that both these novel polyA signals and standard polyA signals become functionally silent when they are positioned close to transcription start sites in either Drosophila or human cells. This suggests that the transcriptional stage when the polyA signal emerges from the polymerase II (Pol II) transcription complex could determine whether a putative polyA signal is recognized as functional. The data suggest that this mechanism, which probably prevents cryptic polyA signals from causing premature transcription termination, depends on low Ser2 phosphorylation of the C-terminal domain of Pol II and inefficient recruitment of processing factors.

Acknowledgements

I am most grateful to my supervisor Dr Saverio Brogna for his immense effort and patience in guiding me through my study. I also thank Prof Jeff Cole and Dr Brian Ford-Lloyd, and all others involved in awarding me the Dorothy Hodgkin Postgraduate Award.

I owe my deep gratitude to Dr Jikai Wen and Ms Preethi Ramanathan, who helped me greatly in virtually all aspects of my lab work. I thank all past and present members of the Brogna lab for their encouragements. I also thank Dr Alicia Hidalgo, Dr Matthias Soller, Dr Stephen Dove and their lab members for valuable discussions and sharing equipments. I am grateful for scientific discussions and suggestions from Prof Nick Proudfoot, Prof Steve Buratowski and Dr Domenico Libri.

All above made my four years in Birmingham an invaluable memory. Nonetheless, I would like to thank my family, especially my parents, for their unconditional support and love.

Table of Contents

Chapter 1 Introduction

The hallmark for eukaryotic gene expression is that the precursor mRNA transcript (pre-mRNA) undergoes a series of complex processing reactions in the nucleus before being exported to the cytoplasm and translated. Although the single reactions can be biochemically separated *in vitro*, in the cell pre-mRNA processing events are interlinked with one another and with transcription, but they are also linked with downstream events such as translation and mRNA degradation in the cytoplasm (Moore and Proudfoot, 2009; Perales and Bentley, 2009).

The pre-mRNA is synthesised by RNA polymerase II (Pol II). The first processing event to occur on newly transcribed pre-mRNA is capping of the 5’ end (Shuman, 2001). Once the nascent transcript emerges from Pol II, a RNA 5’ triphosphatase (RT) converts the triphosphate of the first nucleotide to a diphosphate. Then, a guanylyl transferase (GT) fuses GMP to the terminal phosphates to form an unusual 5' to 5' triphosphate linkage. Finally, a methyl transferase (MT) methylates the N7 of the transferred guanine, forming the cap structure often abbreviated to m^7G (Shatkin and Manley, 2000; Shuman, 2001). The immediate function of the cap is to protect the mRNA from being targeted by 5’-3’ exonucleases (Beelman and Parker, 1995; Shuman, 2001). In the nucleus, the cap structure is bound by the cap binding complex (CBC), which consists of two proteins, CBP20 and CBP80. The CBC is required for mRNA export and possibly for the pioneer round of translation (Ishigaki et al., 2001). In the cytoplasm, the cap is essential for efficient translation

initiation as it is recognised by the eukaryotic translation initiation factor 4E (eIF4E) (Rhoads, 2009; Sonenberg, 2008).

The pre-mRNA of eukaryotes contains sequences (introns) that are not present in the mature mRNA. Introns are removed by pre-mRNA splicing. In brief, accurate splicing chiefly relies on recognition 5' splicing site AG|*GU*RAGU and 3' splicing site Y*AG*|RNNN at the intron-exon borders. The reaction is catalysed by the spliceosome, a large assembly made of U1, U2, U4, U5, and U6 snRNPs, plus several associated splicing factors (Jurica and Moore, 2003). In addition to removing introns, splicing factors interact with 3' end processing factors (Proudfoot et al., 2002). For example, U2 snRNP is shown to interact with 3' end processing factor CPSF (Kyburz et al., 2006). In particular relevance to this thesis, U1 snRNP, which binds to 5' splicing sites, has been shown to inhibit pre-mature 3' end processing in human cells (Kaida et al., 2010). The link between splicing and 3' end processing is further discussed below.

With the exception of metazoan replication-dependent histone genes (Dominski and Marzluff, 2007), all protein-encoding mRNAs undergo 3' end processing, which generates a polyA tail at the 3' end of each transcript. The polyA tail can be of different lengths (60 nt to 250 nt) depending on the organism and the specific mRNA. The functions of 3' end processing include promoting transcription termination, conferring mRNA stability, facilitating mRNA export and translation (Colgan and Manley, 1997; Richard and Manley, 2009). The polyA tail is added by a cleavage/polyadenylation reaction catalysed by a multi-subunit protein complex –

the polyA complex – which recognises the polyA signal on the nascent transcript and triggers the reaction (Shi et al., 2009). Details of 3’ end processing will be discussed in this chapter.

1.1 Pre-mRNA 3’ end processing

1.1.1 Sequence requirements

The sequences on the transcript that dictate the site of cleavage/polyadenylation are referred to as the polyA signal. The polyA signal is present at the 3’ end of the gene. In human systems, where the polyA signal is most conserved, the most apparent sequence motif is the hexamer AAUAAA that lies 10-30 nt upstream of the cleavage site (Colgan and Manley, 1997), but more than ten sub-optimal variants have also been identified (e.g. AUUAAA) (Tian et al., 2005). Around 20-40 nt downstream of the cleavage site lies a less conserved U-rich or GU-rich motif termed the downstream sequence element (DSE). The flexibility in the sequence composition of DSE is suggested to compensate the use of sub-optimal hexamers (Nunes et al., 2010; Zarudnaya et al., 2003). Between the hexamer and the DSE, an endonucleolytic cleavage occurs, preferably but not necessarily after a CA dinucleotide (Chen et al., 1995). The efficiency of 3’ end processing is regulated by additional auxiliary sequences (Hu et al., 2005; Legendre and Gautheret, 2003). For example, an upstream U-rich sequence element (USE) and a UGUA element are found to have stimulatory role in the reaction (Gilmartin et al., 1995; Moreira et al., 1998; Yang et al., 2010). It has been suggested that the structural context of the

RNA upstream of the AAUAAA hexamer is also important for recognition by the polyA complex (Graveley et al., 1996). Moreover, a recent study reported that some human polyA signals require only a potent DSE and an A-rich upstream sequence (Nunes et al., 2010). Contrary to the earlier understanding, all these studies suggest a large diversity in polyA signals. In yeast, the sequence requirement for a polyA signal is even less stringent (see below). Therefore, the question of how cells can accurately and reliably recognise polyA signals in a genome context remains to be addressed.

1.1.2 Trans-acting factor requirement

The large multi-protein polyA complex that recognises the polyA signal consists of a core of 14 protein subunits in human cells (Colgan and Manley, 1997; Proudfoot, 2004; Wahle, 1995). The total assembly of the polyA complex might contain up to ~85 proteins (Fig 1.1.2.1) (Mandel et al., 2008; Shi et al., 2009). Identification and characterisation of the core components of the polyA complex was chiefly achieved via an *in vitro* cleavage assay (Colgan and Manley, 1997). In the polyA complex, the cleavage and polyadenylation specificity factor (CPSF) directly recognises the AAUAAA via its 160 kDa subunit CPSF-160 (Murthy and Manley, 1995). Another two subunits, CPSF-100 and CPSF-73, are also required for the reaction (Gilmartin et al., 1995; Murthy and Manley, 1992). CPSF-30 appears to be redundant *in vitro* (Bienroth et al., 1991; Murthy and Manley, 1992). A fifth component of CPSF, hFip1p, facilitates in linking CPSF and the polyA polymerase (PAP) (Preker et al.,

1995). CPSF-73 is the endonuclease that carries out the cleavage reaction (Mandel et al., 2006). The binding of CPSF to the polyA signal is strengthened in the presence of the cleavage stimulatory factor (CstF), which contains CstF-77, CstF-64 and CstF-50. CstF-64 binds to U/GU-rich sequences of the DSE (MacDonald et al., 1994; Takagaki and Manley, 1997), while CstF-77 has been shown to interact with CstF-64, CstF-50 and CPSF-160, bridging the two protein complexes (Murthy and Manley, 1995; Takagaki and Manley, 1994). CstF-50 is required for cleavage and binds the C-terminal domain (CTD) of the largest subunit of Pol II (Fong and Bentley, 2001; Takagaki and Manley, 1994). Cooperative interactions between CPSF and CstF specifies the cleavage site (Murthy and Manley, 1995). Recently, it has been shown that CstF-77 undergoes dimerization mediated by its own HAT-C domain (Bai et al., 2007). Dimerization of CstF-77 lead to the proposal that the entire CstF may dimerize in the polyA complex, resulting in two copies of CstF-64 available to bind the DSE simultaneously (Bai et al., 2007). In addition to CPSF and CstF, cleavage factor I (CF I) and cleavage factor II (CF II) are also essential for the cleavage of the pre-mRNA (Takagaki et al., 1989). CF I (which consists of CF I-68, CF I-59, and CF I-25) can recognise UGUAN sequence elements upstream of the hexamer and promote assembly of the polyA complex, even in the absence of A(A/U)UAAA (Venkataraman et al., 2005). Recently structural study has shown that CF I-25 directly interact with UGUAAA or UUGUAU (Yang et al., 2010). CF II consists of Clp1 and Pcf11. Clp1 interacts with both CPSF and CF I. Pcf11 interacts with the Pol II CTD and plays a critical role in transcription termination

(Licatalosi et al., 2002; West and Proudfoot, 2008; Zhang et al., 2005). PAP adds the polyA tail, and is generally required for the cleavage reaction (Gilmartin and Nevins, 1989; Takagaki et al., 1988, 1989). PolyA binding protein II (PABP II) binds to the newly synthesised polyA tail and enhances the processivity of PAP (Bienroth et al., 1993; Wahle et al., 1993).

Remarkably, for such a general and important cellular process, the specificity of the 3' end processing chiefly relies on the polyA complex recognising loosely defined sequence elements. Even the most conserved AAUAAA/AUUAAA hexamer in human system could be replaced by A-stretch and still trigger the reaction adequately in a reporter system (Nunes et al., 2010). This lack of specificity not only provides opportunity for higher frequency of alternative 3' end processing, but probably also calls for specificity control mechanisms.

In human system, where splicing is common for most protein coding genes, the removal of last intron and the definition of terminal exon helps to restrict 3' end processing to the last exon (Proudfoot et al., 2002). For example, splicing factors and polyA factors together form a terminal exon definition complex (EDC), which 'licenses' the polyA signal in the terminal exon (Rigo and Martinson, 2008; Rigo and Martinson, 2009). However, for intron-less genes, splicing factors are not involved in this pathway (Rigo and Martinson, 2009). Other specificity mechanisms may be involved in this regulation, insights of which may be obtained by investigating the dynamic modifications on the elongating transcription complex.

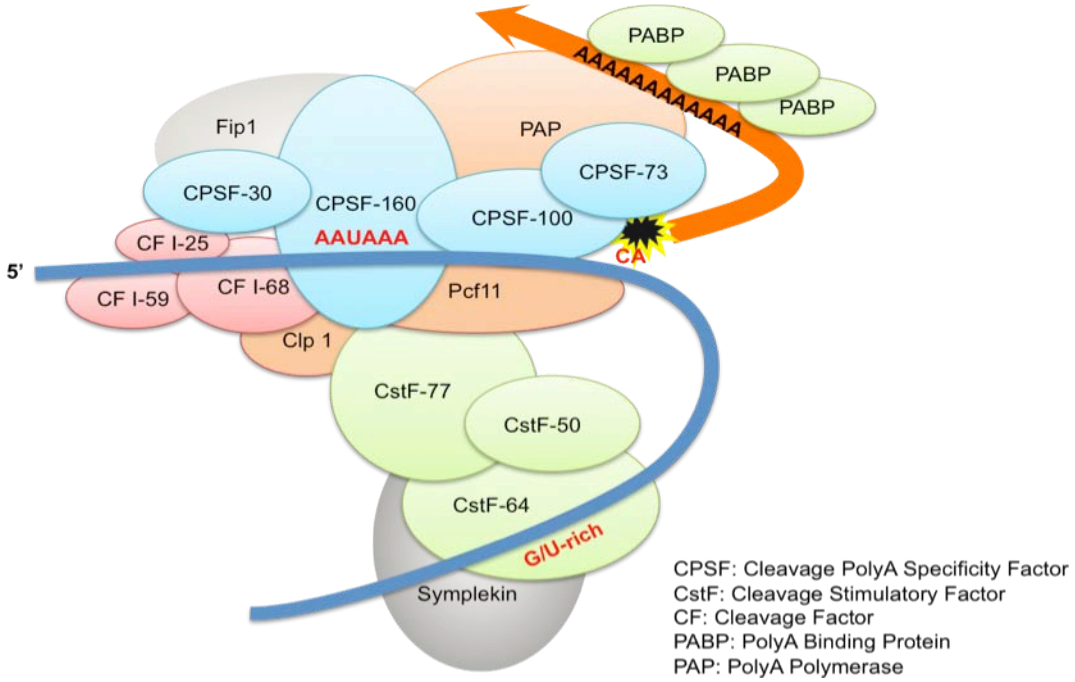


Fig 1.1.2.1 Schematic of the human core polyA complex. The blue curved line represents the RNA substrate, and the 5' end is indicated. G/U-rich represents the DSE. The orange curved line represents the polyA tail added at the cleavage site, indicated by CA. The protein factors involved are indicated as ovals. The Pol II CTD is not shown. This illustration is modified from (Dominski, 2007).

1.1.3 Similarities and differences between metazoan and yeast

The core proteins required for the cleavage/polyadenylation reaction are conserved from yeast to humans even though the polyA signal sequences differ (Mandel et al., 2008; Millevoi and Vagner, 2009; Proudfoot, 2004). The yeast polyA complex consists of the cleavage and polyadenylation factor (CPF), the cleavage factor IA (CFIA) and the cleavage factor IB (CFIB). CPF includes homologues to all human CPSF subunits and many additional factors including PAP (Pap1p in *S. cerevisiae*). The CFIA consists of four subunits: Rna14p and Rna15p are equivalent to human CstF-77 and CstF-64, whilst Pcf11p and Clp1p have counterparts in the mammalian CFII. CFIB has only one member, Hrp1p, and does not have a homolog in mammals. Despite the similar protein composition of the polyA complexes, the polyA signal sequences are distinct between yeast and humans. Most significantly, there is no consensus sequence such as the AAUAAA hexamer in yeast. Instead the required sequence elements are the relatively less conserved AU-rich efficiency element (EE), the A-rich proximal element (PE), the upstream U-rich element (UUE) and the downstream U-rich element (DUE) (Fig 1.1.3.1) (Millevoi and Vagner, 2009; Proudfoot, 2004). Between UUE and DUE is the cleavage site, defined by a pyrimidine followed by multiple adenosines: Y(A)n. Protein-sequence interactions are also different: Thh1p, the counterpart of the mammalian CPSF-160, binds to the cleavage site instead of the A-rich PE, which may be considered the counterpart of the mammalian AAUAAA (Dichtl et al., 2002). Meanwhile, Rna15p, the

counterpart to CsrF-64, recognises the A-rich PE in the upstream region instead of DUE, which may be seen as the equivalent of the mammalian DSE (Gross and Moore, 2001). A concise illustration of human and yeast polyA signals is shown in Fig 1.1.3.1; for detailed comparison between organisms see reviews (Millevoi and Vagner, 2009; Proudfoot, 2004). In summary, yeast polyA signals are less conserved than mammalian polyA signals.

The exact composition of Drosophila polyA complex has not been systematically assessed. However, most human polyA factors have close homologs in Drosophila, implying close similarity between human and Drosophila system (Wahle, 1995). Sequence elements for Drosophila polyA signal is generally considered similar as in mammals, although preliminary data in this lab suggest that a AATAAA to AAGAAA mutation, contrary to established mammalian data, is not sufficient to completely abolish the functionality of the polyA signal. Results in this thesis also imply Drosophila polyA signals might not be as conserved as human polyA signals (Results chapters).

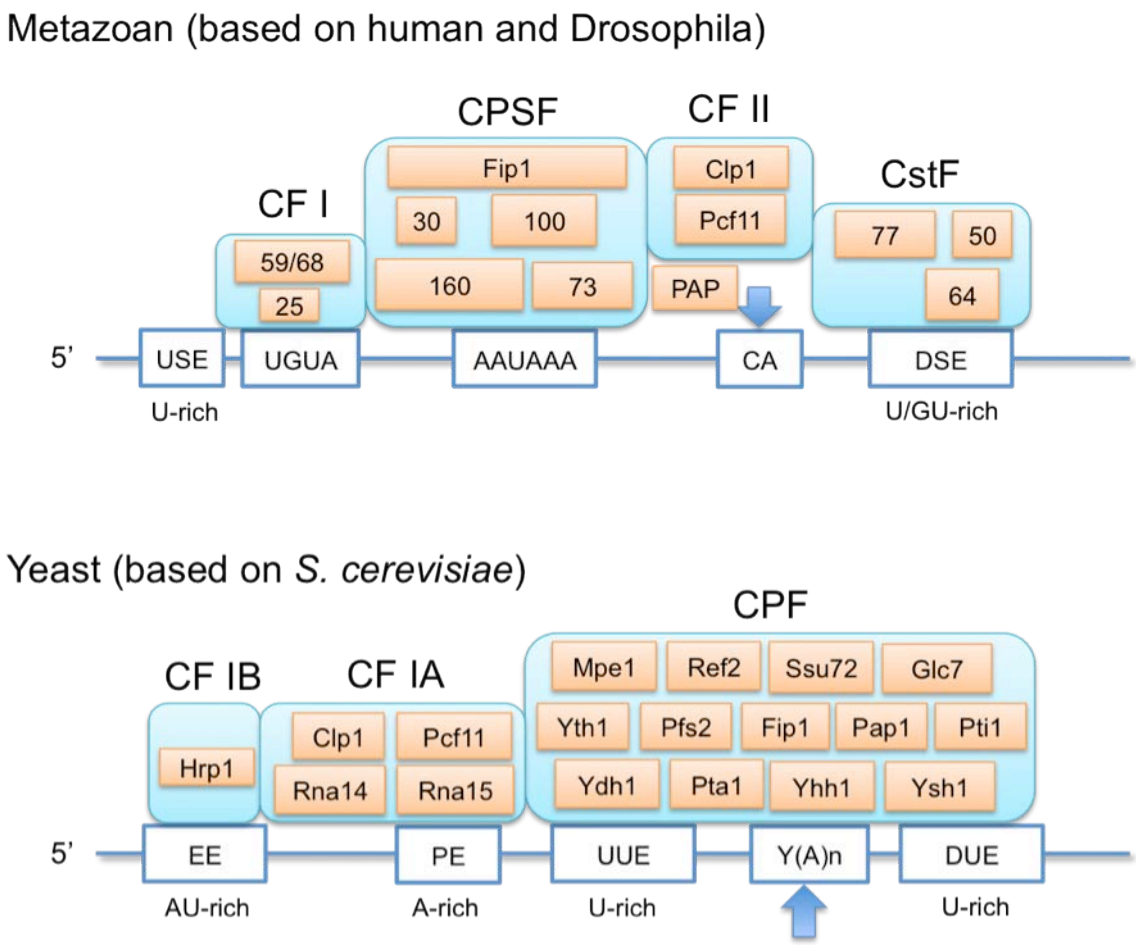


Fig 1.1.3.1 Comparison of metazoan and yeast polyA signals. Schematics show sequence elements and associated polyA factors. This simplified illustration compares the differences of sequence elements required for efficient 3' end processing in different organisms. For details of the elements and factors see text. This illustration is based on two published reviews (Millevoi and Vagner, 2009; Proudfoot, 2004).

1.1.4 Integration of 3’ end processing and transcription

The 3’ end processing reaction, like all other pre-mRNA processes, are interwoven with the transcription complex. The structure of Pol II plays a critical role in this integration. The holoenzyme Pol II is a 550 kDa complex of 12 protein subunits in yeast (Cramer et al., 2000; Cramer et al., 2001; Gnatt et al., 2001; Myer and Young, 1998). The largest subunit, Rbp1p, consists of a central globular structure with the central active site that opens up the DNA template and catalyses the polymerase reaction. The central core has an entry channel for the nucleotides and an exit channel for the transcript. Under the RNA exit channel lays the CTD, which is a relatively unstructured domain that appears separated from the main body of Pol II (Cramer et al., 2001; Gnatt et al., 2001). It contains multiple heptad amino acid repeats that fall in the consensus of YSPTSPS. There are 52 heptad repeats in mammals, 42 in Drosophila, and 26 in *S. cerevisiae* (Moore and Proudfoot, 2009; Zhang and Gilmour, 2006). The CTD serine residues are subjected to reversible phosphorylation during the transcription cycle and directly interact with components of pre-mRNA processing machineries (Fong and Bentley, 2001). The phosphorylation of specific serine residues correlates to specific transcriptional stages. Serine 5 phosphorylation (Ser5P) is abundant during the early stage of transcription elongation (up to the first few hundreds of nucleotides) then declines further downstream (Komarnitsky et al., 2000). The phosphorylated Serine 2 (Ser2P) is closely associated with the later productive stage of transcription elongation,

during which splicing and cleavage/polyadenylation occur. Ser5 and Ser2 are targeted by specific kinases and phosphotases. Cdk7, or Kin28 in yeast, a subunit of transcription factor II H (TFIIH), phosphorylates Ser5 residues at initiation, whereas the Cdk9 subunit of P-TEFb (positive transcription elongation factor), phosphorylates Ser2 residues later during elongation (Peterlin and Price, 2006); Ctk1 catalyses the same reaction in yeast. The phosphatase for Ser5P is the Pol II binding protein Rtr1 (Mosley et al., 2009), whereas the phosphatase for Ser2P is Fcp1 (Cho et al., 2001; Ghosh et al., 2008). The transition from high Ser5P to high Ser2P coincides with Pol II switching from abortive elongation to the productive elongation stage (Buratowski, 2009; Ni et al., 2008; Phatnani and Greenleaf, 2006). Recently, it was reported that Cdk7 of TFIIH could also phosphorylate Ser7 residues (Akhtar et al., 2009; Glover-Cutter et al., 2009; Kim et al., 2009). Ser7P level appears high near promoters, similar to Ser5P, implying a possible association with early transcriptional events. A brief illustration of the CTD phosphorylation state during the transcription cycle is shown in Fig 1.1.4.1.

The polyA signal is recognised cotranscriptionally by components of the cleavage and polyadenylation complex. Although the cleavage/polyadenylation reaction typically takes place at the end of a transcription cycle, some key factors are recruited to the transcription complex at earlier stages. For example, immunoprecipitation experiments demonstrated that CPSF is brought to the preinitiation complex via TFIID – a transcription initiation factor. After transcription starts, CPSF dissociates from TFIID and becomes associated with the elongating

polymerase (Dantonel et al., 1997). In addition, Chromatin immunoprecipitation (ChIP) experiments demonstrated that CPSF-73 and CstF-77 are enriched at transcription start sites, in addition to that at 3' ends (Glover-Cutter et al., 2007). CFIm also appears be recruited at early stage of transcription (Venkataraman et al., 2005). Many other polyA factors are loaded on the Pol II CTD prior to 3' end processing in both humans and yeast (Ahn et al., 2004; Hirose and Manley, 1998; Licatalosi et al., 2002; McCracken et al., 1997; Zhang et al., 2005). CTD Ser2P is required for the recruitment of many polyA factors (Ahn et al., 2004; Licatalosi et al., 2002). Pcf11, in particular, directly interacts with the CTD (Meinhart and Cramer, 2004; Zhang et al., 2005; Zhang and Gilmour, 2006). The level of Ser2P dramatically drops to a basal level downstream of the polyA signal, the same region where polyA factors dissociate from the transcription complex (Ahn et al., 2004; Cui et al., 2008; Garrido-Lecca and Blumenthal, 2010; Kim et al., 2004a; Ni et al., 2004; Zhang and Gilmour, 2006).

Recognition of the polyA signal and 3'end processing is not only essential for polyadenylation, but is also a key determinant for Pol II termination (Richard and Manley, 2009; Rosonina et al., 2006). It was demonstrated that Pol II occupancy is reduced a few hundred nucleotides downstream of the polyA site (Connelly and Manley, 1988; Logan et al., 1987; Whitelaw and Proudfoot, 1986). In particular, it was shown that the cleavage reaction, independent of polyA tail addition, is required for transcription termination (Birse et al., 1998). Furthermore, the strength of the polyA signal directly correlates to termination efficiency (Osheim et al., 1999).

Despite intensive research in this field, the mechanism of Pol II termination is still not fully characterised (Ghazal et al., 2009; Kim et al., 2004b; Rondon et al., 2009; West et al., 2004; West et al., 2008). The consistent observation that transcription of the polyA signal is required for efficient termination is interpreted as Pol II somehow becoming competent for termination only after transcribing the polyA signal (Connelly and Manley, 1988; West et al., 2008). A clear change of the Pol II elongation complex upon transcription of the polyA signal is that CTD Ser2P drops dramatically to the basal level. The polyA factor Pcf11 appears to be a key player in linking transcription of the polyA signal and termination: Pcf11 binds the CTD and, at least *in vitro*, causes dissociation of both Pol II and the nascent transcript from the DNA (Zhang et al., 2005; Zhang and Gilmour, 2006). In addition, interaction between RNA and Pol II at early stages of polyA complex assembly has been suggested to prime the polyA factors to bind to Pol II (Rigo et al., 2005). Together, the concerted input from RNA, Pol II and polyA factors promote termination.

As mentioned before in section 1.1.2, the splicing machinery and 3' end processing machinery together defines the terminal exon (Rigo and Martinson, 2008; Rigo and Martinson, 2009). This coupling clearly increases the chance for exonic polyA signals at the 3' end to be processed. For genes without introns, the composition of transcription elongation complex might play a crucial role in preventing pre-mature 3' end processing. In this thesis, we chiefly discuss the situation of intron-less genes in Drosophila cells.

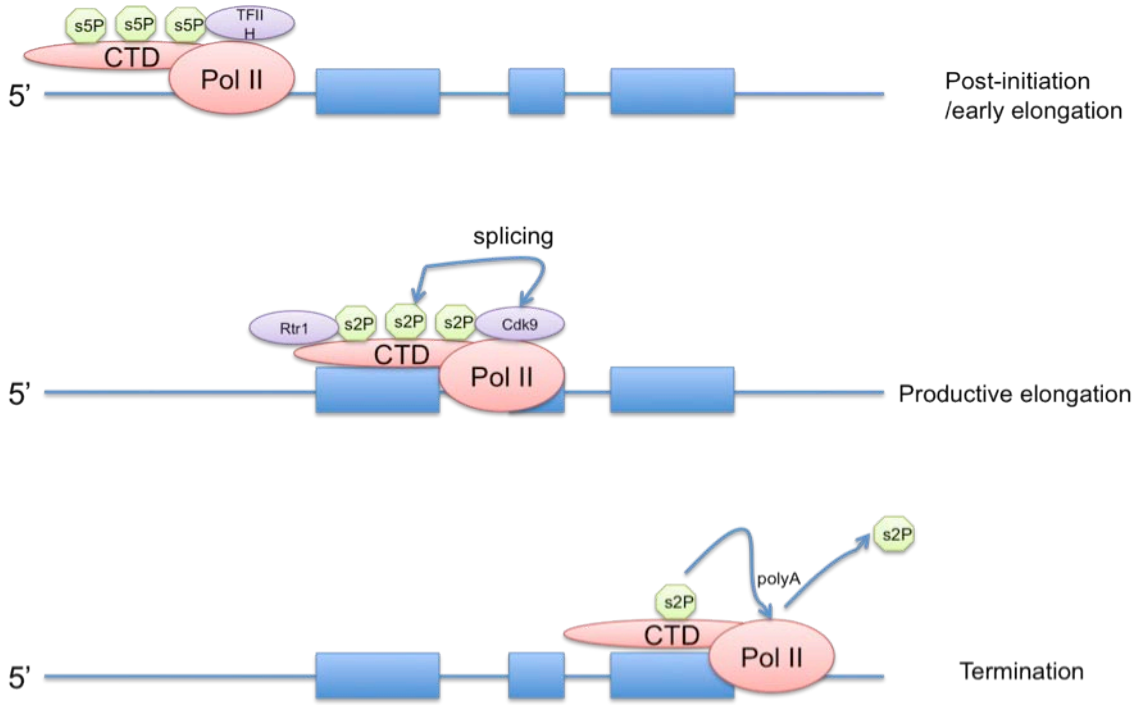


Fig 1.1.4.1. Schematic of the changes in phosphorylation status of Pol II CTD during transcription initiation, elongation and termination. At initiation, TFIIH phosphorylates Ser5, signalling recruitment of the capping enzymes. During the productive elongation phase, Rtr1 dephosphorylates Ser5P while the Cdk9 subunit of P-TEFb phosphorylates Ser2. The Ser2P level gradually increases and peaks at the polyA site. Prior to Pol II transcription termination, the Ser2P level drops to near basal level.

1.2 Alternative polyadenylation and polyA site selection mechanisms

1.2.1 Alternative polyA signals

As discussed above, the sequence requirements for polyA signals are relatively loose. It is therefore not a surprise to see the studies described below which describe the identification of large number of polyA signals in various parts of the transcribed regions. Putative polyA signals occur frequently in the A-U rich regions such as introns (Tian et al., 2007), UTRs (Lee et al., 2008), and intergenic regions (Lopez et al., 2006). Moreover, variants of the hexamer AAUAAA can also be functional. Bioinformatical studies have shown that the AAUAAA hexamer only accounts for 53% and 59% of total polyadenylation events in human and mouse respectively. The second most used hexamer AUUAAA accounts for ~16% in both human and mouse (Tian et al., 2005). Other detected hexamers include UAUAAA, AGUAAA, AAGAAA, AAUAUA, AAUACA, CAUAAA, GAUAAA, AAUGAA, UUUAAA, ACUAAA and AAUAGA (Beaudoing et al., 2000; Tian et al., 2005). Notably, ~54% of human genes have alternative polyA signals and there are on average 2.1 used polyA signals for every gene (Tian et al., 2005). Intronic polyadenylation events are found in ~20% of human genes (Tian et al., 2007). Transcription occurs also in intergenic regions: 3500 human genes are predicted with intergenic polyA signals (Lopez et al., 2006), while extra long 3' UTRs up to 10 kb have been experimentally detected in mammalian cells (Moucadel et al., 2007). A small

number of polyadenylation events in 5’ UTRs have also been listed, although the efficiency and function of those events are unknown (Tian et al., 2007). Furthermore, the number of transcripts with alternative polyA sites might have been underestimated because all the above bioinformatic analysis relied on EST databases, which probably represent only 10% of transcripts, with many tissues and developmental stages libraries lacking (Gilat et al., 2006). Indeed, recent deep sequencing data revealed that human alternative splicing and alternative polyadenylation are twofold to threefold more frequent when comparing between tissues than comparing between individuals (Wang et al., 2008a). In addition, the sequence requirement for a polyA signal is less conserved than previously understood. A recent study has reported that a functional polyA signal requires only an A-rich upstream sequence and a DSE, suggesting that noncanonical polyA signals are more common then previously envisaged (Nunes et al., 2010).

1.2.2 Alternative polyadenylation influences gene expression

Alternative polyadenylation influences many aspects of gene expression (Edwalds-Gilbert et al., 1997). Tandem polyA signals in a 3’ terminal exon will produce transcripts with the same coding region but different lengths of 3’ UTRs, which may lead to different stabilities, translation efficiencies and localisations. It has been shown that effective transcription termination, which is promoted by efficient 3’ end processing, enhances gene expression, and the effect is particularly dramatic when weak or noncanonical polyA signals are present (West and Proudfoot,

2009). Usage of early polyA signal could facilitate early release of Pol II from the template for the next round of transcription; therefore increasing the overall efficiency of gene expression.

The eukaryotic initiation factor 2α (eIF-2α), a key factor for protein synthesis, has two common mRNAs of 1.6 and 4.2 kb, which are produced by alternative usage of two polyA signals in the terminal exon. The 1.6 kb transcript is less stable but is more readily translated *in vitro* (Mao et al., 1992; Miyamoto et al., 1996). The ratio between the two mRNA species differs between tissues and stages in cell cycles. For example, expression of the shorter mRNA is enhanced upon T cell activation, ultimately producing more protein per primary transcript. The shorter 1.6 kb transcript also appear to be more efficiently processed and exported than the 4.2 kb mRNA (Edwalds-Gilbert et al., 1997).

When alternative polyA signals are combined with alternative splicing, different coding regions can be produced, resulting in different protein products (Tian et al., 2005; Tian et al., 2007). Immunoglobulin heavy chain genes (α, δ, ε, γ and μ) are also alternatively polyadenylated to produce two mRNAs. The use of the proximal polyA site produces mRNA encoding a secreted form of the antibody, whereas the use of the distal polyA site generates mRNA for the membrane-bound antigen receptor (Goodnow et al., 1988; Rogers et al., 1986). During activation of B cells, there is a switch from producing the membrane-bound form to the secreted form. It has been shown that high level of CstF-64 in plasma cells enhances the recognition of the relatively weak proximal polyA site and therefore produces the secreted form

of the protein (Takagaki et al., 1996).

On a global scale, recent genomic studies have also shown general gene expression regulation by alternative polyA signals. For example, a global analysis suggests that proliferating cells tend to prefer proximal polyA signals to maximise gene expression efficiency, since experimentally forced use of distal polyA signal reduces gene expression (Sandberg et al., 2008). In agreement with this observation, cancer cells preferentially use proximal polyA signals to avoid long 3' UTRs that are targeted by miRNAs. Short 3' UTRs resulted in increased expression levels and promoting oncogenic transformation (Mayr and Bartel, 2009). The mechanism for selecting alternative polyA signals in the same 3' UTR remains unclear. The CTD might play an important role in the selection, given its involvement in the processing and its changing conformation during the transcription cycle. An insightful example from a study in yeast is that deletion of Ctk1p, the CTD Ser2 kinase, resulted in readthrough of an otherwise active proximal polyA signal and activation of a normally un-transcribed distal polyA signal (Ahn et al., 2004).

1.3 Polyadenylation/oligoadenylation and mRNA quality control

Transcription is coupled with RNA quality control mechanisms. A central player in this process is the exosome, a multi-subunit protein complex containing several exoribonucleolytic proteins and RNA binding proteins. It is required for the degradation of aberrant pre-mRNAs and the processing of certain types of cellular RNAs (such as ribosomal RNAs and small nuclear/small nucleolar RNAs)

(Houseley et al., 2006; Schmid and Jensen, 2008; Vanacova and Stef, 2007). The degradation of RNA with an abnormal 3' end appears to be coupled to transcription.

In *S. cerevisiae*, mRNA export factor mutants produced transcripts that are hyper-polyadenylated (Jensen et al., 2001), while polyA polymerase mutant produced transcripts that fail to acquire polyA tails (Hilleren et al., 2001). Both the hyper-polyadenylated and the hypo-polyadenylated transcripts appear sequestered at the site of transcription; this retention requires the nuclear exosome as a mutation in its nuclear subunit Rrp6p leads to the accumulation of defectively polyadenylated transcripts in the cytoplasm and their translation (Hilleren et al., 2001; Libri et al., 2002). Further evidence indicates that the exosome is present at the site of transcription. In particular, exosome subunits have been visualised at sites of active transcription on the polytene chromosomes in Drosophila, together with the Pol II elongation factor Spt6 (Andrulis et al., 2002). Moreover, yeast mutants of Rrp6p and cleavage and polyadenylation factors (Rna14p and Rna15p) showed inefficient transcription elongation, demonstrating the integrated effects of the nuclear exosome and the polyA factors on mRNA biogenesis (Luna et al., 2005). The function of the exosome in transcription is unclear. It is plausible that the exosome is associated with elongating transcription complex as a surveillance mechanism, while lack of the exosome might lead to accumulation of aberrant mRNA as well as defective transcription elongation.

Degradation can be triggered not only by the detection of defective (hyper or hypo) polyadenylation, but also depending on which polyA polymerase has added

the polyA tail. Typically, the canonical PAP adds a polyA tail to stabilise the mRNA. However, in yeast, polyadenylation by the non-canonical polyA polymerase Trf4p (or Trf5p) in the TRAMP complex (Trf4p/Air2p/Mtr4p or Trf5p/Air2p/Mtr4p polyadenylation complex) leads to recruitment of the nuclear exosome, which rapidly degrades the newly synthesised transcript (LaCava et al., 2005; Vanacova et al., 2005; Wyers et al., 2005). The length of the A-tail in this class appears shorter (no longer than 8 As, David Tollervey, unpublished data) than the tail on stable mRNA, although the exact in vivo mechanism differentiating the two is unknown. Notably, the A-tail addition by TRAMP follows transcription termination induced by Nrd1/Nab3/Sen1 complex instead of the canonical cleavage/polyadenylation complex (Steinmetz et al., 2001). This termination relies on Nrd1p binding to GUAA/G motifs and Nab3p binding to UCUU motifs (Carroll et al., 2004). The Nrd1/Nab3/Sen1 complex interacts with Pol II CTD with high level of Ser5P and therefore triggers termination at 5' end of genes (Gudipati et al., 2008; Vasiljeva et al., 2008). This is in contrast to the cleavage/polyadenylation-mediated termination, which requires high level CTD Ser2P at 3' end of genes. The fission yeast ortholog of Trf4p, called Cid14p, has polyA polymerase activity and appears to be involved in degradation of transcripts generated from naturally silenced heterochromatic domains (Buhler et al., 2007; Wang et al., 2008b).

In Drosophila, however, dTrf4-1 and dTrf4-2 polyadenylate snRNA as in yeast but their involvement in mRNA stability is unknown (Nakamura et al., 2008). Moreover, a TRAMP-like system to destroy aberrant RNAs has not yet been

reported in mammalian cells. Whether there is a system in higher eukaryotes that generates and/or degrades short transcripts generated at 5' end or genes remains unknown. Results in this thesis suggest that pre-mature cleavage and polyadenylation might be simply prevented by the state of early elongation complex. Adding to that splicing activities can also help restricting usage of polyA signals to the terminal exon as discussed in previous sections.

1.4 Work that led to this thesis

PolyA signals are typically found in the 3' UTRs of genes. However, in a previous study in this lab, a functional polyA signal was unexpectedly discovered in the sequence of 5' UTR of the *Ultrabithorax* (*Ubx*) mRNA in *D. melanogaster* (Ramanathan et al., 2008). When the 5' UTR sequence of the *Ubx* transcript was inserted in the intergenic spacer of an *Adh-Luc* dicistronic reporter, an *Adh* mRNA was produced, suggesting that the sequence must contain a functional polyA signal (Ramanathan et al., 2008). It was noticed that in this sequence there are seven AATAAA hexamers (shown in Fig 1.4.1). This unexpected finding indicates that a 5' UTR sequence can trigger 3' end processing, at least when positioned at 3' end of a reporter gene.

As discussed above, the sequence composition of a polyA signal is not as strict as initially thought. The unexpected finding of a polyA signal from a 5' UTR sequence further increased our curiosity. We then hypothesised 5' UTR polyA signal might be common, given the A/U-rich nature of UTR sequences. More polyA

signal regulatory mechanism might be uncovered through this approach. Following this notion, we predicted large number of potential polyA signals in 5' UTRs of Drosophila by a bioinformatic approach. Experiments based on reporter genes suggest that the distance between the polyA signal and transcription start site is a factor to determine whether the polyA signal is active. Further study on possible protein factors implied that the phosphorylation status of Pol II CTD and lack of cleavage/polyadenylation factors at early transcription elongation stage might contribute to this mechanism of silencing promoter proximal polyA signals.

A promoter proximal polyA signal is present in 5' long terminal repeat (LTR) in the HIV-1 genome, but this polyA signal is not used in HIV transcription. In brief, silencing of the HIV 5' LTR polyA signal dependents on two factors: viral transcription induced by Tat and an immediately adjacent major splicing donor site (Ashe et al., 1995; Weichs an der Glon et al., 1993). Similarities and differences of results in this thesis and published work on HIV 5' LTR polyA signal are discussed in the Discussion.

ATGAATGAACGAGAGGCGCCCACCCCGATAAACTTAAACTGAACGAACACTCAAGAGAGAGCGCAAGAGCGCTCAAAAACAATCTGGTT
TTGAGCGTTTCGCTGGCTCTCTGTTTCTGTTTTCCACTCGTTTTTAGGCCGAGTCGAGTGAGTTGAGTCGGCAGAGCAAAGTCAAAACA
CTGGCAACTGCGATTTGGTGCCACATTCGTTCGATGGCAACGGATTGGATAACAGGCGCGCGCTTTGTTTTATTATCCACATTATCAGC
GGCATTATTGTTATTATTGGCCCTCAGCGCTTTACCGCTCGCCCACGCGTCCGCCCGTGAATGCCGCGCGGAAAAGTCGCTTTCCACTA
GATTGGCGTCCAGATTCGAGGAAATCCGTCAGCAGACGCATTCGCGCCCGTTCGGTCAGCACTATGGCTAATAATCGTTCAAATCGTTA
AAACCATAAAAATAATAATAATAATTGCAATAAC**AATAAA**CATAGTAATAATCGTAACGCTTACGAGCCTTTGATAGTGCCAAGGCAAG
CGCAATCCAAGTATTCAAATTCGAATTCAATTAACAGCAAAGTGCAATTGGCTAAAAACCGAAACCAAAACGCAACAAAGTATACGAAA
CACTTGTGAAACCGTACAAACAATTGTGGAAAAAAATTTAAAGATTATTAAGATTGAAGTCTC**AATAAA**CATTAGTGCTTA**AATAAA**TT
TAAAACGACCCGCGTGGAGAGTGC**AATAAA**AAGAATAACTTTTGA**AATAAA**TATTTACCAAACAGAAAAATATTTTATAAATATTTAAA
TAAGTGAAAAACAAATTGGTTACTCTGAAACAAAGAAAATTCAAATTGGTGCTAAAACAAAGGAGAAAAAATTTCAAGAATATTTTACA
AATAATAAGACATATTTAACTATATAAAACCAAACTTAATCAACAAAGACTTGGAGTGAAAATATAA**AATAAA**TTTTAAAAAGAGTTTA
ACAAAATTTGTTTAAATCCAAAGGAGGCAAAGGAACAGCACAGAAAGCGAGGAAACACTCA**AATAAA**ATCCGCCAAAAATCGCAGATCC
CTGGAAACCAATTCGTGTGAAATCGGTCAAGCCCCCAACGACTTTTAGCCCGTCTCAGACGGAGCACCGCCAAGATTCTTACCGCCAGC
AGCGCA

Fig 1.4.1 Sequence of the *Ubx* 5' UTR. The seven AATAAA hexamers that could to be part of the functional polyA signal are shown in bold and underlined.

Chapter 2 Materials and Methods

Most protocols used in this thesis are as described in Molecular Cloning 2nd edition (Sambrook et al., 1989). Chemicals and reagents were purchased from Sigma-Aldrich, VWR or Fluka. Solutions where prepared with deionised H_2O (Elix 5 Water Purification System, Millipore), followed by sterilisation by autoclaving (121°C for 15 minutes) or filtration (0.22 μm, Millipore).

2.1 DNA cloning

2.1.1 PCR, DNA purification and DNA cloning

In PCR for cloning purposes, the Phusion high fidelity DNA polymerase (FINNZYMES) was used according to manufacture's instruction. DNA was purified by polyethylene glycol (PEG), High Pure PCR Template Preparation Kit (Roche) or Silica Bead DNA Gel Extraction Kit (Fermentas). PEG purification protocol is as follows:

1. Add equal volume of the PEG solution (13% PEG8000 (w/v), 0.6 M NaAc, and 6mM $MgCl_2 \cdot 6H_2O$) to the DNA sample and mix by vortexing. Incubate on bench for 10-20 minutes at room temperature.
2. Centrifuge at 13,200 rpm for 20 minutes and then remove the supernatant by pipetting.
3. Wash the DNA pellet in 1 ml of 96% ethanol and centrifuge at 13,200 rpm for 10 minutes. Remove the liquid by pipetting.

4. Air-dry the pellet and dissolve in H_2O or TE buffer for storage.

Ligations of DNA fragments used T4 DNA ligase (NEB) according to manufacture's instruction. In a ligation reaction, the molar ratio between the fragment and the target plasmid was typically 10:1.

2.1.2 Transformation to *E.coli* and plasmid preparations

Plasmids or ligation reactions were transformed into *E. coli*, chemi-competent cells, strains DH5α, TOP10 or XL1-Blue. The competent cells were purchased from Invitrogen or made by using the Rubidium Chloride method:

1. Grow 1 ml overnight culture of DH5α or XL1-blue cells in a 37 °C shaker.
2. Next morning, transfer the overnight culture into 100 ml LB with 10 mM $MgCl_2$ and 10 mM $MgSO_4$ and incubate in a 37 °C shaker until OD_{600} reaches 0.5-0.6.
3. Incubate cell on ice for 15 minutes.
4. Centrifuge cells at 5000 rpm for 10 minutes at 4 °C. Resuspend in 33 ml ice-cold Rb Buffer 1 (100 mM RbCl, 50 mM $MnCl_2{\cdot}4H_2O$, 80 mM KAc, 10 mM $CaCl_2{\cdot}2H_2O$, 15% glycerol, pH 5.8, filter sterilized) and incubate on ice for 1 hour.
5. Pellet the cells and remove the solution. Resuspend in 8 ml ice-cold Rb Buffer 2 (10 mM RbCl, 10 mM MOPS, 75 nM $CaCl_2{\cdot}2H_2O$, 15% glycerol, pH 6.8, filter sterilized). The cells are now ready to use for heat shock transformations.

6. Distributed the competent cells into microcentrifuge tubes on ice and then store at -80 °C.

In a heat shock transformation, ~5 ng plasmid or 5 µl ligation reaction was added to 50 µl ice-cold competent cells and incubated on ice for 15 minutes. The sample was then heat shocked at 42 °C for 30 seconds followed by immediate cooling on ice for 2 minutes. Then add 300 µl of NZY medium (10 g/L NZ amine, 5 g/L yeast extract and 5g/L NaCl, pH 7.5) and incubate at 37 °C for one hour. The sample was then plated onto LB agar plate with 100 µg/ml ampicillin for overnight growth.

To make a small-scale preparation of plasmid (mini-prep), a single colony was inoculated in 2 ml LB medium with 100 µg/ml ampicillin and incubated in a 37 °C shaker overnight. Plasmid preps are typically carried out by the boiling method:

1. Centrifuge 1.5 ml overnight culture in a microcentrifuge tube at 5,000 rpm for 5 minutes.
2. Resuspend the pellet from in 100 µl STET (8% sucrose, 5% Triton X-100, 50 mM EDTA and 50 mM Tris, pH 8.0) with 10 µl 10mg/ml lysozyme.
3. Incubate in boiling water bath for 30 seconds.
4. Centrifuge at 13,200 rpm for 10 minutes. Then discard the pellet by a toothpick.
5. Add equal volume of iso-propanol. Mix well and centrifuge at 13,200 rpm for 10 minutes.
6. Discard the liquid and wash the pellet with 1 ml 70% ethanol.

7. Centrifuge at 13,200 rpm for 10 minutes and discard the solution.
8. Air-dry the pellet and dissolve in H_2O with 0.1 mg/ml RNase A.
9. Incubate at room temperature for 20 minutes before storage or further applications.

For plasmids to be sequenced, minipreps were obtained using either GeneJET plasmid miniprep kit (FERMENTAS) or QIAprep spin miniprep kit (QIAGEN) following manufactures' instructions.

Plasmids for transfection were obtained using PureLink HiPure Plasmid Midiprep/Miniprep Kit (Invitrogen) following manufacture's instruction.

Plasmid DNA concentration was measured by NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies).

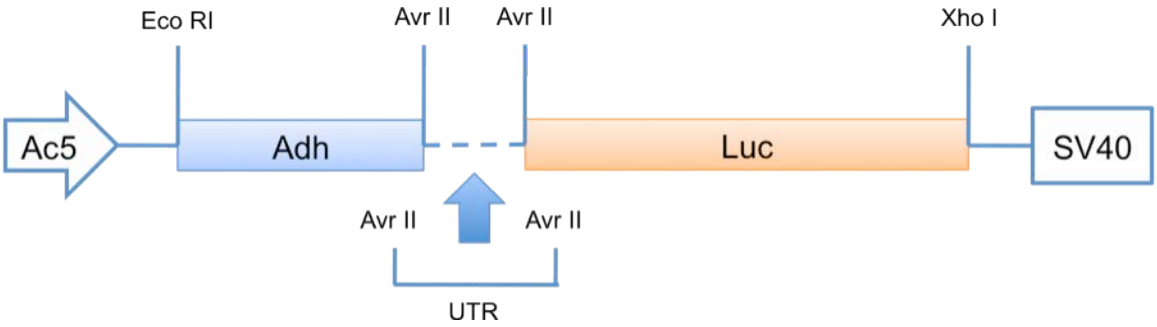
In normal PCR for verification purposes, such as colony-PCR, GoTaq (Promega) or BIOTAQ (Bioline) were used following manufactures' instructions. Restriction digestions were carried out using enzymes purchased from New England Biolabs (NEB) following manufacture's instruction.

2.2 Plasmid constructions

The plasmid constructs are derivatives of the dicistronic *Adh-Luc* reporter previously described (Ramanathan et al., 2008). The backbone of the plasmid is pAc5.1/V5-His A (Invitrogen), in which transcription is driven by the Actin 5C (Ac5) promoter and terminated by SV40 polyA signal (Fig 2.2.1). In brief, the sequence of *Adh* (either genomic or cDNA version) and *Luc* were inserted between the Eco RI and Xho I

sites. The coding regions of the two genes are separated by an Avr II site, one just before the stop codon of *Adh* and the other after the start codon of *Luc*. In the initial construct, the intergenic region is that of endogenous *Adh-Adhr* spacer (Brogna and Ashburner, 1997). This construct is used as positive control as the *Adh* polyA signal is known to be a strong polyA signal (Brogna and Ashburner, 1997).

Similar to the *Adh* polyA signal, the ten selected 5' UTR sequences that contain putative polyA signals (UTR-1 to UTR-10) and the five 5' UTRs that do not (Neg-1 to Neg-5) were PCR amplified from fly genomic DNA with flanking Avr II sites and inserted between the two genes (Fig 2.2.1). The stop codon of *Adh* and start codon of *Luc* are added in the primers to maintain complete open reading frames for both genes. The list of primers used to amplify the 5' UTR sequences is shown in the table in Fig 2.2.1.



UTR-1	CG1322_FW	GGG CCTAGG TAA GCG TTC GCT TTT TCT ACA
	CG1322_RV	GGG CCTAGG CAT GGT TGT TGC TTT ATT TTG GGG
UTR-2	CG7530_FW	GGG CCTAGG TAA GCG TTG CGA GAG GTG GAG
	CG7530_RV	GGG CCTAGG CAT GGT GCC ACT CGG TAG CCT GAT
UTR-3	CG6433_FW	GGG CCTAGG TAA CGT TAA TGC CAC CGA TCA
	CG6433_RV	GGG CCTAGG CAT GGT TTG TGG TCC AAT TTG CGG
UTR-4	CG5758_FW	GGG CCTAGG TAA AAT CGG TGC GGT TCA GTT
	CG5758_RV	GGG CCTAGG CAT GGT CGC TCA AAT CTG ATC GCA
UTR-5	CG6179_FW	GGG CCTAGG TAA ACA CCG TGT CCA TCT ACC
	CG6179_RV	GGG CCTAGG CAT GGT TTC CTG GAT TTG GCA GCG
UTR-6	CG9164_FW	GGG CCTAGG TAA ACC CAA CGA GTG CGA ACC
	CG9164_RV	GGG CCTAGG CAT GGT GCC GTC TTT GCA TTA CTG
UTR-7	CG17299_FW	GGG CCTAGG TAA TTA TTG CCG TAG CCG TTG
	CG17299_RV	GGG CCTAGG CAT GGT GCC GCC TTT GTC TTT GCT
UTR-8	CG17046_FW	GGG CCTAGG TAA TTG GGC AGA CTG GAG TGA
	CG17046_RV	GGG CCTAGG CAT GGT CAT GCG TCG AAT GGG AAT
UTR-9	CG7628_FW	GGG CCTAGG TAA CAA TGA AGT TTA AGC GCA
	CG7628_RV	GGG CCTAGG CAT GGT ATT TGG TTT TCG GTG TTC
UTR-10	CG17117_FW	GGG CCTAGG TAA ATT TTT GAC TGC GAA GCG
	CG17117_RV	GGG CCTAGG CAT GGT CTC TGG GAG CGA CGT CTA

Fig 2.2.1 The *Adh-Luc* dicistronic reporters. Schematic of the structure of the *Adh-Luc* based constructs with locations of key restriction sites indicated. Ac5 represents the Drosophila Actin 5C promoter; SV40 represents the polyA signal in the plasmid. The dotted line between *Adh* and *Luc* coding regions indicates the location where the 5' UTR sequences (UTR-1 to UTR-10) were inserted. The table below lists the primers used to PCR amplify the ten 5 UTRs sequences from genomic DNA.

To produce the construct derivatives with polyA signals at different distances from the 5' end, Adh-P1, Adh-P2 and Adh-P3, a Bgl II site was introduced immediately after codon 64, 126 or 203 of the *Adh* coding sequence (Fig 2.2.2). To introduce the Bgl II site the left and right halves of the *Adh-Luc* sequence were PCR amplified with one primer carrying a Bgl II site. PCR fragments of the two halves were digested by Bgl II and ligated to each other. The ligation product of *Adh-Luc* carrying the inserted Bgl II site was PCR amplified and cloned between the Eco RI and Xho I sites of pBlueScript II KS+ (Stratagene). The UTR-9 and SV40 polyA signals were PCR amplified with flanking Bgl II sites and inserted into P1, P2 and P3. The resulting fragments of *Adh-Luc* with insertions were cloned back into the original pAc plasmids (Fig 2.2.2).

The Adh-UTR-9-ΔP2, Adh-UTR-9-ΔP3, Adh-SV40-ΔP2 and Adh-SV40-ΔP3 constructs are derivative of the Adh-UTR-9-P2, Adh-UTR-9-P3, Adh-SV40-P2 and Adh-SV40-P3 described just above (Fig 2.2.3). Essentially, these constructs were generated by deleting the beginning of the *Adh* coding sequence, nt 1-192 (corresponding to codons 1 to 64). This deletion was generated by PCR using primers that omit nt 1-192 of *Adh* and cloning the fragments back into the backbone of the original constructs. To achieve this, I used a sense primer at nt 193 of *Adh* with flanking Eco RI site.

The shortened derivative of UTR-9, S-UTR-9, was PCR amplified with primers 5'-GGGAGATCTGCGCAAATATGGCTGTTTAGA-3' (S-UTR-9_FW) and 5'-GGGAGATCTAATACTGATTTCACTTCTTGC-3' (S-UTR-9_RV). To

generate the S-UTR-9-ΔTAAΔAATAAA, I used a PCR-ligation-PCR scheme similar to the procedure used to delete the TAA and the AATAAA.

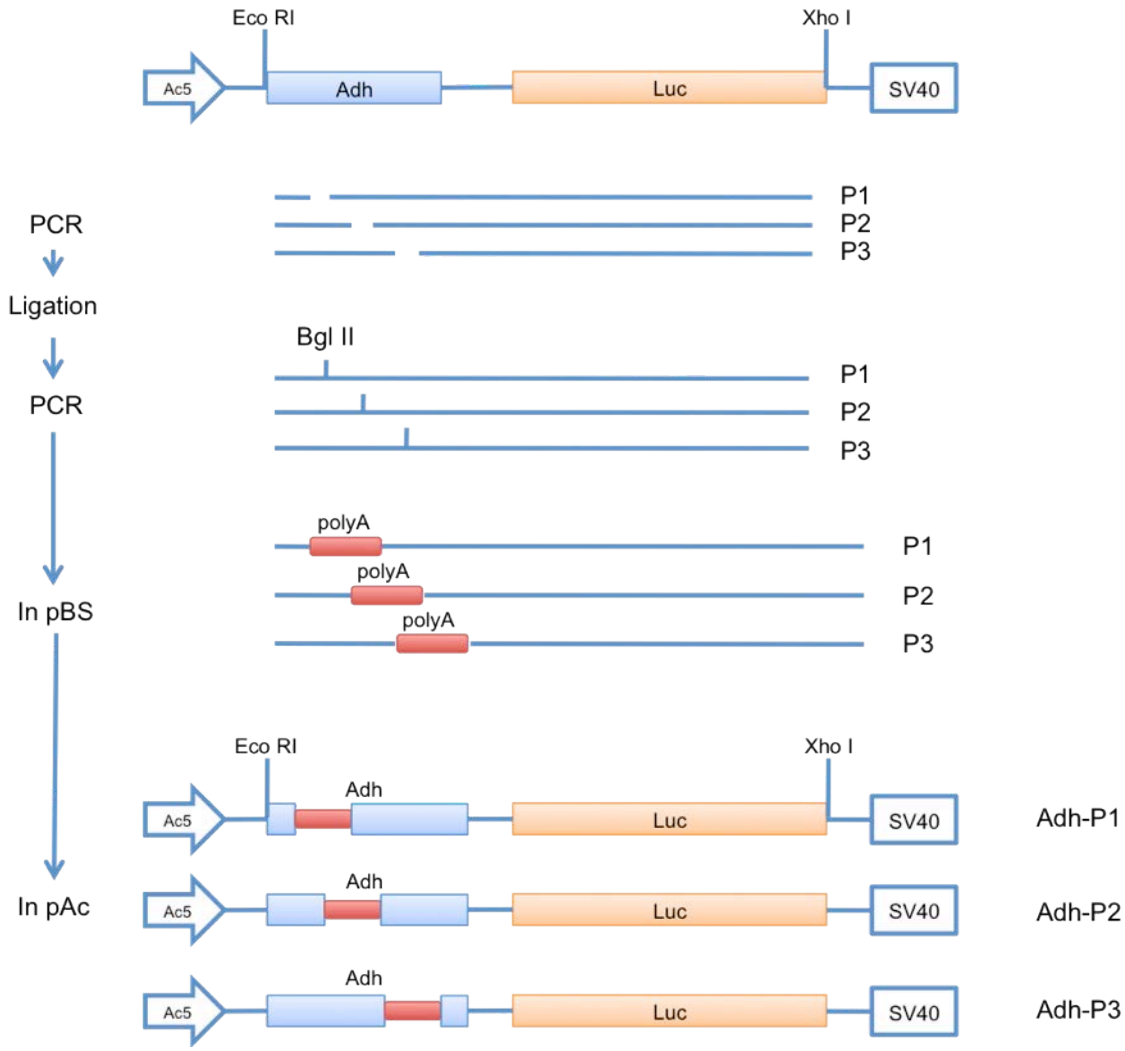


Fig 2.2.2 Cloning of reporters with polyA signal inserted at positions P1, P2 and P3 in the *Adh* gene. Original structure of *Adh-Luc* with polyA signal is shown on the top. On the left is shown the flow chart with the cloning strategy. Short vertical lines in the second PCR fragments indicate the Bgl II sites created. The inserted sequences are shown as thinner red boxes. Experimental procedures are described in the text.

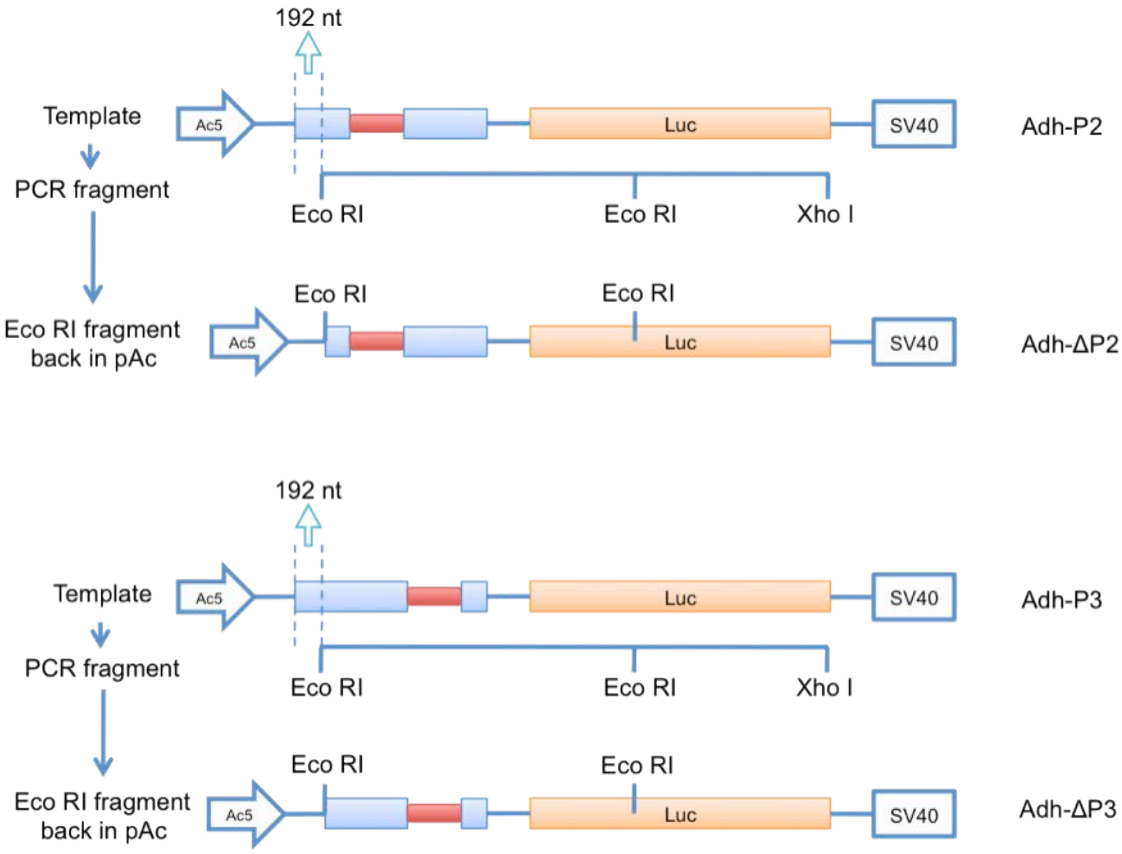


Fig 2.2.3 Cloning of the Adh-ΔP2 and Adh-ΔP3 reporters. Labelling scheme similar to Fig 2.2.2. These constructs were generated by deleting nt 1-192 of the *Adh* coding sequence. The deletion was generated by PCR using primers that omit nt 1-192 of *Adh* and cloning the fragments back into the backbone of the original constructs.

The *lacZ* based reporters were made as described by the schematic in Fig 2.2.4. The *lacZ* sequence was PCR amplified from plasmid pAcV5/His/LacZ (Invitrogen) with primers flanked with a Kpn I site at the beginning and an Xho I site at end. The fragment was cloned into pAc5.1 using the same two sites in the polylinker. The insertion point for LacZ-P1 is the Kpn I site upstream of the coding region. To create the LacZ-P2 and LacZ-P3, Avr II sites were introduced after codon 49 and 149 of *lacZ* coding sequence by similar cloning strategy as Fig 2.2.2. The sequence of bovine growth hormone gene's (BGH) polyA signal was inserted at P1, P2 and P3.

The *Luc* based of reporters were made with a similar strategy (Fig 2.2.5). The *Luc* sequence was PCR amplified from the *Adh-Luc* plasmid flanked with Kpn I site and Xho I site, followed by cloning into pAc5.1. In Luc-P1, the UTR-4 was inserted in the Kpn I site upstream of the *Luc* coding region. To create the Luc-P2 and Luc-P3, Avr II sites were introduced after codon 64 and 293 of *Luc* by similar cloning strategy as Fig 2.2.2. The sequence of UTR-4 was inserted at P1, P2 and P3.

The human expression plasmids were generated by cloning the reporter cassettes in the pAc plasmids, Adh-SV40-P1, Adh-SV40-P2 and Adh-SV40-P3, into the backbone of pcDNA 3.1+ (Invitrogen). Essentially, the Eco RI fragments from the pAc plasmids were moved into the Eco RI site in the pcDNA plasmid (Fig 2.2.6).

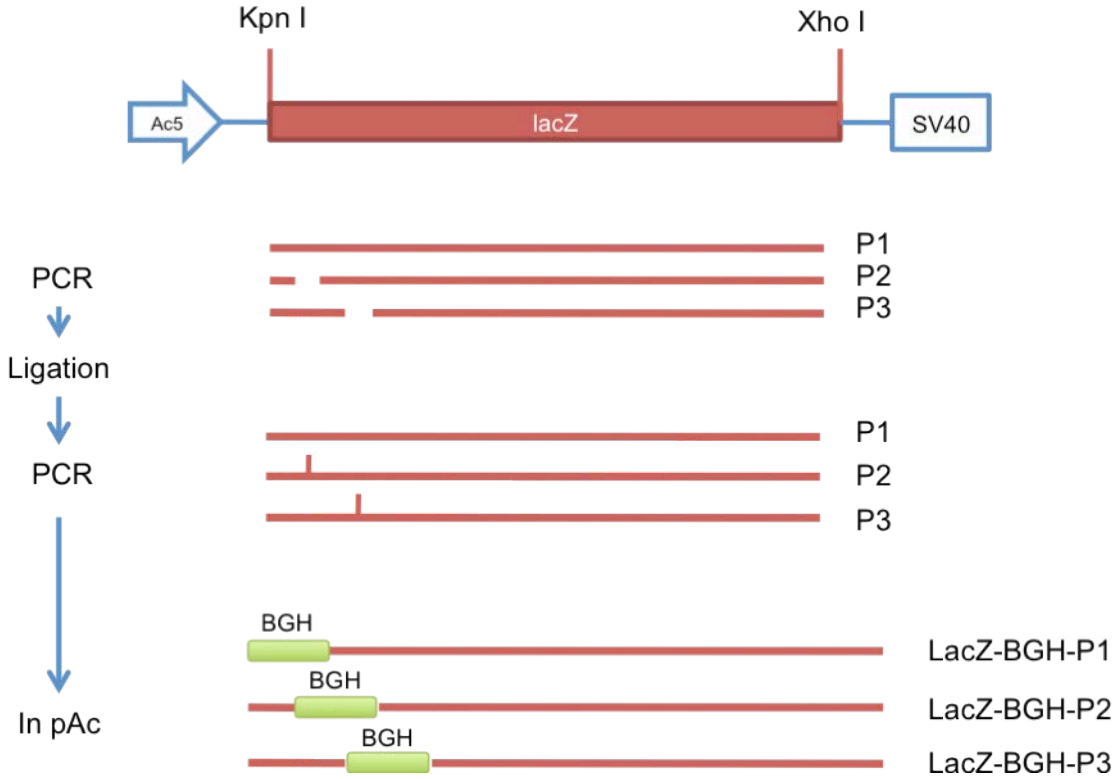


Fig 2.2.4 Cloning of the LacZ-BGH-P1, LacZ-BGH-P2 and LacZ-BGH-P3 reporters. Labelling scheme similar to Fig 2.2.2. LacZ-P1 uses the Kpn I site. LacZ-P2 (after codon 49) and LacZ-P3 (after codon 149) contain Avr II sites by PCR-ligation-PCR scheme similar as Fig 2.2.2. The fragments were then cloned back into pAc between the Kpn I and the Xho I sites. Then PCR fragments of the BGH polyA signal with Kpn I flanked (for P1) or Avr II flanked (for P2 and P3) was inserted into corresponding sites.

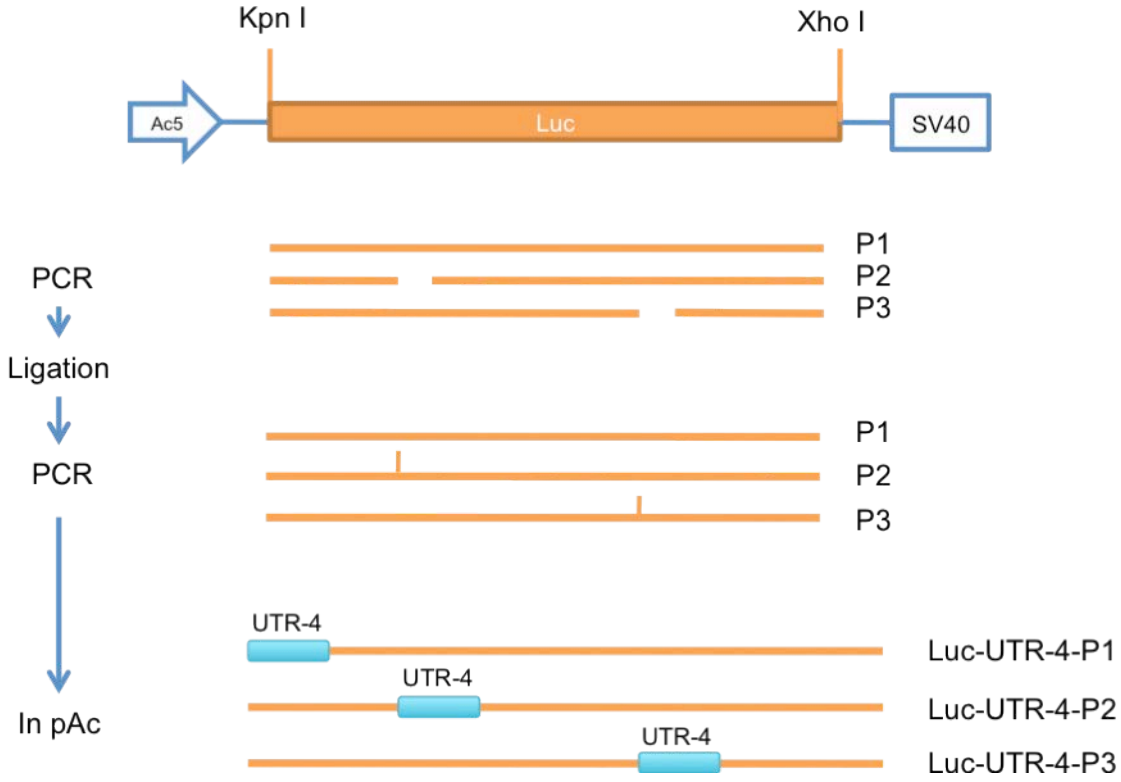


Fig 2.2.5 Cloning of the Luc-UTR-4-P1, Luc-UTR-4-P2 and Luc-UTR-4-P3 reporters. Similar cloning strategy and labelling scheme as in Fig 2.2.4, except that Luc-P2 is after codon 64 and Luc-P3 is after codon 203. UTR-4 with corresponding flanking sites was inserted at each point.

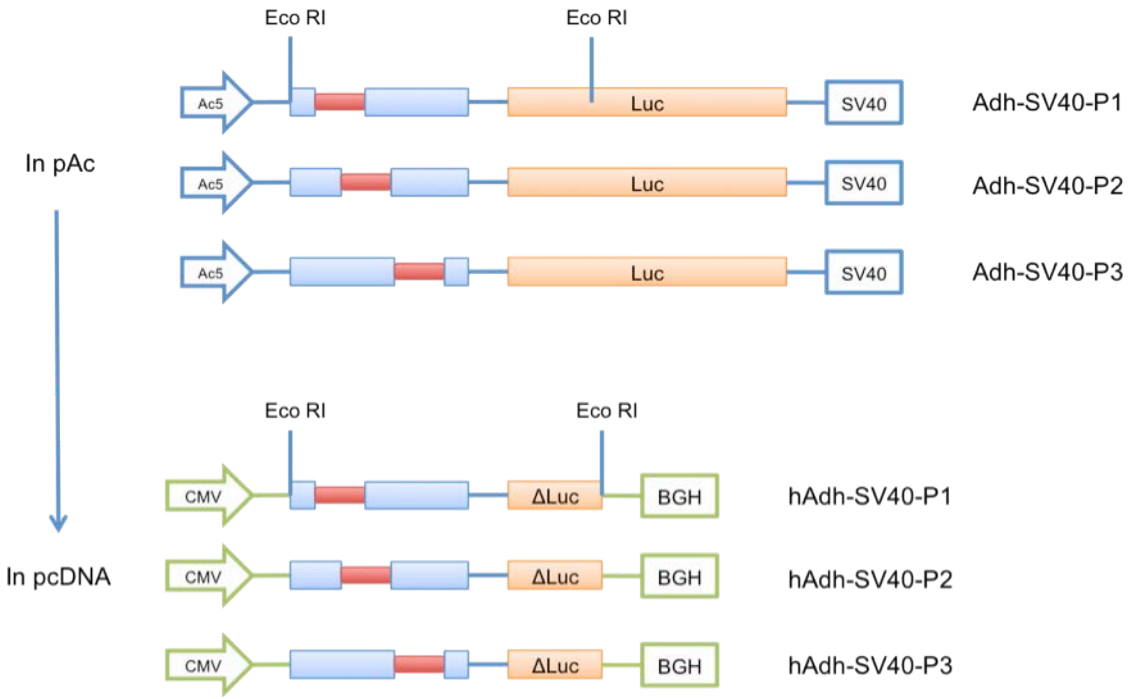


Fig 2.2.6 Cloning of the hAdh-SV40-P1, hAdh-SV40-P2 and hAdh-SV40-P3 reporters. From the pAc5.1-based reporters Adh-SV40-P1, Adh-SV40-P2 and Adh-SV40-P3, the Eco RI fragments were digested and cloned into pcDNA3.1.

2.3 Transfections of S2 and 293T cells

Drosophila Schneider 2 cells were grown at 27 °C with no CO_2 in Insect Xpress medium (Lonza) supplemented with 4% heat-inactivated fetal bovine serum (FBS, Lonza) and 1x penicillin/streptomycin/glutamine mix (P/S/G, Lonza). Heat inactivation of FBS was carried out for 30 minutes at 65 °C in a water bath. Parental S2 cells were typically grown in T25 cm^2 cell culture flasks (Falcon) and split every 3 – 4 days to maintain 20% - 100% of confluency.

Transient transfections of plasmids were carried out using a dimethyldioctadecylammonium bromide (DDAB, Sigma) mediated transfection using a protocol similar to that has been previously published (Han, 1996; Ramanathan et al., 2008). For preparation of DDAB solution, 4mg/ml of DDAB/H_2O was prepared and sonicated 17-20 rounds of 60-80 seconds intervals, with 2 – 3 minutes of cooling on ice between each round.

On the day before transfections, 3 x 10^6 cells/well were seeded into 6 well plates, so the next morning the cells would be attached to the plates and reached almost 100% confluency. For each well, the transfection mix contains 2μg of experimental plasmid, 1μg of EGFP expressing plasmid, 4μl of 4mg/ml DDAB solution, topped up to 125μl with empty IX medium (without FBS and P/S/G). The mix was well mixed and left standing at room temperature for 30 minutes. During the 30 minutes, the seeded cells were washed by empty IX medium twice followed by addition of 1ml of empty IX medium. Then the transfection mix was added into each well and

mixed by swirling and tilting gently. Cells with transfection mix were then incubated at 27 °C with no CO_2 for five hours before the empty medium with transfection mix was replaced with complete medium (with serum and P/S/G). Cells were further incubated 27 °C with no CO_2 for 24 to 48 hours before analysis.

The human HEK 293T cells were cultured at 37 °C with 5% CO_2 in DMEM medium supplemented with 10% fetal bovine serum and l-glutamine (Lonza). Cells were split every three to five days to maintain 10% to 80% of confluency. Transfections into 293T cells were carried out using FuGENE HD Transfection Reagent (Roche) following manufacture's instructions.

293T cells were seeded in 6 well plates the day before transfections, so the next morning the cells would be attached to the wells and reached 60% to 100% confluency. For each well, the transfection mix contains 2μg of experimental plasmid, 1μg of EGFP expressing plasmid for normalising transfection variations between wells, 4μl of the FuGENE reagent, with H_2O topped up to total volume of 100μl. The mix was well mixed and incubated at room temperature for 30 minutes. During the 30 minutes, the seeded cells were washed by fresh medium twice followed by addition of 1ml of fresh medium. The 100μl transfection mix was then added to the well. Cells were then incubated at 37 °C with 5% CO_2 for 24 hours before harvested for analysis.

2.4 RNA extraction and Northern blotting

Total RNA was extracted from a fully confluent well of a 6-well plate of cells using

TRI reagent (Molecular Research Center Inc) following manufacture's instruction.

1. Add 1ml TRI reagent to each well after removing the medium.
2. Resuspend cells in TRI reagent by pipetting up and down a few times then transfer to a 1.5 ml microcentrifuge tube. Incubate for 5 minutes at room temperature.
3. Add 100μl of 1-Bromo-3-chloropropane (BCP, Sigma) and vigorously shake the tube for 15 seconds. Leave the sample standing for 10 minutes.
4. Centrifuge at 13,200 rpm at 4°C for 15 minutes.
5. After centrifugation, transfer the aqueous phase (typically ~450μl) into a fresh tube. Discard the lower phases.
6. Add equal volume of iso-propanol and mix well. Then centrifuge at 13,200 rpm for 15 minutes at 4 °C.
7. Wash the pellet by 70% ethanol.
8. Air-dry the pellet and resuspend in DEPC treated H_2O. Concentration of total RNA was determined by NanoDrop.

To make DEPC-treated H_2O, DEPC (diethyl pyrocarbonate) was added into de-ionised H_2O at volume ratio of 1/1000 followed by vigorous shaking. After over night incubation in a flow hood, the solution was autoclaved for 30 minutes at 121 °C. All H_2O used in RNA related experiments were DEPC treated.

For Northern blotting analysis, 5μg of total RNA were separated on (1% to 1.6%) denaturing agarose gel electrophoresis in the presence of 2.2M of formaldehyde (37%, Sigma). The sample for loading into gel contains 5μg of total

RNA, 3µl of 10x MOPS buffer (200 mM MOPS, pH 7.0, 80 mM NaAc, and 10 mM EDTA, pH 8.0), 1µl of 30x loading buffer, 5.25µl of 37% formaldehyde and 15µl formamide (Sigma). Prior to loading, the samples were heated to 65 °C for 15 minutes then cooled on ice. Electrophoresis was carried out at 10-15 V/cm in 1x MOPS buffer.

After separation, edges of the gel were cut off and the gel was washed twice by DEPC treated H_2O for 20 minutes followed by washing in 20x SSC buffer (3M NaCl, 300 mM sodium citrate, pH 7.0) for 20 minutes. Then the gel was blotted over-night by capillary transfer in 20x SSC solution onto nylon membranes (Hybond-N, Amersham) followed by UV cross-linking. The membrane was pre-hybridised in a rotating oven at 68 °C with 30 ml HYBSOL (0.15M NaCl, 0.01M NaH2PO4, 0.001M EDTA, 7% SDS and 10% PEG 8000) with 300 µl of freshly boiled 1mg/ml sheared salmon sperm DNA (ssDNA, Sigma) and 30 µl of 250 mg/ml Heparin (Sigma) in the hybridisation tube. After 3-4 hours of pre-hybridisation, the solution is replaced with 20 ml HYBSOL with ssDNA and Heparin.

After changing the solution, radiolabelled probe is added and incubated overnight in the rotating oven at 68 °C. For the probe, 10µl of Labelling 5x Buffer (including random synthetic hexadeoxynucleotide primers, Promega) is mixed with 50 ng PCR products for a probe template and heated in a boiling water bath for two minutes followed by cooling on ice. Then 2µl of dATP/dTTP/dGTP mixture (15 mM each), 1µl DNA polymerase I large fragment (Klenow, NEB), and 3µl

α-32P-dCTP (Perkin Elmer) were added to the probe solution for 30 min incubation at room temperature. Before adding into hybridisation, the probe was purified through a G50 Sephadex column followed by heating to 100 °C for three minutes and cooling on ice.

After the overnight hybridisation, the membrane was washed in 2x SSC + 0.1% SDS with four time intervals: 2 minutes, 5 minutes, 30 minutes and another 30 minutes. Then it was washed in 0.2x SSC + 0.1% SDS for 30 minutes before being sealed in Saran film and exposed to Kodak phosphor storage screen by autoradiography. Typically, the screen was scanned on a phosphor imager (Bio Rad) after 16 – 24 hours of exposure. Quantification analysis was done on Quantity One (Bio Rad).

When more than one probes were required, the labelling on the membrane was stripped off by boiling in 0.1% SDS for 15 minutes before being labelled by the next probe.

The hybridization probes were PCR fragments labelled by random hexamer priming using 32P-dCTP. Primers for producing the PCR fragments for *Adh* probe, *Luc* probe and *Egfp* probe are as (Ramanathan et al., 2008). The Adh probe targets the entire *Adh* coding region. PCR primers to amplify the *Adh* fragment were 5'-GGGAATTCACCATGTCGTTTACTTTGACCA-3' (Adh_start_FW) and 5'-CCGCCTAGGGCCGGAGTCCCAGTGCTT-3' (Adh_stop_RV) from plasmid carrying *Adh* cDNA sequence. The Luc probe targets 1 – 508 nt of *Luc* coding region. PCR primers to amplify the *Luc* fragment were:

5’-GGGcctaggGAAGACGCCAAAAACATAA-3’ (Luc_start_FW) and 5’-ATGTGACGAACGTGTACATCG-3’ (Luc_508_RV) from plasmid carrying *Luc* cDNA sequence. The Egfp probe targets the entire *Egfp* coding region. The UTR-9 probe was PCR amplified from plasmid carrying UTR-9 insertion with primers 5’-GGGCCTAGGTAACAATGAAGTTTAAGCGCA-3’ (UTR-9_FW) and 5’-GGGCCTAGGTAAATTTGGTTTTCGGTGTTC-3’ (UTR-9_RV). The UTR-4 probe was PCR amplified from plasmid carrying UTR-4 insertion with primers 5’-GGGCCTAGGTAAAATCGGTGCGGTTCAGTT-3’ (UTR-4_FW) and 5’-GGGCCTAGGTAACGCTCAAATCTGATCGCA-3’ (UTR-4_RV).

2.5 Circular-RT-PCR

The circular-RT-PCR were performed as previously described (Brogna, 1999). Schematic illustration of the procedure is shown in Fig 2.5.1. 5μg of total RNA was decapped by treatment with Tobacco Acid Pyrophosphatase (TAP, EPICENTRE) at 37 °C for two hours. Then the sample was purified by two volume of ethanol and 1/10 volume of NaAc pH 5.3. The 5’ end and 3’ end are ligated by T4 RNA ligase (NEB) at relatively low concentration: 5μg total RNA in 100μl ligation reaction. Ligation reaction was incubated at 37 °C for two hours followed by phenol/chloroform extraction.

SuperScript III (Invitrogen) was used for the following reverse transcription reaction, a primer that anneals to exon 3 of *Adh* was used to synthesis continued first strand cDNA that spans across the ligation point:

5’-CATCATAGGGGTAGAAGGTG-3’ (anti-sense). Then two rounds of PCR with nested primers were applied to map the cleavage site of *Adh* polyA signal. The first round PCR was with primers 5’-CATAACATTAGTTCATAGGGTT-3’ (sense) and 5’-CAGACCAATGCCTCCCAGAC-3’ (anti-sense) and the second round was with primers 5’-GATGCACACTCACATTCTTCTC-3’ (sense) and 5’-GACCGGCAACGAAAATCACG-3’ (anti-sense). For the experiments to map cleavage sites in UTR-4 and UTR-6, the primers used were: 5’-ATCCCACCCAGCCATCGTTG-3’ (sense) and 5’-CAGACCAATGCCTCCCAGAC-3’ (anti-sense).

At the beginning of the c-RT-PCR analysis, an optional step to remove the polyA tail was carried out as controls, as the long stretch of As might interfere the accuracy of the reverse transcription or PCR reaction. To the total RNA, oligo (dT) and RNase H treatment was carried out at 37 °C for 30 minutes to remove the polyA tail. Then the sample was phenol/chloroform purified to original volume before de-capping.

After c-RT-PCR, the resulted PCR products were cloned into a pBlueScript (pBS) based T-vector followed by sequencing with M13 forward primer that anneals to the pBS.

To make the T-vector, we followed a protocol as (Holton and Graham, 1991). 10μg of pBlueScript II KS+ was digested by Eco RV restriction enzyme for five hours at 37 °C. Then the digested plasmid was incubated with Taq polymerase in a PCR reaction with only dTTP and incubated for two hours at 72 °C. Then the

plasmid was purified and treated with T4 DNA ligase over night at room temperature. This step ligated all the plasmids that failed to be tailed. Linear and circular plasmids were separated on agarose gel electrophoresis: the single band of 3kb, corresponding to liner form of pBS, was purified from gel and used for cloning PCR fragments.

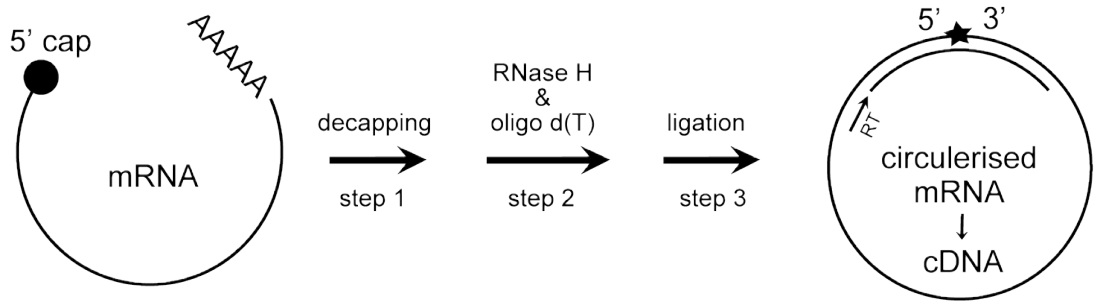
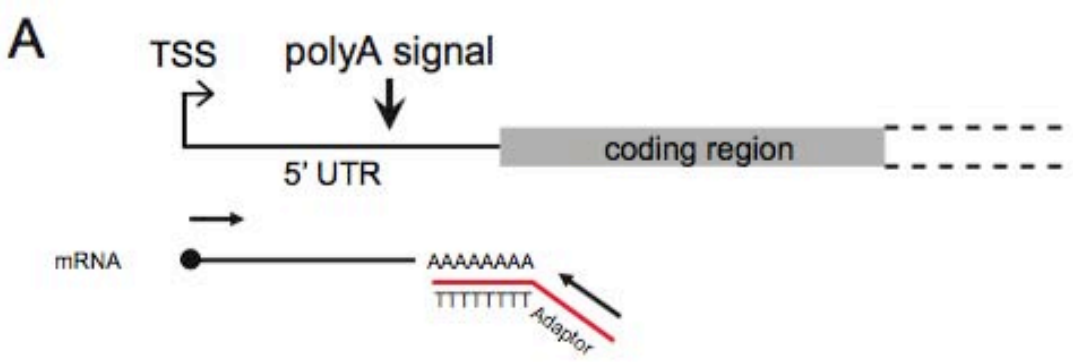


Fig 2.5.1 Schematic for circular-RT-PCR. The open circle on the left represents an mRNA with a 5’ cap and a polyA tail. Decapping, polyA tail removal and ligation are sequentially carried out. Note that the polyA tail removal step is omitted when polyA tail length is measured. On the circularised mRNA, the star indicates ligation point with joint 5’ end and 3’ end of mRNA. Arrow with RT represents gene specific reverse transcription primer.

2.6 Adapter-RT-PCR

Adaptor-RT-PCR was carried out following a published protocol (Moucadel et al., 2007). In the reverse transcription reaction, an adaptor-oligo(dT) is used: 5'-TAGAATTCAGCATTCGCTTCTTTTTTTTTTTTTTTT(C/G/A)-3'. In the following PCR reactions, nested gene specific sense primers and adaptor targeting anti-sense primer (5'-TAGAATTCAGCATTCGCTTC-3') are used (Fig 2.6.1). After two rounds of PCR using nested primers anneal to the 5' end of the 5' UTRs (see Fig 2.1.1 for first round FW primers and the Table in Fig 2.6.1 for second round FW primers), all visible bands on 2% agarose gel were purified and cloned into a T-vector followed by sequencing.

A



B

CG1322_35_FW	TTA ATT GAA GTT TCG GCG
CG7530_67_FW	TTT CCA ATC CCT GTT CCG
CG6433_57_FW	GCG GAT CGC ACG ACT ATA
CG5758_36_FW	CGC CAA AGG TGC TAA GCT
CG6179_61_FW	GTC CCC GCC GAA ATA ATA
CG9164_50_FW	ATT GCG CAT ACG ACG TGT
CG17299_63_FW	AAT CCA GTG TGC GCA ATG
CG17046_62_FW	ATT TAA TTG CGC GAT GAG
CG7628_86_FW	TGG ATT TTT GTT TCG ACC
CG17117_36_FW	GTG TCT GTG TTG CGT TGC

Fig 2.6.1 The adapter-RT-PCR assay. (A) Schematic of the assay. (B) List of 5' UTR-specific primers used in the second round of PCR. First round of primers were same as FW primers in Fig 2.1.1. The reverse primer was 5'-TAGAATTCAGCATTCGCTTC-3' in both rounds.

2.7 RNAi

Sequences of double strand RNA (dsRNA) probes were obtained from GenomeRNAi (Gilsdorf et al., 2010). To make the dsRNA probes, the corresponding PCR fragments were PCR amplified from genomic DNA extracted from S2 cells. All PCR primers carry T7 promoter sequence (5'-TAATACGACTCACTATAGGGA-3') at the 5' ends. The 3' halves of the primers specific for the targets are listed in Table 2.7.1. The dsRNAs were produced using T7 RiboMAX Express RNAi System following the protocol provided (Promega).

The initial PCR DNA fragments with T7 promoter sequence attached at both ends were purified with the High Pure PCR Product Purification Kit (Roche). DNA concentration was determined by NanoDrop; typically, 100 ul of ~100ng/μl DNA was obtained. For the dsRNA synthesis, 10μl RiboMAX Express T7 2X Buffer, 8μl template DNA and 2μl Enzyme Mix were incubated at 37 °C for 30 minutes in a thermal cycler. The mix was then heated to 70 °C for 20 minutes then slowly cooled down to room temperature. Then 1μl of 1 unit/μl RNase-Free DNase was added to the mix and incubated 37 °C for 30 minutes to remove the DNA template. Then 1μl of freshly 1 : 200 diluted RNase solution (4 mg/ml) was added to the reaction and incubated 37 °C for 30 minutes to remove any remaining single stranded RNA. Then the reaction was purified by ethanol precipitation (2.5 volumes of 95% ethanol and 0.1 volume of 3M sodium acetate (pH 5.2) were added and incubated at -20 °C for

30 before spun at top speed in a microcentrifuge for 20 minutes). The pellet was washed by 500μl of 70% ethanol followed by air-drying. Finally the pellet was resuspended in 200μl DEPC treated H_2O. The concentration of dsRNA was determined by measuring absorbance at 260nm on the NanoDrop. 1μl of the dsRNA was examined by gel electrophoresis to check the integrity of dsRNA.

For RNAi in S2 cells, 7.5μg of corresponding dsRNA was added to each well immediately after transfection of the plasmid DNA. After five hours incubation with the transfection mix, the medium was replaced with 1ml of fresh serum-free medium. Then, the dsRNA was added to the medium followed by gentle swirling and tilting. The dsRNA was incubated with transfected S2 cells for 30 minutes, followed by addition of 2ml complete medium. Cells were further incubated for 3 days before harvested for RNA extraction.

Primer	Tm	Sequence	Size
rrp6_ds_FW	58	GCCTGCTGAACTTTTTCGAC	400
rrp6_ds_RV	57	AGCCGACACAAGAAGAGGAA	
upf1_ds_FW	58	ACGTTTCAGGTTGCGGTATC	499
upf1_ds_RV	58	GAACTTCATCTACTCGCGCC	
Dis3_ds_FW	57	CGTAACGATTGACACATGGC	551
Dis3_ds_RV	57	TGATGACGCTCTTGTGGAAG	
Ski6_ds_FW	57	CTCGATGACGGTTTCCAAGT	552
Ski6_ds_RV	57	AAACTGGGAGTCTTCGAGCA	
Mtr3_ds_FW	58	GGGCGATCTGTTCAAAGGTA	529
Mtr3_ds_RV	58	GGTTCCGCCTACATGGAGTA	
rrp40_ds_FW	58	GAAAGACACACGCGGATTTT	347
rrp40_ds_RV	57	GATATGGAGGCTGTGTCCGT	
rrp46_ds_FW	58	GTATAGCTTTGCGGCTGTCC	296
rrp46_ds_RV	57	ACGTCTCATTGACGTCCTCC	
rrp42_ds_FW	56	TCGGACAATCCGTATGACTG	333
rrp42_ds_RV	58	TCCTACTATTTGCGGTCGCT	
Csl4_ds_FW	58	CCATATTCAGCGTCCTCTGG	541
Csl4_ds_RV	57	TTCATCTAACCATCCCAGCC	
rrp4_ds_FW	57	GCAGTAGGATGGAGTCGAGC	366
rrp4_ds_RV	58	TTTGTTTACCTCGCTTTGGC	
rrp45_ds_FW	57	GAGGTTGCTGCTCCTGTTTC	384
rrp45_ds_RV	57	TGCTTTTCCACCACTATCCC	
pcf11_ds_FW	58	GCGAAGTGGCTTTCCTAGTG	641
pcf11_ds_RV	57	TCTCCCAAAAGGAATGATGC	
cpsf_ds_FW	58	TCGGCTGGTTAACCGTAAAG	435
cpsf_ds_RV	58	GTTCTGGAGCTAAGGCATCG	
cstf_ds_FW	58	CAGGAGACGGCTTTAAGTGC	657
cstf_ds_RV	58	ATTGGGTAGAGAAGCTCGCA	
lacZ_ds_FW	60	CTGTCGTCGTCCCCTCAAAC	547
lacZ_ds_RV	58	CGTTTCACCCTGCCATAAAG	
Rtr1_ds_FW	58	CGTTCCCAAGCAAAAGTACAG	200
Rtr1_ds_RV	58	CGATGGTCAAGTATTTCGTGG	
Cdk9_ds_FW	57	CATACTGTTGTCCTGGGGCT	375
Cdk9_ds_RV	58	CAGCTATGCGGCTCCTTTAC	
CycT_ds_FW	57	CATGGATGGTGGTACAGCAG	703
CycT_ds_RV	57	AACTCCGATGACCAGTTTGG	
Fs(1)h_ds_FW	57	TCCTCATCCGAGTTGGATTC	540
Fs(1)h_ds_RV	57	GAACAAGGAGAAGCTGTCGG	
Cbp20_ds_FW	56	TCGCATCTGTGGAATTAAGC	623
Cbp20_ds_RV	56	TGGGTGCAATCTTCTGTGAC	
Cbp80_ds_FW	57	CATGATCGATGTCTCCAACG	633
Cbp80_ds_RV	58	ATATGAAAGAGCTCGGCGAA	
Fcp1_ds_FW	58	CCGAATCTTCGGAACGATAA	362
Fcp1_ds_RV	57	CACCAGATGCTGAAAAAGCA	
T7-promoter		to the 5' end: TAATACGACTCACTATAGGGA	

Table 2.7.1 Primers used to generate dsRNA probes. Target genes of RNAi are indicated by primer names. The Tm of each primer is indicated. Sizes of PCR products are listed in the right column.

2.8 Real-time PCR

Real-time PCR was performed using SYBR Premix Ex Taq II (TAKARA) on the ABI PRISM 7000 real-time PCR system. Firstly, oligo(dT) primed reverse transcription of 5μg of total RNA was carried out by using SuperScript III (Invitrogen). Quantitations of the transcripts were produced with the ABI Prism 7000 SDS software by analysing amplification profiles of the real-time reactions. Primers for the measured genes are listed in Table 2.8.1. For genes treated with dsRNA but with no primers listed in Table 2.8.1, the primer pairs used for amplifying the DNA fragments for the corresponding dsRNA (in Table 2.7.1) were used. All the validating primers target exon regions. The levels of mRNA are normalised against that of *Rpl32*, which is amplified by primers 5'-CGCCGCTTCAAGGGACAGTAT-3' (Rpl32_FW) and 5'-TCTTGAGAACGCAGGCGACCG-3' (Rpl32_RV).

Before carrying out real-time PCR, GoTaq (Promega) PCR was used to confirm that single band was produced with each primer pair. Semi-quantitative PCR using GoTaq with 18 – 25 cycles (empirically determined for each tested gene) was also used to compare mRNA levels.

Primer	Tm	Sequence
Trf4-1_Val_FW	59	CCATTTGTTGGCTGGTTTCA
Trf4-1_Val_RV	62	GCCTTGTTCTCCGGTCGTTT
Trf4-2_Val_FW	60	CCGACCAACGACATAGGGAG
Trf4-2_Val_RV	55	CAAGGTCGGCAACTATGTCT
rrp6_Val_FW	53	GCCCTTTACCTAAGCTATCC
rrp6_Val_RV	55	ACCATTAGTTCGGTTTCTGC
upf1_Val_FW	62	CCTGGACATGGACGACAACG
upf1_Val_RV	59	CTGTTCGCTGGGTTGCTTTA
Dis3_Val_FW	58	GACCGCACCAGGAACTTCTA
Dis3_Val_RV	56	GGTGCCCTCCAAGTAGTTTT
Ski6_Val_FW	64	TGCGGGAATCTGCCTCAATG
Ski6_Val_RV	62	CGCAAAGTCAGAGGCACTGC
Mtr3_Val_FW	55	CAGTGCGGATATCTCAGTCC
Mtr3_Val_RV	63	TCCTCCCGGATGTTCGATTG
rrp40_Val_FW	66	CTTTGAGGCGGCCAGCAAGA
rrp40_Val_RV	64	TGTCGATTTCCGCACATCCC
rrp46_Val_FW	59	TTTACCAGCGAAAATGGACG
rrp46_Val_RV	64	AATTTACGGCGCAGGCATCA
rrp42_Val_FW	60	GCGTGGAGGATGACTTTCGT
rrp42_Val_RV	63	TGGCGATTCGTAGGCGTTCT
Csl4_Val_FW	54	TGAGTGAACAGCAGGATGAG
Csl4_Val_RV	58	CTTGGCGAATTTTGGAGTTG
rrp4_Val_FW	60	TCTGCCAGGCGGAGAACTAC
rrp4_Val_RV	60	CGTACTGGATGCTGGTGTCG
rrp45_Val_FW	65	CGGGCAAATCGAAGCCAGAG
rrp45_Val_RV	54	GCCGTCCTTATACCAGACAT
pcf11_Val_FW	59	ATCGCTATGTTCGCAATGGA
pcf11_Val_RV	58	TCGTGGGATTTGAGTTGAGC
cpsf_Val_FW	61	GGTTGGACGGGTCGCTATTT
cpsf_Val_RV	61	AACGCAAAGCGTAAGCACGT
cstf_Val_FW	57	CAATGTCCATCCGAACGATA
cstf_Val_RV	57	TGACAATACTGACCCGTTGC
Pcf11.Val2.FW	55	ACATCAACTACGCCCACATC
Pcf11.Val2.RV	61	AGGCATCCCAGGAACCAAAC
Cdk9_Val_FW	70	CTCCAGCAGCCTTCGGGGTCG
Cdk9_Val_RV	67	GCCAAGCCAAAGTCAGCCAGC
CycT_Val_FW	62	AGCCAGTGCCTCAGTCTCAGC
CycT_Val_RV	51	ATGGACACAGACTCTCCTTTA
Fs(1)h_val_FW	63	GTGAGCCACCGCCTCGTTAC
Fs(1)h_val_RV	60	ACCTGGTCCGCTGGTAACTG
Cbp80_val_FW	59	TCTGCTACGGCTCCATTTTG
Cbp80_val_RV	59	ATCTCCTCGCCGCTATATCC
Fcp1_val_FW	59	CGCTACAGAAGCACCCAAAG
Fcp1_val_RV	58	ACCGCCACTAGATGCGTTAT

Table 2.8.1 Primer list for RT-PCR validation of RNAi depletions. The Tm of each primer is indicated.

Results chapters:

Chapter 3 PolyA signals are found at the beginning of Drosophila genes

3.1 PolyA signals are predicted in the 5’ UTR of many Drosophila transcripts.

Following the unexpected finding that a 5’ UTR sequence showed significant function of a polyA signal when located at 3’ end of a reporter gene (Introduction, Section 1.4), we hypothesised that other genes might also contain polyA signals in 5’ UTRs because untranslated regions are typically A-U rich. It was therefore decided to search for putative polyA signals using two computer programs developed for predicting mammalian polyA signals: Polya_svm and Polyadq (Cheng et al., 2006; Tabaska and Zhang, 1999).

Polyadq identifies AATAAA or ATTAAA then scans the downstream region (1-100 nt) for a putative DSE (Tabaska and Zhang, 1999). Limitations of Polyadq include that it does not scan for other known hexamer variants or for additional flanking sequence motifs present in polyA signals (Hu et al., 2005; Tian et al., 2005). The more recent program Polya_svm searches for polyA sites by using a window-based scoring scheme across a wider region. The scoring result reflects fitness for 15 cis elements which were identified within the 100 nt upstream and downstream of known human polyA sites (Cheng et al., 2006; Hu et al., 2005).

Polya_svm showed 33.8% higher sensitivity and better specificity than Polyadq in predicting human polyA signals (Cheng et al., 2006). For detailed algorisms, see Methods sections of the original papers from developers (Cheng et al., 2006; Tabaska and Zhang, 1999).

The following bioinformatic analysis for predicting Drosophila polyA signals was performed in collaboration with Prof Gos Micklem and Dr Matthew Garret (University of Cambridge). Because both Polyadq and Polya_svm were developed for predicting mammalian polyA signals, we firstly tested whether they can correctly identify the polyA signals within annotated Drosophila 3' UTRs. Both programmes were tested on available 3' UTRs of *D. melanogaster* transcripts. The whole dataset of *D. melanogaster* 3' UTRs and 5' UTRs were downloaded from Flybase (server: ftp://flybase.net/genomes/, folder Drosophila_melanogaster/, RELEASE dmel_r4.3_20060303). The sequences were extended by 150 nt at the 3' end with the corresponding genomic sequence to allow the programmes to evaluate downstream elements. Both programs were run as previously described (Cheng et al., 2006; Tabaska and Zhang, 1999). In the Polya_svm results, the E-value represents the probability of being a polyA signal and the higher the probability, the lower the E-value.

In the predictions, all target sequences are transcribed sequences mapped by EST. Great majority of 5' UTRs and 3' UTRs appear intron-less. In the 150 nt extension of genomic sequence without annotation of possible splicing sites. This could bring a fraction of false positives if downstream sequence motifs were located

in intronic regions. Therefore, results from these polyA signal predictions will be solely based on the sequences. At this stage, we proceeded to test whether the programmes were applicable for polyA prediction in Drosophila. Results from later stages shows that the predictions are only moderately accurate and probably under represents the number of true hits. Given the lack of prediction accuracy, we did not further investigate this issue.

From the collection of known *D. melanogaster* 3' UTRs, Polya_svm identified polyA signals in 7587 3' UTRs (corresponding to 6053 individual genes) out of the total 13562 annotated 3' UTRs (10019 genes) (Fig 3.1.1 and Fig 3.1.2). Among these, 186 3' UTRs (148 genes) contain multiple polyA signals and 7401 3' UTRs (5905 genes) contain single polyA signals. However, to what extend these predicted a polyA signals overlap with the ones genuinely used is not further validated. Because the programme already seemed to under-predict and only a small fraction of 3' UTRs were predicted to contain multiple hits, it seems Polya_svm probably under estimated true Drosophila polyA signals. Polyadq identified fewer polyA signals: 2278 3' UTRs (1841 genes) contain single polyA signals while 242 3' UTRs (195 genes) contain multiple polyA signals. The overlap between Polya_svm and Polyadq results covers 1594 genes, which is over 60% of Polyadq prediction. Pie charts of above data are presented in Fig 3.1.1 and Fig 3.1.2. These results indicate Polya_svm and Polyadq can predict Drosophila polyA signals to an acceptable level. Polya_svm shows higher sensitivity, as shown in human predictions (Cheng et al., 2006). However, both programmes missed a sizable

fraction of genuine polyA signals in 3’ UTRs. Two factors may contribute to this: one is that sequence requirement for Drosophila polyA signals are different than in human genes; another is the limited prediction power of the programmes – Polya_svm only predicted 15,469 polyA sites (52.8% of the total) that are within 24 nt from the real human polyA sites (Cheng et al., 2006). In general, both Polya_svm and, to lesser extent, Polyadq can predict genuine polyA signals in Drosophila.

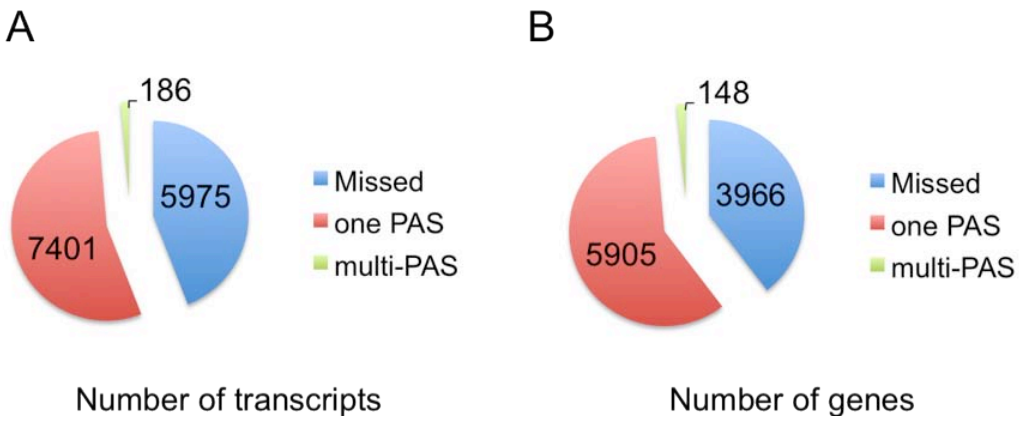


Fig 3.1.1 Polya_svm identifies PolyA signals in Drosophila 3' UTRs. (A) Pie chart showing proportion of transcripts with one (red sector), none (blue) or multiple (green) polyA signals. Analysis based on 13562 annotated 3' UTRs. (B) Hits as in A but re-classified by the number of genes carrying predicted polyA signals. No PAS represents number of genuine polyA signals missed by the prediction programme.

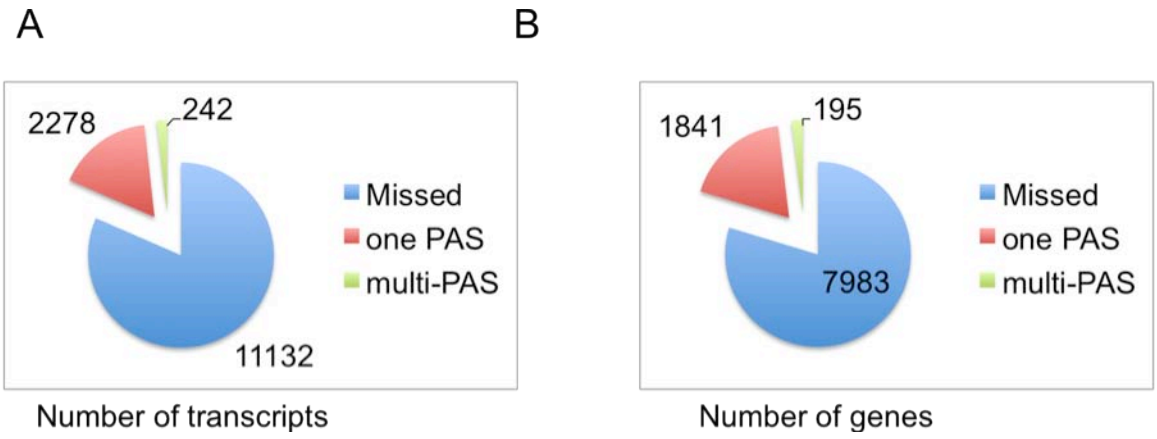


Fig 3.1.2 Polyadq identifies PolyA signals in Drosophila 3' UTRs. (A) Pie chart showing proportion of transcripts with one (red sector), none (blue) or multiple (green) polyA signals. Same data input as Fig 3.1.2 (B) Hits as in A but re-classified by the number of genes carrying predicted polyA signals. No PAS represents number of genuine polyA signals missed by the prediction programme.

We then searched for polyA signals in all known 5' UTRs. The 5' UTR sequences were also downloaded from Flybase (server: ftp://flybase.net/genomes/, under folder Drosophila_melanogaster/, RELEASE dmel_r4.3_20060303). The sequences were also extended at the 3' ends with 150 nt of the corresponding genomic sequences. This extension could include intronic regions, although the results only show hits within the UTRs. Out of 18911 annotated 5' UTRs (including alternative transcripts), Polya_svm predicted 3389 polyA signals (corresponding to 2380 individual genes). Of these, 397 5' UTRs (321 genes) showed multiple polyA signals (Fig 3.1.3). Polyadq identified polyA signals in 1101 5'UTRs, corresponding to 876 individual genes. Of these, 112 (94 genes) showed multiple signals (Fig 3.1.4). 5' UTRs of 483 genes are predicted to contain polyA signals by both programmes (Fig 3.1.5). Furthermore, the number of putative polyA signals is likely to be an underestimate because both programs missed polyA signals in the 3' UTRs of experimentally verified transcripts. The full lists of hits identified by both programmes are available in the Brogna lab and the submitted manuscript. In summary, these analyses clearly indicate that polyA signals are common in 5' UTRs.

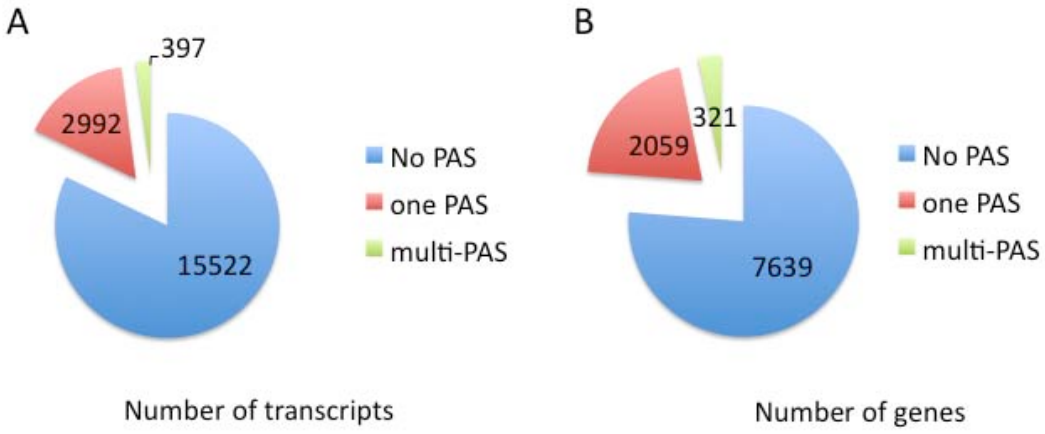


Figure 3.1.3 Polya_svm identifies PolyA signals in Drosophila 5' UTRs. (A) Pie chart showing proportion of transcripts with one (red sector), none (blue) or multiple (green) polyA signals. Analysis based on 18911 annotated 5' UTRs. (B) Hits as in A but re-classified by the number of genes carrying predicted polyA signals.

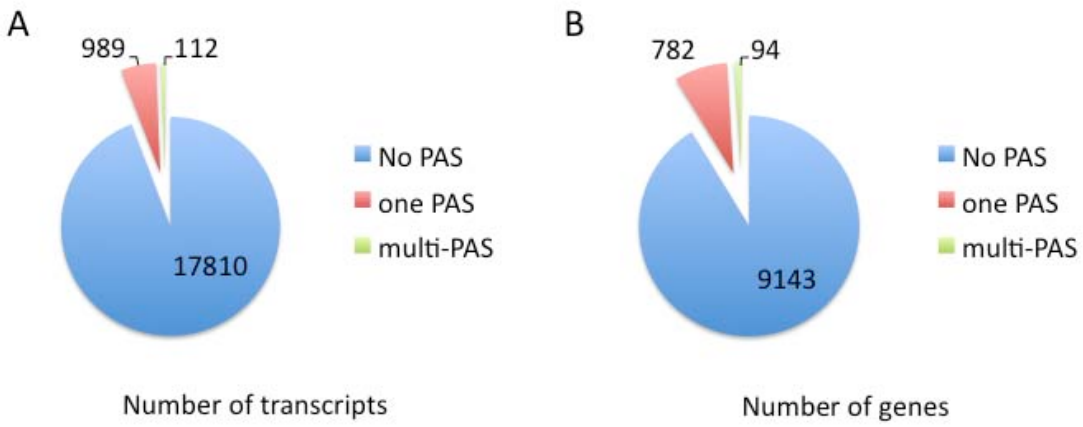


Figure 3.1.4 Polyadq identify PolyA signals in Drosophila 5' UTRs. (A) Pie chart showing proportion of transcripts with one (red sector), none (blue) or multiple (green) polyA signals. Analysis based on 18911 annotated 5' UTRs. (B) Hits as in A but re-classified by the number of genes carrying predicted polyA signals.

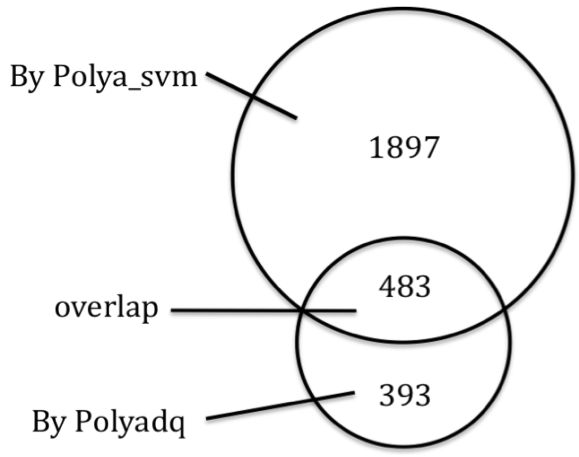


Figure 3.1.5 Number of genes shared 5' UTR hits by both Polya_svm and Polyadq. Venn diagram represents the data in Fig 3.1.3B and Fig 3.1.4B. 483 genes are predicted to carry 5' UTR polyA signals by both programmes, but the exact position on the sequence may be different.

3.2 Experimental validation of 5’ UTRs polyA signals

We next sought to test the functionality of the predicted polyA signals located in 5’ UTRs. From the list of Polya_svm hits, we obtained a sub-set of 5’ UTRs that are longer than 200 nt as they should be easier to clone. From this sub-set I selected ten sequences for experimental testing (CG1322, CG7530, CG6433, CG5758, CG6179, CG9164, CG17299, CG17046, CG42575 and CG17117; more information in Fig 3.2.1. Expression data in Appendix 4). The sequences are thereafter referred to as UTR-1 to UTR-10. Their lengths range from 342 nt (UTR-1) to 2706 nt (UTR-6). Although the lengths are significantly longer than common Drosophila 5’ UTRs, we proceeded without shortening them as the priority at this point was to test their functionality of polyA signals. Some sequences have a single polyA signal predicted (UTR-3, UTR-4 and UTR-9), others have several (UTR-1, UTR-2, UTR-5, UTR-6, UTR-7, UTR-8, UTR-10). The positions the signals refer to either the locations of the AATAAA or ATTAAA (Polyadq hits) or to the positions of the most likely polyA site (Polya_svm hits). Because of the inconsistency and inaccuracy amongst the two programmes (both bioinformatically and later experimentally), the indicated positions and strengths do not always reflect true polyA signals and their strengths.

A

UTR	Gene	Length (nt)	Position of polyA signals	Polya_svm E-value	Polyadq score
1	CG1322	342	229, 244, 271, 331	0.070	0.124
2	CG7530	894	148, 207, 383, 425	0.017	Neg
3	CG6433	442	134	0.031	0.327
4	CG5758	375	146	0.440	0.104
5	CG6179	638	408, 576, 595	0.036	0.159
6	CG9164	2706	83, 103, 136, 1234, 2192	0.062	0.758
7	CG17299	571	105, 111, 165, 248, 322, 350, 384, 557	0.093	0.501
8	CG17046	691	123, 489, 592	0.036	0.511
9	CG7628	339	198	0.041	0.270
10	CG17117	1252	79, 177, 293, 361, 489, 530, 812	0.047	0.463

B

UTR	Gene	Description of genes from *Flybase*
1	CG1322	*Zn finger homeodomain 1*
2	CG7530	unknown function
3	CG6433	*quail*
4	CG5758	unknown function
5	CG6179	unknown function
6	CG9164	unknown function
7	CG17299	*SNF4/AMP-activated protein kinase gamma subunit*
8	CG17064	*mars*
9	CG7628	*phosphate transporter, inorganic phosphate transporter*
10	CG17117	*homothorax*

Fig 3.2.1. Lists of experimentally characterised 5' UTRs. (A) First column gives short name of the sequences as used in this study. Second column reports the 5' UTR gene name. Third column indicates the lengths of the sequences. Fourth column indicates at which position the predicted polyA signal. Nucleotide distance is from the 5' end of the sequence. Last two columns show the E-value from Polya_svm prediction and scores from Polyadq. The lower the E-value is, the stronger is the polyA signal. The higher the Polyadq score is, the stronger is the polyA signal. (B) Annotated functions of the ten genes.

To experimentally test these putative polyA signals, I PCR amplified the 5'UTR from adult fly genomic DNA and then cloned the fragments in the intergenic spacer of a dicistronic *Adh-Luc* reporter. This reporter is derived from the *Adh-Adhr* dicistronic gene in Drosophila and has previously been described (Brogna, 1999; Ramanathan et al., 2008). In this plasmid construct an Avr II site located between the coding region of *Adh* and *Luc* was used as cloning site. The principle of using the dicistronic reporter for this analysis is that if the inserted sequence functions as a polyA signal, a monocistronic mRNA encoding only the *Adh* gene would be produced. Otherwise, transcription would read through the intergenic region and a longer dicistronic mRNA encoding both *Adh* and *Luc* should be produced by the downstream SV40 polyA signal in the plasmid (Fig 3.2.2 A). Reporters containing the original *Adh* polyA signal serve as positive controls. Negative controls are shown in Fig 3.2.4 later in this chapter. Two parallel sets of reporters were made, with either the genomic intron-containing *Adh* or the cDNA of *Adh*. This is to monitor if upstream splicing can affect these predicted polyA signals because of the dynamics between the two processes as discussed in Introduction (Kyburz et al., 2006; Millevoi et al., 2006; Niwa et al., 1990; Proudfoot et al., 2002; Rigo and Martinson, 2009; Tian et al., 2007).

The reporters were transfected into Drosophila S2 cells and total RNA was isolated 24-48 hours after transfection. Total RNA was analysed by Northern blotting with either *Adh* or *Luc* specific probes, as indicated in Fig 3.2.2. These measurements of steady state transcripts would reflect mostly the levels of stable

mRNA. The Adh probe is a PCR fragment corresponding to the full coding sequence of the *Adh* gene; this was amplified from the intron-less version of the reporter. The Luc probe is a PCR fragment that spans nt 1-508 of the *Luc* coding sequence. A plasmid carrying EGFP was co-transfected and the level of Egfp mRNA was used to normalise transfection variation. The Egfp probe is a PCR fragment of the full EGFP coding region.

Northern blots in Fig 3.2.2 show that the reporters with the original *Adh* polyA signal produced abundant level of monocistronic *Adh* mRNA (detectable by Adh probe only) and very low level of dicistronic *Adh-Luc* mRNA (detectable by both Adh and Luc probe) due to the presence of the strong Adh polyA signal (Fig 3.2.2 B, lanes 1-2 and 13-14). Intron-containing and intron-less reporters showed similar level of monocistronic mRNA while the intron-less reporter produced ~ 7 to 8 fold more dicistronic mRNA (lanes 1 vs 2, 13 vs 14). We did not have size markers on the gels, but instead relied on the sizes of transcripts produced by reporters in lane 1-2 and 13-14, which was confirmed when first used (Brogna and Ashburner, 1997; Ramanathan et al., 2008). The reporters carrying the selected 5' UTRs also produced monocistronic *Adh* mRNA, confirming all of these sequences can function as polyA signals (Fig 3.2.2 B, lanes 3-12 and 15-24). The levels of *Adh* mRNA produced by the 5' UTRs vary between sequences, reflecting the difference of strengths of polyA signals. Notably, UTR-4 contains the strongest polyA signal (Fig 3.2.2 B, lanes 9-10), despite Polya_svm predicted UTR-4 to be the weakest among the ten predicted (highest E-value, Fig 3.2.1). These results indicate that functional polyA

signals are indeed present in Drosophila 5' UTRs.

The activity of the intergenic polyA signal also determines the level of the readthrough transcript. The Northern blot analysis indicates that the levels of the *Adh-Luc* mRNA negatively correlate with the strengths of the intergenic polyA signals (Fig 3.2.2). For example, the reporters with UTR-1 and UTR-10 produced five to ten fold less *Adh* mRNA than those with the *Adh* polyA signal but 11 – 25 fold more *Adh-Luc* mRNA (lane 3-4, lane 23-24). Conversely, the reporters with UTR-2 and UTR-4 produced more *Adh* mRNA and less *Adh-Luc* mRNA comparing to the *Adh* polyA signal. However, the comparison of dicistronic *Adh-Luc* mRNA is complicated by the fact that the insertions are of different lengths. The readthrough transcripts with different sequence composition and carrying different lengths of insertions might attract different protein factors and form different mRNPs, consequently leading to different stabilities or different turnover rates. Data in Fig 3.2.2 only measures steady state mRNA and therefore those possibilities could not be assessed.

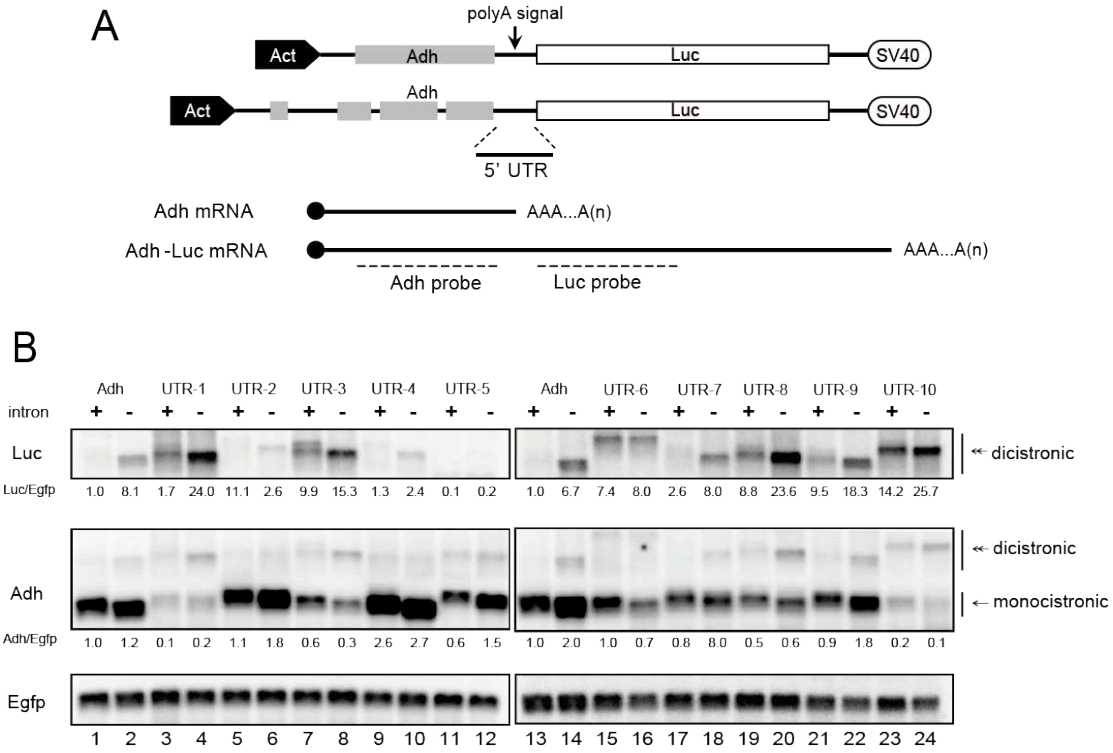

A
polyA signal
Act
Adh
Luc
SV40
Act
Adh
Luc
SV40
5' UTR
Adh mRNA
AAA...A(n)
Adh -Luc mRNA
AAA...A(n)
Adh probe
Luc probe
B
Adh
UTR-1
UTR-2
UTR-3
UTR-4
UTR-5
Adh
UTR-6
UTR-7
UTR-8
UTR-9
UTR-10
intron
Luc
Luc/Egfp
1.0 8.1 1.7 24.0 11.1 2.6 9.9 15.3 1.3 2.4 0.1 0.2 1.0 6.7 7.4 8.0 2.6 8.0 8.8 23.6 9.5 18.3 14.2 25.7
dicistronic
Adh
Adh/Egfp
1.0 1.2 0.1 0.2 1.1 1.8 0.6 0.3 2.6 2.7 0.6 1.5 1.0 2.0 1.0 0.7 0.8 8.0 0.5 0.6 0.9 1.8 0.2 0.1
dicistronic
monocistronic
Egfp
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Figure 3.2.2 5' UTR sequences contain functional polyA signals. (A) Schematics for the *Adh-Luc* dicistronic reporters. Boxes represent exons, lines UTRs and introns. Act stands for the Drosophila actin-5C promoter and SV40 for the SV40 late polyA signal present in pAc5.1/V5-His (Invitrogen). The cDNA version of *Adh* (upper diagram) encodes the full Drosophila *Adh* ORF. The genomic version of *Adh* starts from adult transcription start site and includes the 5' UTR exon (Brogna and Ashburner, 1997). Twenty-two reporters were constructed with the ten 5' UTRs listed in Fig 3.2.1 and with the *Adh-Adhr* spacer; half of the reporters carry the genomic *Adh* sequence, the others the cDNA derivative. A schematic of the expected monocistronic and dicistronic transcripts is drawn. Northern blot probes are indicated by dotted lines below. (B) Northern Blots analysis of total RNA from transfected S2 cells. Cells were co-transfected with a plasmid expressing EGFP to normalise for transfection variations. The *Adh-Luc* mRNAs were first detected with the *Luc* specific probe (top panel), and then the filter stripped and re-probed for *Adh* (middle panel); single arrowed line point to the *Adh* monocistronic mRNA and doubled arrowed line the readthrough dicistronic *Adh-Luc* transcript. Relative quantification of the signal intensities is normalised against Egfp band and reported below. Intensities of the bands are relative to that of *Adh* (middle panel) or *Luc* (top) in lanes 1 or 13 respectively.

Surprisingly, the presence of introns in the reporters has little influence on the usage of intergenic polyA signals (Fig 3.2.2). In some cases, the intron-containing reporters produced more *Adh* mRNA (UTR-3, UTR-6, UTR-10) but the opposite was found with other reporters (Adh, UTR-2, UTR-5, UTR-9). The level of *Adh-Luc* mRNA, however, is lower in most intron-containing reporters. This observation implies splicing might bring 3' end processing factors to its proximity. This restricting force might come from interactions between splicing factors and cleavage/polyadenylation factors (Kyburz et al., 2006; Millevoi et al., 2006).

It was noticed that the band densities of the *Adh-Luc* mRNA detected by Luc probe and Adh probe showed certain degree of discrepancy. For example, the relatively high level of *Adh-Luc* mRNA in lane 4 and lane 20 detected by Luc probe is under represented when later detected by the Adh probe (Fig 3.2.2). A possible explanation is that during the striping of Luc probe, a proportion of the mRNA also detach from the membrane, resulting in less contrasted results when later labelled by the Adh probe. We have observed similar issue with previous work in the lab (Ramanathan et al., 2008). Given the focus at this point is the ability of producing a monocistronic Adh mRNA, we did not further investigate this issue.

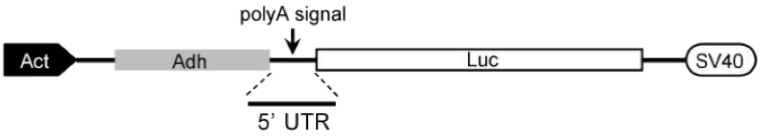
As negative controls, we tested five 5' UTR sequences that are predicted not to contain polyA signals (the bioinformatics was also in collaboration with Gos Micklem and Matthew Garret). Polyadq counter-select 5' UTRs that do not contain AAUAAA/AUUAAA, and the resulting sub-set was then scored by Polya_svm. Among the lowest scoring sequences, five 5' UTRs of 400-600 nt long were selected

for testing as above (Neg-1 to Neg-5, Fig 3.2.3A). The Northern blot analysis suggests that four sequences (Neg-2, Neg-3, Neg-4 and Neg-5) contain no functional polyA signals, as they did not produce detectable amount of *Adh* mRNA (Fig 3.2.3C, panel Adh). Instead, high levels of the dicistronic *Adh-Luc* mRNA are produced compared to the *Adh* polyA signal (Fig 3.2.3C, panel Luc). Similar results were observed with the intron-containing *Adh-Luc* reporter, indicating upstream splicing events do not activate the intergenic polyA signals (data now shown). Unexpectedly, Neg-1 produced a significant amount of *Adh* mRNA (Fig 3.2.3B, lane 2). By analysing the sequence of Neg-1, it was noticed that despite the lack of AATAAA/ATTAAA, a GATAAA hexamer is located at 200 nt from the 5' end. This hexamer accounts for 1.75% of human and 1.16% of mouse polyA signals (Tian et al., 2005) and therefore should explain the polyA activity detected. Again, this unexpected polyA activity from Neg-1 implies that there are more 5' UTR polyA signals than Polyadq and Polya_svm predicted.

A

UTR	Gene	Length
Neg-1	CG10192	416
Neg-2	CG2336	412
Neg-3	CG10808	458
Neg-4	CG7359	413
Neg-5	CG8171	561

B



C

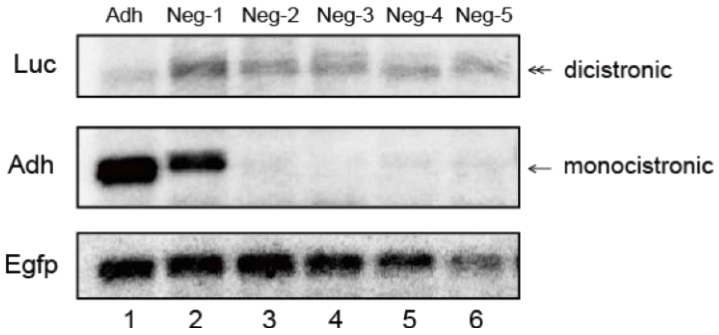


Figure 3.2.3 5' UTRs predicted not to contain polyA signals do not show polyA activity. (A) List of tested negative hits. Table shows genes of origin and lengths of the 5' UTRs. (B) Schematic of the reporters similar as Fig 3.2.2. (C) Northern blots analysis of the transcripts generated by the reporters shown in A, probes as in Fig 3.2.2.

To further characterise the 3’ ends of the mRNAs produced by the 5’ UTR-containing reporters, we used the circular-RT-PCR assay followed by cloning and sequencing of the PCR products to map the 3’ end of the mRNA (Brogna, 1999; Couttet et al., 1997) (Fig 3.2.4B). Expression of endogenous *Adh* gene is undetectable in S2 cells; therefore the c-RT-PCR products are expressed by the transfected plasmid. Furthermore, multiple copies of transcripts were sequenced and results were consistent. The results of this assay confirmed that the monocistronic mRNA produced by the *Adh* polyA signal is cleaved at the same position as the endogenous transcript in adult flies (Fig. 3.2.4D) (Brogna and Ashburner, 1997). The *Adh* mRNAs produced by the intron containing and intron-less reporters share the same 3’ ends and therefore the size difference of mRNAs (Fig 3.2.4A) is solely due to the inclusion of the small 5’ UTR exon that is only present in the genomic reporter (Fig 3.2.4C-D). We applied c-RT-PCR to map the 3’ ends produced by UTR-4 and UTR-6 containing reporters in Fig 3.2.2B. We found that in both cases the 3’ ends were generated by cleavage just downstream of the AAUAAA: 20-30 nt after the hexamer within UTR-4, and 12-21 nt after that within UTR-6 (Figure 3.2.4D). The heterogeneity of cleavage site is consistent with other studies surveying EST dataset or full cDNA sequencing (Grzechnik and Kufel, 2008; Tian et al., 2005; West et al., 2006). Sequencing indicates these mRNAs have polyA tail lengths of up to 60-80 nt, similar to that of the endogenous *Adh* transcripts (Brogna, 1999).

In summary, we demonstrated that polyA signals derived from 5’ UTR are functional as predicted: they can drive 3’ end processing downstream of the

AAUAAA hexamer and generate mRNAs with polyadenylated 3’ ends that are indistinguishable from those generated by a standard 3’UTR derived polyA signal.

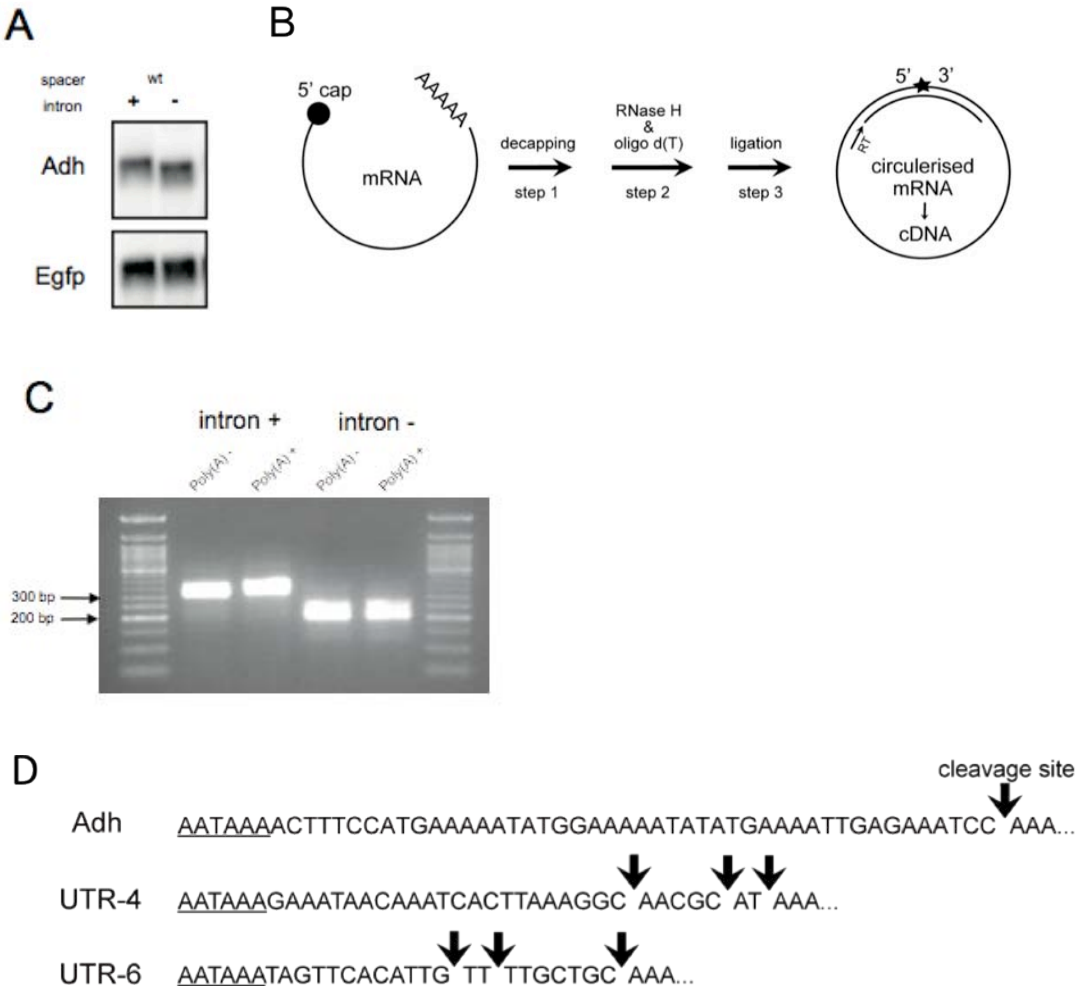


Figure 3.2.4 Characterisation of transcript 3' ends. (A) Northern blot showing size differences between *Adh* mRNA generated by the cDNA and genomic version of the *Adh-Luc* reporter (same as Fig 3.2.2, lanes 1-2). (B) Schematic of circular-RT-PCR, see Material and Methods. (C) Agarose gel showing the c-RT-PCR fragments produced from total-RNA used in A. Nested PCR using Adh specific primers were applied (primers listed in Material and Methods). Samples in lane 2 and 4 were treated with oligo(dT) and RNase H (labelled as polyA-). (D) Sequences of the 3' ends of mRNAs with cleavage sites indicated (based on sequencing of multiple clones: Adh, n=3; for UTR-4, n=4; for UTR-6, n=4).

Chapter 4 PolyA signals located in the 5' UTRs do not produce significant level of stable transcript in endogenous genes

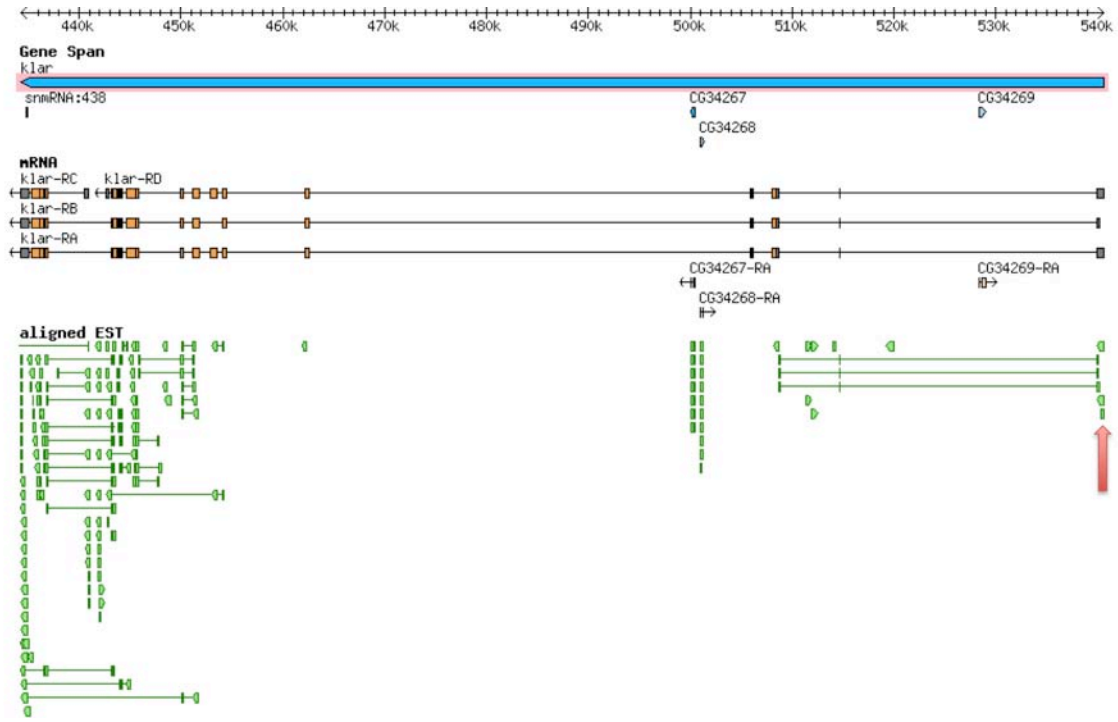
4.1 Available EST datasets suggest no endogenous usage of 5' UTR polyA signals

Since 5' UTR polyA signals are functional in reporter genes, we sought to test whether these polyA signals are used in flies during transcription of the endogenous genes. Firstly, we searched available EST sequences (Expressed Sequence Tag) for transcripts with polyadenylated 3' ends ending in the 5'UTRs. This was accomplished using the GBrowse genome viewer of Flybase where all available ESTs are mapped to the gene (Fig 4.1.1A shows the region around CG17046/UTR-8). We searched for EST tags ending in the ten 5' UTR we have tested experimentally. For nine of the ten genes, no oligo(dT) primed ESTs were found which mapped in their 5' UTRs. However, for CG17046 (UTR-8) (Fig 4.1.1A), among the 30 3' EST tags for this gene, 29 are located at 3' end but one is located in the 5' UTR (indicated by red arrow). Investigating the sequence of this EST (Genebank number EC267859.1), it was noticed that the AATAAA is 107 nt upstream of the 3' end, which exceeds the normal distance between the hexamer and cleavage site (10-40 nt). However, it is unlikely that this 3' EST was generated by miss-annealing of the oligo during the reverse transcription reaction because the 3'

end of the EST sequence is not followed by A-stretch in the genome. This EST sequence was not investigated further. In summary, the analyses of the EST datasets show no evidence of polyadenylation in the endogenous 5' UTRs.

However, the EST datasets being analysed here cannot represent complete collection of transcripts, with transcripts at low level may be under represented (Gilat et al., 2006). To examine the full potential 5' UTR polyA signal in the future, study of high sensitivity (such as high-throughput sequencing) would be required.

A



B

GGCACGAGCGAGGTTTGCGGCGCGCATTGGGCAGACTGAAGTGAACTA
AGTAAATCTGCAGCGTTCTAATTTAATTGCGCGATGAGCGCGTAAAAA
CACACCTGCGTAGCGCTGC**AATAAA**ACCAATGTAGATTTTTGTTTA
TATGTTCGCGTTTCCATCGTATTTTTCGAGTCTTATTCGTTTTAATTG
TTAGTGCTCGTACGCTAAGTGAATTGCTAGCA

Figure 4.1.1 Evidence for a putative transcript ending in the 5'UTR of CG17046 (*Klar*) gene. (A) GBrowse (version FB2010_04) view of CG17046 with available ESTs (transcription right to left). Each green bar represents an EST sequence. The red arrow points to the only 3' EST that ends within the 5' UTR. (B) Sequence of the arrow highlighted EST in A. The AATAAA is shown in bold with bigger font.

4.2 RT-PCR show non-detectable or extremely low level of stable mRNA produced by the 5' UTR polyA signals

To further investigate whether 3' processing can occur in the 5' UTR of the endogenous genes, we employed an adaptor-oligo(dT)-RT-PCR to identify 3' ends of polyadenylated mRNA (Moucadel et al., 2007). The RT reaction is primed by an adaptor-oligo(dT) primer. The following PCR uses a reverse primer for the adaptor and a gene specific forward primer.

We assayed total-RNA from S2 cells, embryos, 1st instar larvae, 2nd instar larvae and adult flies. After 30 cycles standard PCR, no visible bands were visible on the agarose gels. Clear visible bands were only produced after two rounds of PCR (totalling 50-60 cycles) with nested primers - a similar pattern of bands was seen with different RNA samples (Fig 4.2.1B). It was noticed that some PCR products appear specific to certain RNA source. However, given that the level of mRNA those PCR products represent are extremely low (no PCR products are visible on agarose gel after 30 cycles), we could not reliably quantify those products based on the gels. Additionally, great majority of the products are proved to be results of miss-annealing of the primer during reverse transcription reaction. Together, the inconsistency might be due to the fact that these experiments were done separately and the lack of repeats.

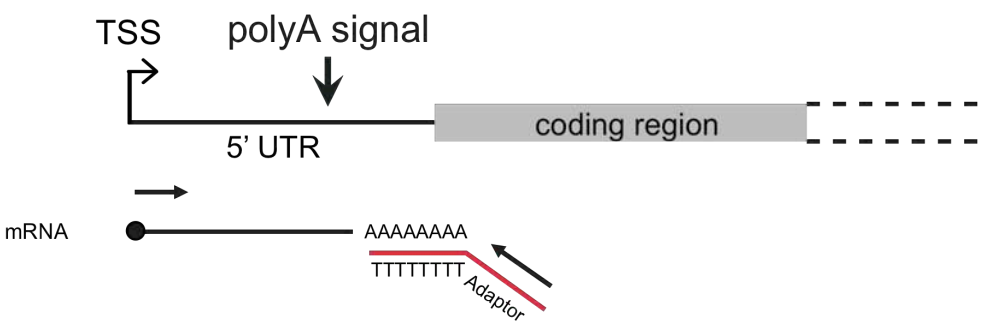
Because previous validation experiments showed inaccuracy of the prediction programmes, we did not reply on any prediction results to expect certain sizes of

products. Instead, all visible bands from the nested PCR that are within the lengths of the UTRs were purified, sub-cloned into a plasmid and sequenced. Alignments of sequence confirmed that almost all the PCR products are gene specific. However, most products are generated by miss-annealing of the oligo(dT) to stretches of eight or more As in the 5' UTRs. Out of the 37 predicted 5' UTR polyA signals, we found evidences for only three polyadenylated transcripts that might have been produced by their activation: UTR-2 (of CG7530), UTR-5 (of CG6179), and UTR-8 (of CG17046). The corresponding PCR products are circled in Fig 4.2.1B. The sequence of the circled band within UTR-2 is shown in Fig 4.2.1C. Here the cleavage site is 23 nt downstream of the AATAAA. The genomic sequence downstream of the cleavage site does not contain A-stretch, ruling out miss-annealing of the oligo(dT). Surprisingly, neither Polya_svm nor Polyadq identified this hexamer as part of a polyA signal – the hexamer is 296 nt from the 5' end. Instead, Polya_svm predicted polyA signals at 148 nt, 207 nt, 383 nt, and 425nt, whilst Polyadq identified this hexamer as an negative hit. Nevertheless, the levels of usage of these polyA signals are extremely low as the product is only detectable after nested PCR.

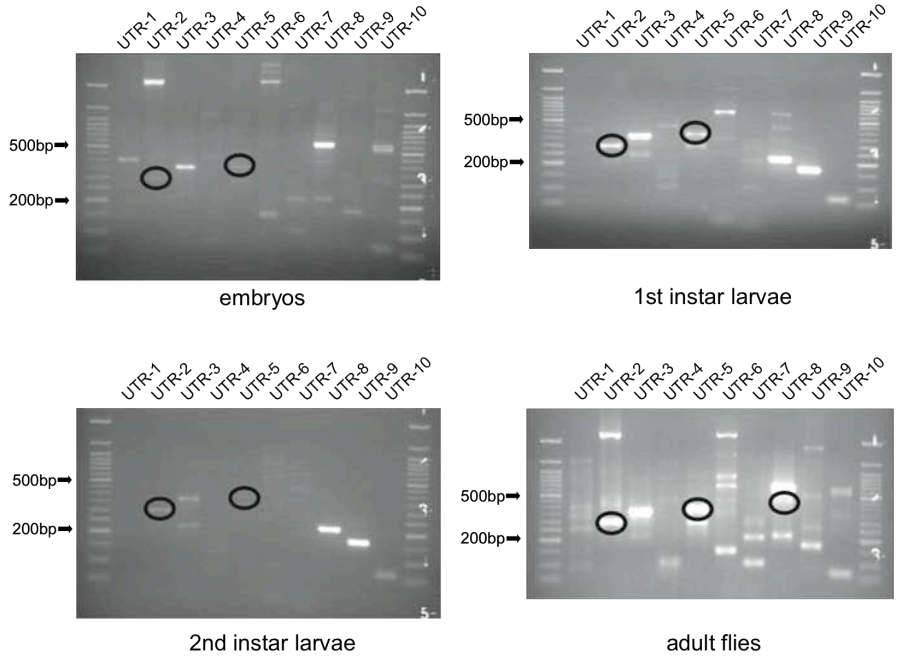
The highlighted bands produced in UTR-2 and UTR-5 are seen in all development stages tested whereas the one in UTR-8 could be detected only in adult flies. It also has to be noted that this assay could only map 3' end of polyadenylated mRNA and could not distinguish whether the transcription is initiated from promoter of the corresponding gene or a distant promoter of upstream genes.

However, the low level of these transcripts may be subjected to rapid degradation, especially if the polyadenylation was defective (see Introduction). Given the limited quantification power of the nested PCR approach, this issue was not further addressed in its endogenous context in this thesis. The reporter systems used later show that RNAi against exosome is not sufficient to increase the level of transcripts produced by early polyA signals. However, RNAi system also has the problem of not being able to achieve 100% depletion. In the future, highly sensitive experimental approaches and more thorough disruption of degradation mechanisms might better address this possibility.

A



B



C

sequence of fragment highlighted in lane UTR-2

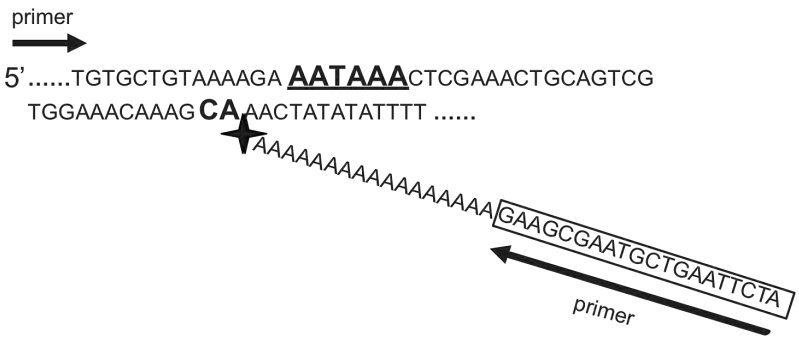


Figure 4.2.1 PolyA signals in 5' UTRs are rarely used in Drosophila. (A) Schematics for adaptor-RT-PCR assay (see Material and Methods for details). (B) Agarose gels showing PCR products produced by the adapter-RT-PCR from total-RNA derived from embryos, 1st instar larvae, 2nd instar larvae and adult flies (equal mix of males and females). All ten 5' UTR used in this study were tested. The bands highlighted by ovals correspond to polyadenylated mRNAs, as detected by sequencing of several clones of the PCR products. The other fragments are derived from miss-annealing of the RT primer in the 5'UTR. (C) Sequence of the UTR-2 polyA signal indicating the position of the polyA site (star) detected by sequencing of clones of the PCR fragment highlighted in B.

4.3 15 case studies of available microarray data do not show gene expression profiles correlate with having polyA signals in 5' UTR

Although the 5' UTR polyA signals do not appear to be detectably used, transcription elongation might still be affected as the presence of the AAUAAA alone was shown to pause Pol II (Nag et al., 2006). Therefore, we investigated whether the presence of polyA signal, not necessarily usage of it, would affect gene expression. To investigate whether there is any correlation between level of transcription and the presence of a 5' UTR polyA signal, expression data based on microarray results were analysed in Flybase (the data are constantly updated, the analysis was carried out in June 2010). This analysis shows that the genes from which the ten 5' UTRs derive are well expressed throughout the fly life cycle. Screen shots are included in Appendix 4. The ten genes with 5' UTR polyA signals (UTR-1 to UTR-10) show indistinguishable expression profiles from the five genes that do not to carry 5' UTR polyA signals (Neg-1 to Neg-5). Furthermore, the three genes with weak but detected polyA activities (UTR-2, UTR-5 and UTR-10) do not have distinguishable expression profiles. This set of very limited search suggest that bearing potentially active polyA signals in 5' UTR of gene do not seem to affect gene expression in a consistent way (see Appendix 4 for details).

However, similar to the RT-PCR results in section 4.2, this observation cannot rule out involvement of possible degradation. Also, the microarray data have limited

power of detecting transcripts of low level. In addition, usually there is only one microarray probe for each gene, targeting part of an exon. Given the complex post-transcriptional regulation pathways, the level of transcript for the particular exon may not reflect the expression of the gene.

In summary, these preliminary analyses in this chapter suggest that presence of potentially active polyA signals in the 5' UTR do not seem to trigger premature 3' end processing or influence gene expression to a drastic level. Therefore, recognition and processing of polyA signals by Pol II would probably require additional input other than the sequence motifs.

Chapter 5 Close proximity to the transcription start site silences polyA signals

5.1 PolyA signals are silenced when positioned close to the transcription start sites in Drosophila cells.

The finding that the 5'UTR polyA signals are very active when placed at the 3' end but seem silent in their natural location at 5' end raised the question whether recognition and processing of polyA signal depends on the sequence relative location in the gene. We investigated the possible positional effect on polyA signals recognition in S2 cells. Initial studies were carried out on with the *Adh-Luc* reporter similar to those I used before. The UTR-9, one of the sequences with a strong polyA signal, was inserted at three different positions in the *Adh* coding region, generating three constructs: Adh-UTR-9-P1, Adh-UTR-9-P2 and Adh-UTR-9-P3 (Fig 5.1.1). In these constructs, the distance between the AAUAAA and transcription start site (TSS) are 509nt, 695nt, and 926nt respectively (Fig 5.1.1A). UTR-9 is chosen for these tests because it contains only one AAUAAA hexamer and has shown reasonably strong polyA activity.

The constructs were transfected into S2 cells and total RNA was analysed by Northern blot as described above. The results show that at P2 and P3 the UTR-9 sequence functions as an efficient polyA signal and produced the expected truncated *Adh* transcripts (Fig 5.1.1B panel Adh, lanes 2-3). In contrast, when located at P1,

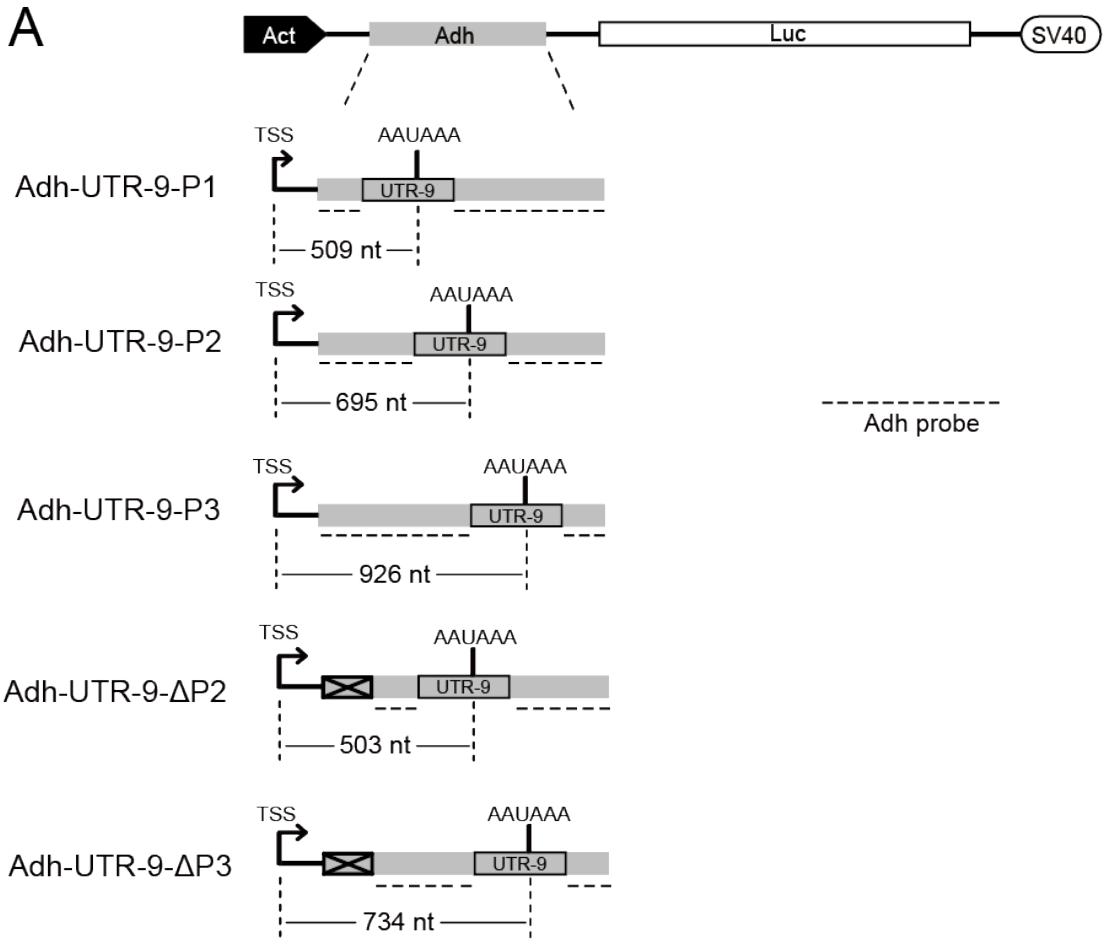
UTR-9 only produced trace amount of the truncated *Adh* transcript (Fig 5.1.1B panel Adh, lane 1). The P1 transcript is barely visible as shown in B, but can be visualised when the filter was exposed for a longer time (data not shown). A readthrough generated by the intergenic *Adh* polyA signal is detected. The presence of this transcript indicates active transcription of the reporter (Fig 5.1.1B panel Adh, readthrough). The low level of readthrough in lane 2 may be explained by the strong P2 transcript. However, readthrough level increased in lane 3 along with the stronger P3 transcript. We do not have experimental explanation for this contradiction, but speculate be that two closely positioned polyA signals (the UTR-9 polyA signal at P3 and Adh polyA signal in the intergenic spacer) might enhance each other's activity.

Next, we tested if the polyA activity at P1 is inhibited by the sequence at the beginning of the *Adh* coding region (nt 1-192 of *Adh* cDNA). The first 192 nt of the *Adh* coding region were deleted from the Adh-UTR-9-P2 construct. In the resulting construct, the AAUAAA is 503 nt from the TSS in the Adh-UTR-9-ΔP2 (Fig 5.1.1A). Indeed, Northern blot shows that the efficiency of the polyA signal at ΔP2 is significantly reduced, to similar level as the P1 (Fig 5.1.1B, lane 4 vs lane 2). However, deletion of the same 192 nt sequence from reporter Adh-UTR-9-P3 to yield Adh-UTR-9-ΔP3 caused only a moderate reduction (Fig 5.1.1B, lane 5 vs lane 3).

It was noticed that the use of the Adh probe (recognising entire *Adh* coding region) might result in different hybridisation efficiencies against P1, P2 and P3

transcript. However, the RT-PCR in the following Fig 5.1.2, which does not have the same issue, shows similar result as the Northern blot. In addition, in later experiments in Fig 5.1.5 and Fig 5.1.6, we avoided this potential under representation by using probes that have identical recognising region amongst all expected transcripts.

These experiments suggest that the UTR-9 polyA signal becomes silent when it is close to the TSS regardless of the upstream flanking sequence. When the AAUAAA is as close as ~500 nt to the TSS, the polyA signal does not seem to be recognised. Whereas when the AAUAAA is ~700 nt or further away from the TSS, the polyA signal is efficiently processed. This observation may explain why 5' UTR polyA signals are not appreciably used in the endogenous genes (Chapter 4).


A
Act
Adh
Luc
SV40
TSS
AAUAAA
Adh-UTR-9-P1
UTR-9
509 nt
Adh-UTR-9-P2
695 nt
Adh probe
Adh-UTR-9-P3
926 nt
Adh-UTR-9-ΔP2
503 nt
Adh-UTR-9-ΔP3
734 nt

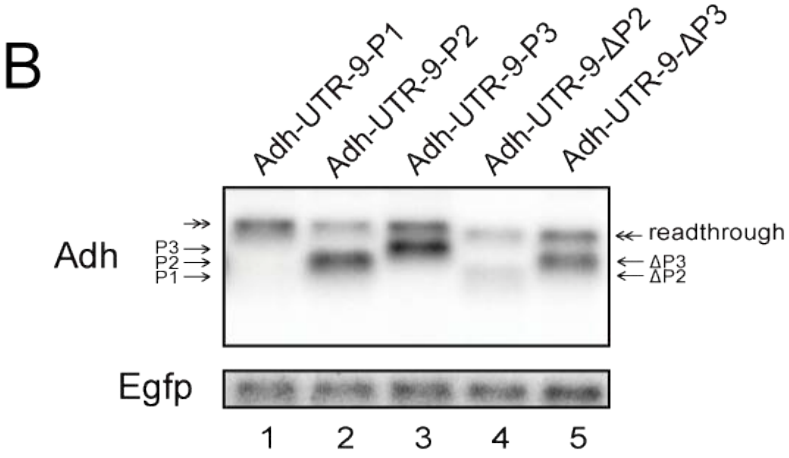

B
Adh-UTR-9-P1
Adh-UTR-9-P2
Adh-UTR-9-P3
Adh-UTR-9-ΔP2
Adh-UTR-9-ΔP3
Adh
P3
P2
P1
readthrough
ΔP3
ΔP2
Egfp
1
2
3
4
5

Fig 5.1.1 PolyA signals become silent when close to 5' end. (A) Schematics of reporters with UTR-9 inserted at different positions in the *Adh* coding region. Distance from TSS to AAUAAA is indicated below. Horizontal dotted line below each reporter indicates Adh probe recognising region. (B) Northern blots of total RNA of S2 cells transfected with the reporters in A; probes as in Fig 3.2.2. Adh panel: top band (doubled arrowed line) is the readthrough mRNA processed at the intergenic *Adh* polyA signal. Truncated transcripts processed at early polyA signals are indicated: P1, P2 and P3; ΔP2 and ΔP3 indicate mRNA derived by the deletion derivative lacking the initial region of *Adh*.

To further characterise the truncated transcripts shown in Fig 5.1.1, the adaptor-RT-PCR assay described in Chapter 4 was applied to map the 3’ ends of the truncated *Adh* mRNA generated by the UTR-9 polyA site. The RT-PCR assay shows similar results as the Northern blot: the reporters with the polyA signal at P2 and P3 produce abundant level of the expected transcripts but much lower level when located at P1 (Fig 5.1.2B). The entire shown area of the gel were evenly stained by EtBr as visualised under UV transilluminator. Because the RT reaction is adaptor-oligo(dT) primed, the results also indicate that the detected transcripts are polyadenylated. Cloning and sequencing of the PCR products show that cleavage takes place at 11-26 nt downstream the AAUAAA as expected (Fig 5.1.2C). These experiments indicate that the 3’ ends are generated by the standard cleavage and polyadenylation reaction.

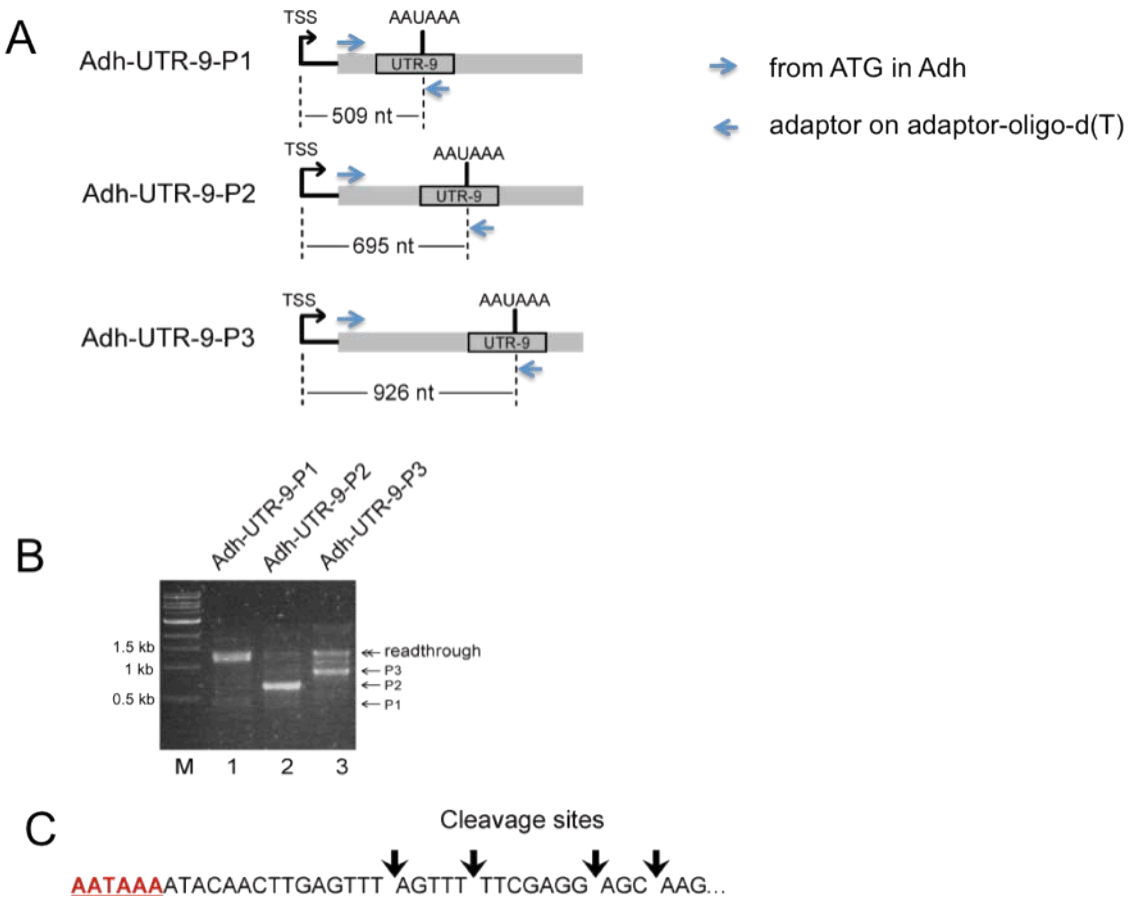
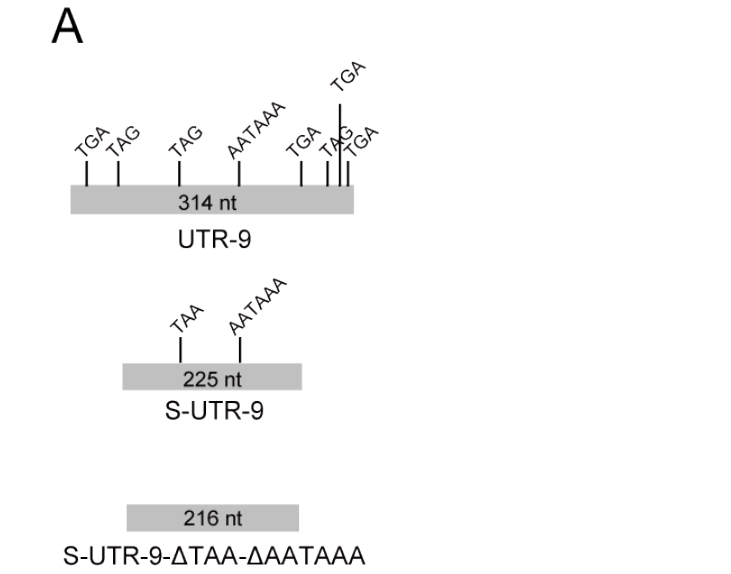
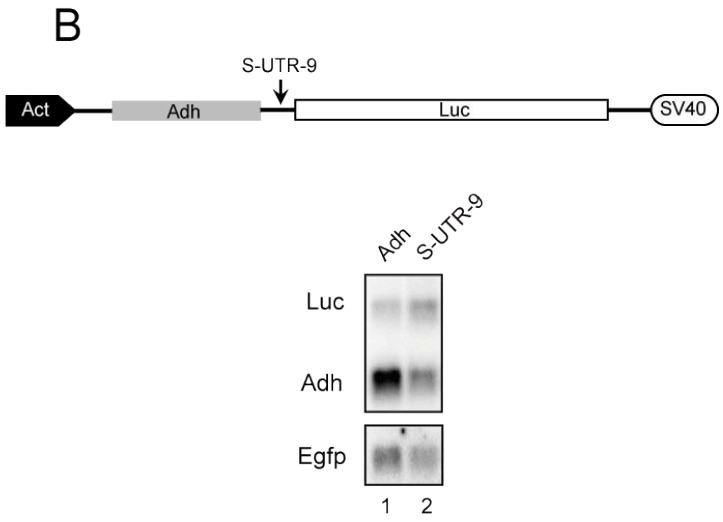


Figure 5.1.2 Characterization of the 3' end of the transcripts truncated at Adh-P1, Adh-P2 and Adh-P3. (A) Schematics showing locations of the primers used in Adaptor-RT-PCR. Same method used as in Fig 4.2.1. (B) Agarose gel showing the DNA fragments produced by the adaptor-RT-PCR assay of total-RNA extracted from cells of transfected with the indicated reporters. (C) Location of the polyA sites in the P1, P2 and P3 transcripts shown in B (based on sequencing of several clones of the P1, P2 and P3 RT-PCR fragments).

As negative controls, we thought to inhibit the polyA activity of UTR-9 by deleting the AAUAAA. A shortened variant of UTR-9 (S-UTR-9) that carries only one in-frame stop codon was produced and the S-UTR-9 showed polyA activity when inserted into the intergenic spacer of *Adh-Luc* reporter as expected (Fig 5.1.3B). The positional-dependent activity when placed at P1, P2 and P3 remained as for UTR-9 (Fig 5.1.3C, lanes 1-3). A further variant of S-UTR-9 was made with both the stop codon and the AAUAAA deleted (S-UTR-9-ΔTAA-ΔAATAAA) (Fig 5.1.3A). Deletion of the stop codon was to prevent the readthrough from being targeted by Nonsense Mediated mRNA Decay (NMD), as it would be helpful to indicate active transcription in this set of reporters. Deletion of the hexamer prevented production of the truncated *Adh* transcripts regardless of its position (Fig 5.1.3C, lanes 4-6). For reasons we do not know, deletion of the AAUAAA did not affect the level of the readthrough transcript. It appears the inserted UTR sequences caused general reduction of steady state mRNA level, which is seen throughout this chapter. But we did not further investigate this issue at this point.


A
TGA
TAG
TAG
AATAAA
TGA
TAG
TGA
TGA
314 nt
UTR-9
TAA
AATAAA
225 nt
S-UTR-9
216 nt
S-UTR-9-ΔTAA-ΔAATAAA


B
S-UTR-9
Act
Adh
Luc
SV40
Adh
S-UTR-9
Luc
Adh
Egfp
1
2

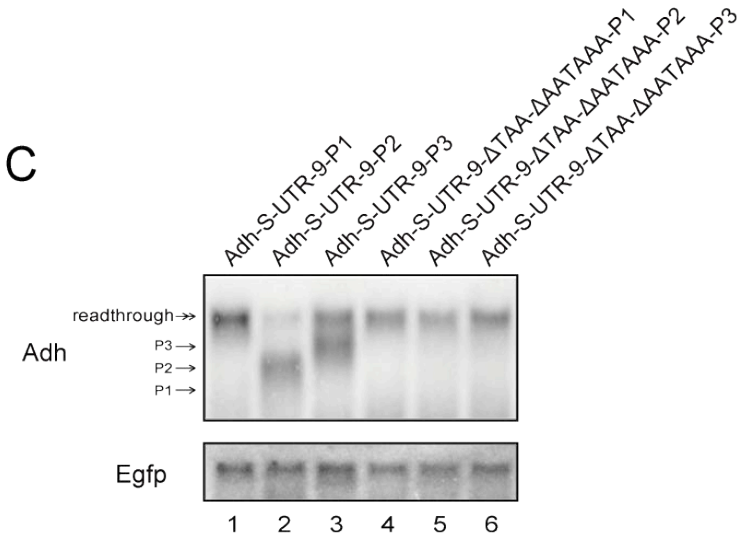

C
Adh-S-UTR-9-P1
Adh-S-UTR-9-P2
Adh-S-UTR-9-P3
Adh-S-UTR-9-ΔTAA-ΔAATAAA-P1
Adh-S-UTR-9-ΔTAA-ΔAATAAA-P2
Adh-S-UTR-9-ΔTAA-ΔAATAAA-P3
readthrough
P3
P2
P1
Adh
Egfp
1
2
3
4
5
6

Figure 5.1.3 The AAUAAA hexamer is required for 3' end processing. (A) Schematics of the UTR-9 derivatives with or without AAUAAA and in-frame stop codons. (B) Northern blot analysis of reporters with the shorter UTR-9 derivative shown in A, inserted between *Adh* and *Luc* as in Fig 3.2.2. (C) Northern blot analysis of cells transfected with reporters containing the S-UTR-9 derivative at positions P1, P2 and P3 in *Adh*. The S-UTR-9-ΔTAAΔAATAAA constructs lack all in frame stop codons and AAUAAA.

Next, to test if the position dependent effect is applicable to other polyA signals, we inserted a strong SV40 polyA signal at positions P1, P2, P3, ΔP2 and ΔP3 as above (Fig 5.1.4A). The results show that the SV40 polyA signal at P1 is impaired, producing much less transcript than at P2 or P3 (Fig 5.1.4B, lanes 1-3). Moving the SV40 signal closer to the TSS by deletion of the first 192 nt of *Adh* also reduced its activity, as in the UTR-9-based reporters (Fig 5.1.4B, lanes 4 vs 2).

Unlike the reporters with UTR-9, the SV40 polyA signal did not produce significant amount of readthrough transcript. This is probably because the SV40 polyA signal is much stronger than the UTR-9 polyA signal (when two sets of samples were compared on a same membrane), as it has been shown that stronger polyA signal induces more efficient Pol II termination (Orozco et al., 2002). But we did not examine efficiency of termination in this study. However, this inverse correlation between the strength of inserted polyA signal and usage of distal polyA signal (readthrough) is also observed in later experiments. It is also possible the readthrough transcripts with different inserted sequences have different stabilities in the cell. However, we did not investigate this issue in this thesis, but instead focused on the product of early polyA signals only.

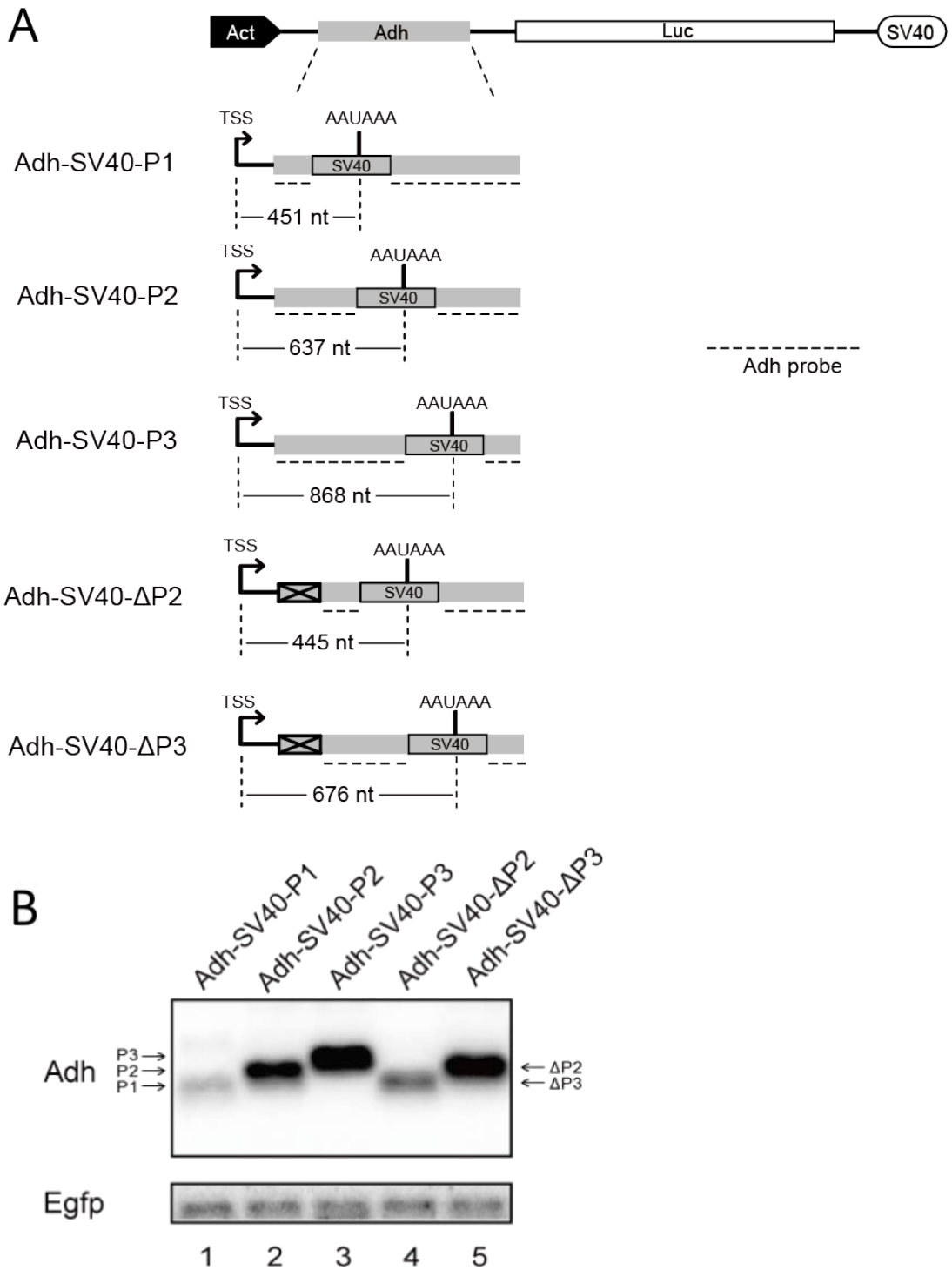

A
Act
Adh
Luc
SV40
TSS
AAUAAA
Adh-SV40-P1
SV40
451 nt
Adh-SV40-P2
637 nt
Adh probe
Adh-SV40-P3
868 nt
Adh-SV40-ΔP2
445 nt
Adh-SV40-ΔP3
676 nt
B
Adh-SV40-P1
Adh-SV40-P2
Adh-SV40-P3
Adh-SV40-ΔP2
Adh-SV40-ΔP3
Adh
P3
P2
P1
ΔP2
ΔP3
Egfp
1
2
3
4
5

Figure 5.1.4 Proximity to the TSS silences also the SV40 polyA signal in the *Adh* reporter. (A) Schematics of reporters with SV40 polyA signal inserted at different positions in the *Adh* coding region. Distance from TSS to AAUAAA is indicated below. (B) Northern blots of total RNA of S2 cells transfected with the reporters in A; probes as in Fig 3.2.2. Truncated transcripts processed at early polyA signals are indicated: P1, P2 and P3; ΔP2 and ΔP3 indicate mRNA derived by the deletion derivative lacking the initial region of *Adh*.

To further investigate the generality of the observation that polyA signals do not function when located close to the 5' end, I analyzed different reporter genes and other polyA signals. I used the *E. coli* β-Galactosidase gene (*lacZ*) with the bovine growth hormone (BGH) polyA signal inserted at three different positions, placing the AAUAAA 204, 404, and 704 nt from the TSS respectively (Fig 5.1.5A). In another set of reporters, the UTR-4 polyA signal from Chapter 3 was inserted at three different positions in the firefly luciferase gene (*Luc*), placing the AAUAAA 253, 445 and 862 nt from the TSS respectively (Fig 5.1.6A). In each set of reporters, the insertion points are labelled as P1, P2 and P3.

With the *lacZ*-based reporters, we found that BGH polyA signal produced very little mRNA at P1 whereas it produced significant amount of mRNA at P2 and P3 (Fig. 5.1.5B, lane 1-3). With the *Luc*-based reporters, the UTR-4 signal did not produce detectable amount of transcript at P1, but became highly active when located further downstream at P2 and P3 (Fig 5.1.6B, lane 1-3). It is noticed that Luc-UTR-4-P3 showed weaker polyA activity than Luc-UTR-4-P2 (Fig 5.1.6B, lane 3 vs lane 2), possibly because it is affected by the presence of the extremely strong SV40 polyA signal downstream in the plasmid. In these two sets of experiments, the probes PCR products of the entire inserts (UTR-4 and BGH polyA signal), which have identical hybridisation area for all reporters within each sets.

In summary, we tested four polyA signals (UTR-9, SV40, BGH and UTR-4) in three reporter genes (*Adh*, *lacZ* and *Luc*) in Drosophila S2 cells. Results show that, in all cases, polyA signals are silenced when located close to the TSS. The required

distance to TSS for polyA signal to be functional varies between 400 nt to 600 nt between the tested reporters, suggesting that gene specific features must determine the exact position where 3’ end processing becomes efficient. Perhaps, the differences are down to the sequence compositions for individual genes.

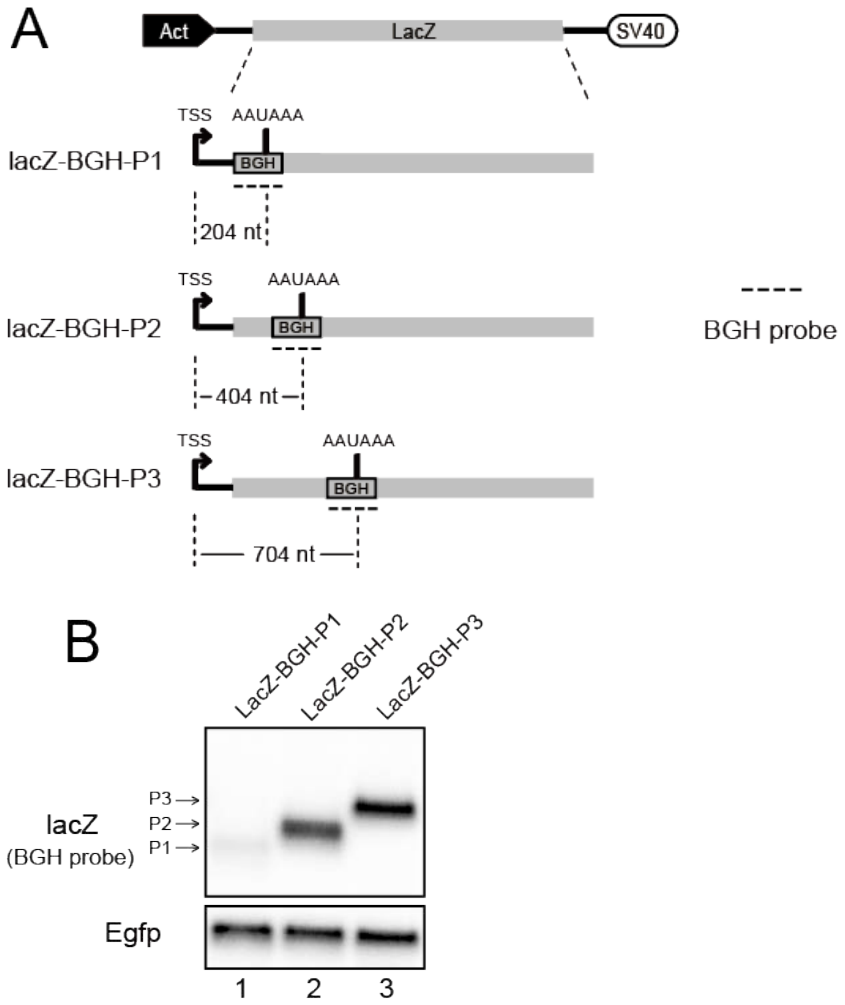


Fig 5.1.5 BGH polyA signal is silenced when positioned close to the 5' end of *lacZ*. (A) Schematics of *lacZ* gene with BGH inserted at positions P1, P2 and P3 with distances between TSS and AAUAAA indicated below. (B) Northern blots of total RNA of S2 cells transfected with reporters in A. In the LacZ panel, truncated transcripts processed at early polyA signals are indicated: P1, P2 and P3. A BGH probe recognising the entire BGH polyA signal insertion was used (for probe details see Materials and Methods).

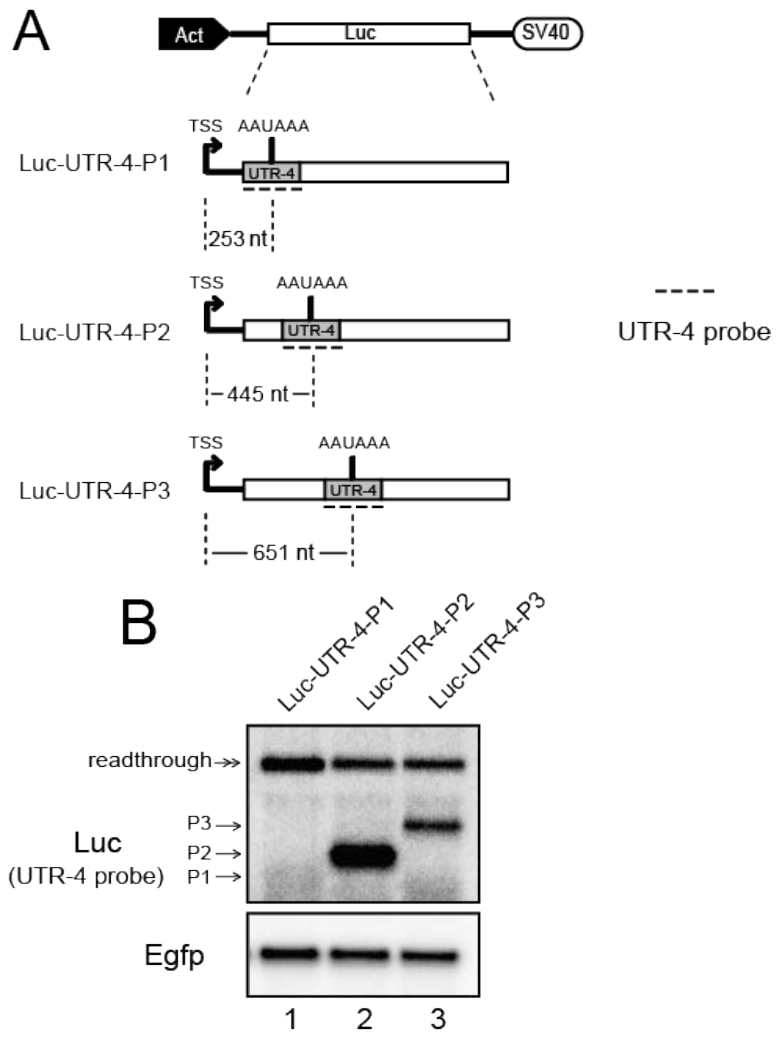


Fig 5.1.6 The UTR-4 polyA signal is silenced at 5' end of *Luc*. (A) Schematics of *Luc* gene with UTR-4 inserted at positions P1, P2 and P3. Distance between TSS and AAUAAA is indicated below. (B) Northern blots of total RNA of S2 cells transfected with reporters in A. A UTR-4 probe that recognises the entire UTR-4 insertion is used for the Luc panel: top band (doubled arrowed line) is the readthrough mRNA processed at the downstream SV40 polyA signal in the plasmid. Truncated transcripts processed at early polyA signals are indicated: P1, P2 and P3.

5.2 PolyA signals close to the transcription start sites are silent also in human cells.

To assess whether the position on the pre-mRNA affects polyA signal recognition in other organisms, we made similar constructs driven by the CMV promoter and tested them in human HEK 293T cells (Fig 5.2.1A). Similar to the Drosophila constructs; the SV40 polyA signal is located at positions P1, P2 and P3 in *Adh*. The distances between TSS and AAUAAA are 451 nt, 637 nt and 868 nt, respectively (Fig 5.2.1A). In these experiments I used a reporter coding for a truncated version of *Luc* – because the SV40 polyA signal generated no *Adh-Luc* dicistronic transcripts, shortening of the *Luc* gene should not affect upstream transcription of the reporter.

The reporters were transiently transfected in 293T cells and the total RNA was isolated 24 hours post transfection. Northern blot in Fig 5.2.1B shows the same positional effect as seen in Drosophila cells: when placed at P1, the SV40 polyA signal showed very low activity, while at P2 and P3 it becomes highly active. These results suggest that polyA signal is inactivated when located close to TSS also in human cells.

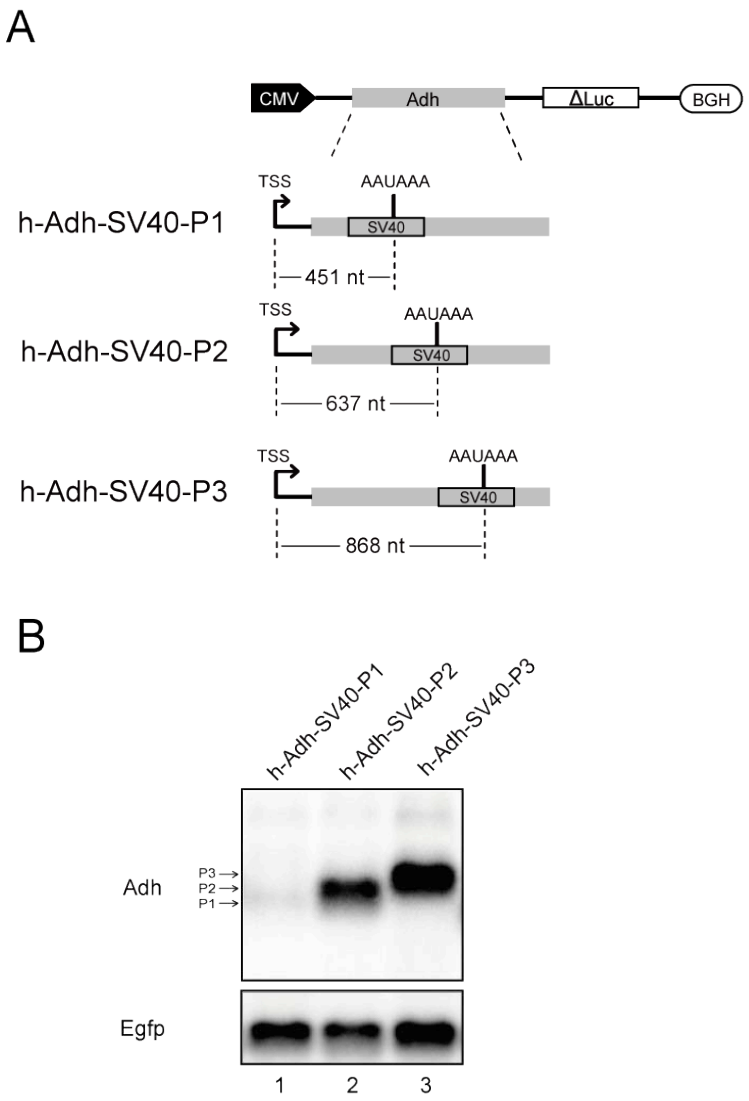


Figure 5.2.1 PolyA signal close to TSS is also silenced in human cells. (A) Schematics of human reporters with SV40 polyA signal inserted at different positions in the *Adh* coding region. Distance from TSS to AAUAAA is indicated below. (B) Northern blots of total RNA of 293T cells transfected with the reporters in A; probes as in Fig 3.2.2. Truncated transcripts processed at early polyA signals are indicated as P1, P2 and P3.

Chapter 6 Transcripts processed at early polyA sites are not exosome substrates

In the following two chapters, I will discuss the results from experiments aimed at understanding the mechanism/s involved in silencing early polyA signals. Because Northern blots detect steady state mRNA, it is possible that in the previous experiments the change in transcripts level are due to differences in mRNA stability. For example it can be argued that transcripts are processed at 5' proximal positions but the resulting transcript are rapidly degraded. Therefore, the question arises whether the early polyA signal is never able to process a transcript or could the transcript be targeted by co-transcriptional quality control mechanism for rapid degradation.

We firstly asked whether the reason for the low steady state accumulation of mRNA polyadenylated at early sites might simply be that such transcripts are unstable; they might be efficiently produced but rapidly degraded by some mRNA quality control pathway. Many recent studies have indicated that aberrant transcripts are rapidly degraded by the nuclear exosome (Houseley and Tollervey, 2009). Therefore, it was interesting to assess whether the exosome targets also prematurely polyadenylated transcripts. We depleted nine exosome subunits (Rrp6, Dis3/Rrp44, Rrp41/Ski6, Mtr3, Rrp40, Rrp46, Rrp42, Csl4 and Rrp4) in S2 cells by RNAi and then transfected them with some of the *Adh-Luc* reporters described above. We also targeted Trf4-1, Trf4-2 as they are the Drosophila homologs of the non-canonical

polyA polymerase Trf4 (Nakamura et al., 2008). Key factors in the polyA complex CPSF-160, CstF-64 and Pcf11 were also included for comparison.

Initially, we carried out an RNAi screen using the Adh-SV40-P1 reporter and found that none of RNAi depletions increased the level of the P1 transcripts processed at the 5' proximal P1 site (Fig 6.1). As expected, depletion of polyA factors CPSF, CstF and Pcf11 reduced the transcript level (Fig 6.1B Lane 3, 13, 14 and 15). However, depletion of Rrp6 showed no recovery of the P1 transcript, instead the levels of the transcripts were further reduced (Fig 6.1B). These Northern blots are exposed for longer time (72 hours) to achieve clear bands for both P1 and readthrough transcripts. In these experiments RNAi also affects the Egfp transcript, however, similar results were observed in many experimental repeats and the constant level of the 18S rRNA indicate no significant variability in the assay (although we could not rule out that Rrp6 knockdown does not affect 18S rRNA production). In summary, these results suggest that the truncated transcripts produced by early 3' processing are probably not subjected to Rrp6/exosome mediated degradation.

Although exosome was the likely degrader for early polyA signal products, we could not rule out degradation possibility altogether. A more thorough screen of other degradation factors may provide a clearer picture.

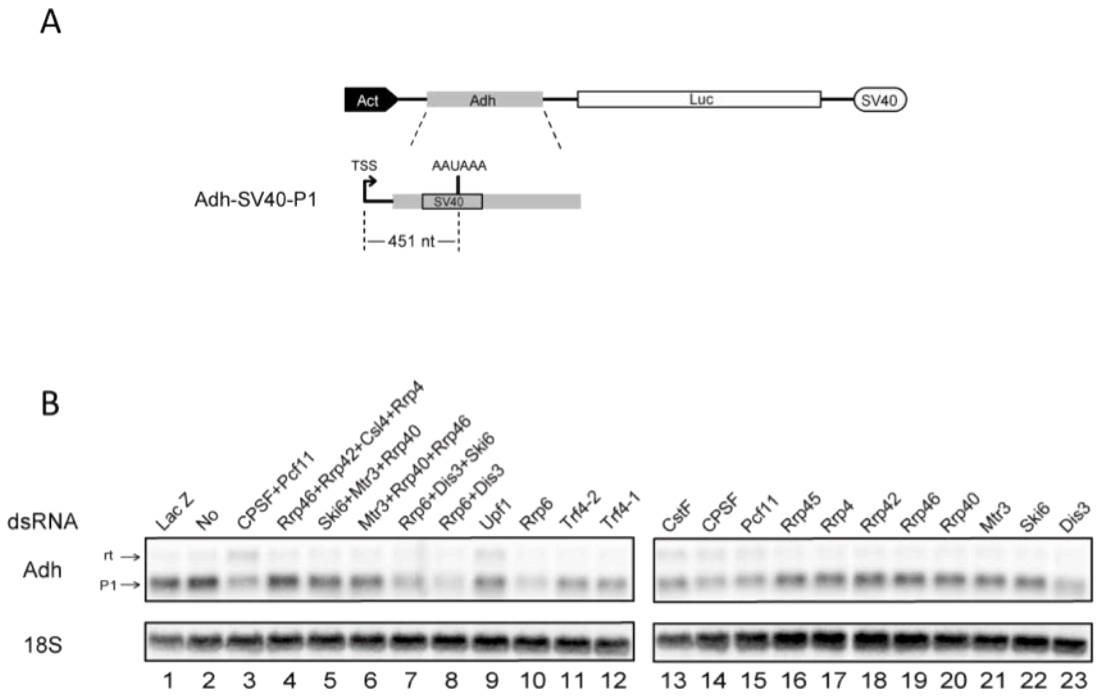


Fig 6.1 RNAi screen for putative factors regulating early polyA signals. (A) Adh-SV40-P1 was used in this screen. (B) Northern bolts of total RNA of S2 cells transfected with Adh-SV40-P1 and the labelled dsRNA. The 18S rRNA was used as loading control. The Adh probe is same as in Fig 3.2.2. The 18S rRNA probe is as described (Chan et al., 2001). The Adh labelling was exposed for 72 hours to visualise both P1 and rt bands.

To investigate further the observation that Rrp6 depletion leads to less mRNA rather than more, polyA signals at other positions were tested with the same Rrp6 depletion. I found that transcripts produced by Adh-SV40-P1, Adh-SV40-P2, and Adh-SV40-P3 are also reduced upon Rrp6 depletion (Fig 6.2A Lane 1-3 vs Lane 4-6); the off-target dsRNA against *lacZ* showed identical results as the samples with no dsRNA treatment (Fig 6.2A Lane 1-3 vs Lane 7-9).

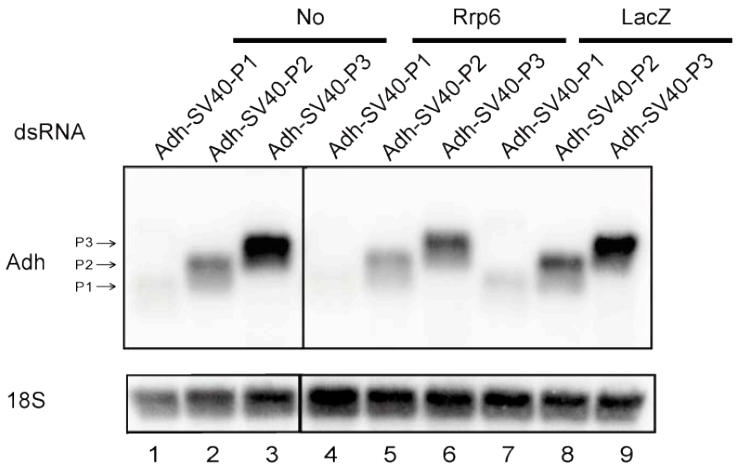
To check if Rrp6 depletion reduces the mRNA levels of the other reporters, the original *Adh-Luc* dicistronic reporter with *Adh* polyA signal was tested. Northern blots show that Rrp6 depletion reduced the level of *Adh* mRNA to 70%, similar to that of CPSF or Pcf11 depletion (75% for CPSF and 92% for Pcf11) (Fig 6.2C). Furthermore, double depletions of CPSF + Rrp6 and Pcf11 + Rrp6 further reduced the level of *Adh* mRNA to 46% and 48% respectively (Fig 6.2C). In addition, the co-transfected EGFP also showed similar reduction pattern. Egfp mRNA was reduced to 76% by Rrp6 depletion, to 75% by CPSF depletion and to 90% by Pcf11 depletion. The double depletions of CPSF + Rrp6 and Pcf11 + Rrp6 further reduced Egfp mRNA to 51% and 50% respectively (Fig 6.2 C, Panel Egfp). Band intensities were normalised against 18S rRNA before comparing with the sample treated with no dsRNA (which is set at 100%).

Overall, it seems that, similar to CPSF and Pcf11, Rrp6 depletion affect general transcription. The mechanism for this unexpected finding is not further studied. Using the RNAi methods, we did not find a degradation related role for the exosome in 3’ end processing. However, RNAi cannot completely remove the level or Rrp6

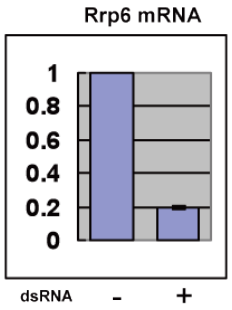
to zero, as the mRNA level is maintained at ~20%. The remaining portion of Rrp6, and other exosome subunits, may still be able to carry out certain level of degradation.

With limited testing of exosome related degradation factors, it seems the early polyA signals might be silenced by pathways other than degradation. The transcripts processed at Adh-UTR-9-P1, -P2, and -P3 was sequenced and did not show evidence of undergoing cryptic splicing as the transcripts were sequenced in Fig 5.1.2. Therefore, the mechanism that silences these promoter proximal polyA signals seems different than the silencing of HIV 5' LTR polyA signal as discussed in Introduction. For further comparisons, see Discussion.

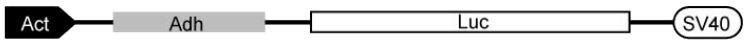
A


No
Rrp6
LacZ
Adh-SV40-P1
Adh-SV40-P2
Adh-SV40-P3
Adh-SV40-P1
Adh-SV40-P2
Adh-SV40-P3
Adh-SV40-P1
Adh-SV40-P2
Adh-SV40-P3
dsRNA
Adh
P3
P2
P1
18S
1 2 3 4 5 6 7 8 9

B


Rrp6 mRNA
1
0.8
0.6
0.4
0.2
0
dsRNA - +

C


Act
Adh
Luc
SV40

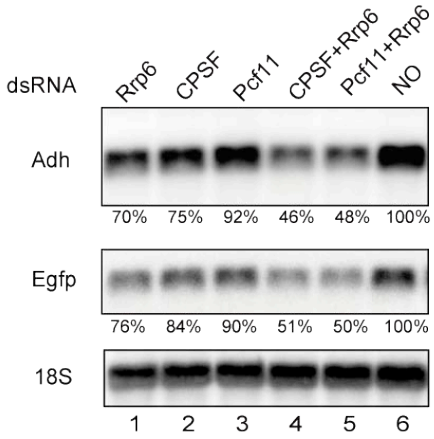

dsRNA
Rrp6
CPSF
Pcf11
CPSF+Rrp6
Pcf11+Rrp6
NO
Adh
70% 75% 92% 46% 48% 100%
Egfp
76% 84% 90% 51% 50% 100%
18S
1 2 3 4 5 6

Fig 6.2 Rrp6 depletion does not recover truncated mRNAs. (A) Northern blot analysis of transcripts in Rrp6-depleted S2 cells transfected with Adh-SV40-P1, Adh-SV40-P2 and Adh-SV40-P3. Mock-experiments without dsRNA incubation (lanes 1-3) or with off-target dsRNA against *lacZ* (lanes 7-9) are shown. Adh probe as in Fig 3.2.2; 18S rRNA probe as in Fig 6.1. (B) Real time RT-PCR measuring level of Rrp6 mRNA depletion relative to control cells not treated by dsRNA. Level of Rrp6 mRNA is normalised by that of Rpl32. (C) Northern blots to total RNA from S2 cells transfected with *Adh-Luc* reporter (as Fig3.2.2, with *Adh* polyA signal in the intergenic region); the cells were treated with dsRNA against Rrp6, CPSF-160, Pcf11, CPSF+Rrp6 and Pcf11+Rrp6. Quantifications are band intensities relative to that in the control not treated by dsRNA; values were standardised by the relative intensity of the 18S rRNA band. The lane with no dsRNA was set as 100%.

Chapter 7 Efficient 3' end processing requires high levels of CTD Ser2P and Pcf11.

We then looked at whether dynamics of the transcription elongation complex and assembly of the polyA complex could affect early polyA signals. It is feasible that the Pol II elongation complex being incompetent of processing 5' proximal polyA signals is because it lacks key components that are only available at later stage of elongation. As reviewed in the introduction, a key change between early and late transcription is the gradual increase in phosphorylation of the Ser2 residues on the Pol II CTD (Ser2P), which is required for 3' end processing. In addition, many key polyA factors interact with the Ser2 phosphorylated CTD (Licatalosi et al., 2002; Zhang and Gilmour, 2006). This may explain why polyA signals are only active when placed further downstream.

7.1 Early polyA signals are more sensitive to Pcf11 depletion.

Early polyA signals might be skipped because key processing factors are inefficiently recruited to short nascent transcripts. We reasoned that depletion of some 3'end processing proteins might affect earlier sites more than later ones. We tested this possibility in S2 cells by depleting CPSF-160, CstF-64 and Pcf11 by RNAi. Depletion of CPSF-160 and CstF-64 caused a general reduction in the transcript levels regardless of the positions of the polyA signals (Fig 6.2C and data not shown). Instead, depletion of Pcf11 appears to affect comparably more the early

polyA signals: in cells partially depleted of Pcf11, the ratios of the truncated transcripts to the readthrough transcripts were clearly reduced for both P2 and P3 mRNAs (Fig 7.1.1, Quantitations in B). Cells treated with dsRNA against *lacZ* gave identical results as those not treated with dsRNA, confirming the specificity of the RNAi knockdown (Fig 7.1.1A). Although weak, the general down regulation of mRNA by Pcf11 depletion also made the Egfp normalization inapplicable between Pcf11 dsRNA treated samples and the no dsRNA samples. As before, the level of 18S rRNA was used to indicate loading variation (Fig 7.1.1A). The observation that Pcf11 depletion affects relatively early polyA signals (Adh-P2 and Adh-P3) much more than the read-through suggests that Pcf11, required for efficient 3' end processing, is progressively recruited to the transcription complex during Pol II elongation. Supporting this notion, the progressive recruitment of Pcf11 was presented in a study using dsRNA knock-down and ChIP measuring Pcf11 level in S2 cells (Zhang and Gilmour, 2006).

However, the limited effect of RNAi depletion could mean the results in Fig 7.1.1A are unspecific or indirect. Unfortunately, we did not successfully acquire the dPcf11 anti-body to measure protein level in our RNAi experiments. Overexpression of Pcf11 would further support specificity of Pcf11's function, but attempts of making Pcf11 overexpression plasmid were not successful.

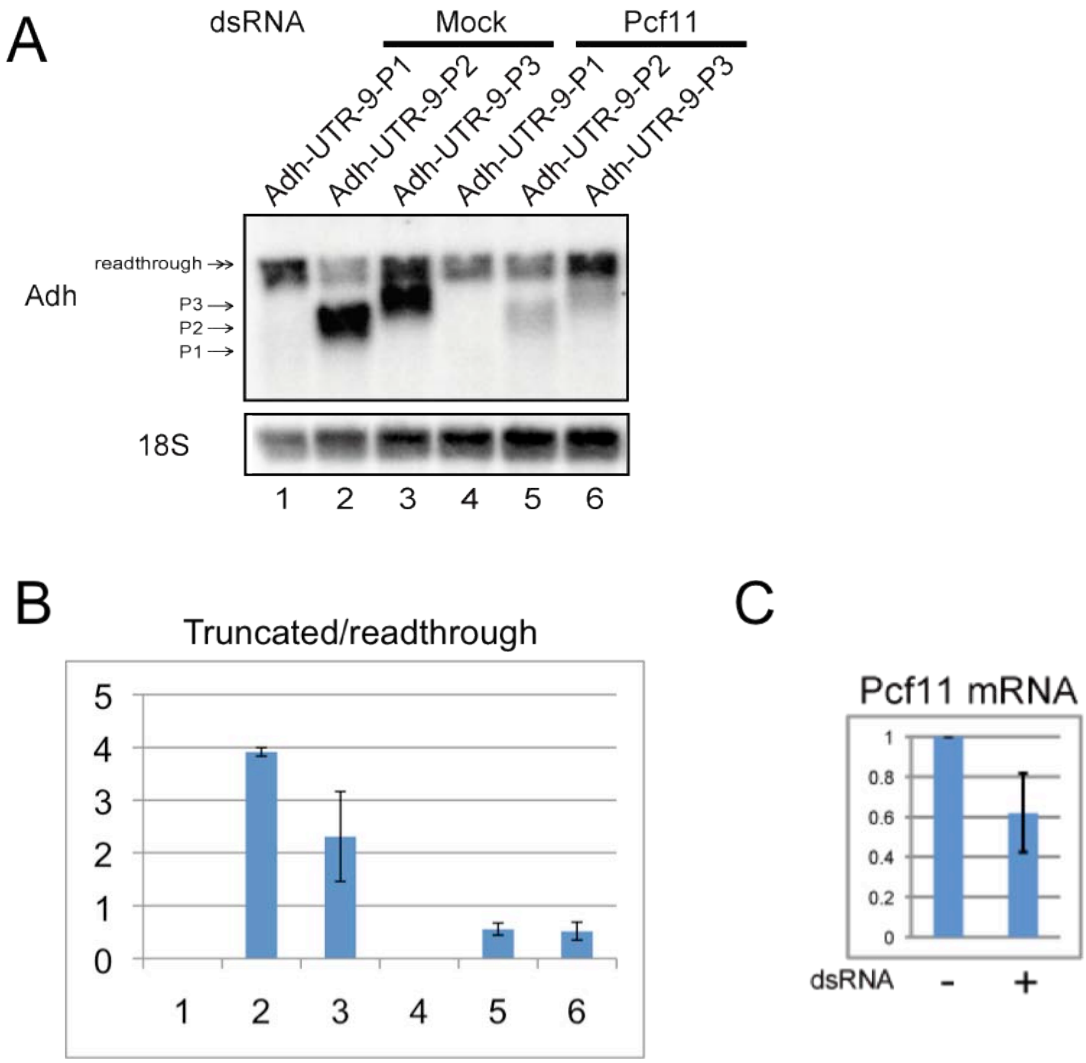


Fig 7.1.1 Early polyA signals are more sensitive to Pcf11 depletion. (A) Northern blot analysis of total-RNA extracted from Pcf11 depleted S2 cells, transfected with the Adh-UTR-9-P1, Adh-UTR-9-P2 and Adh-UTR-9-P3. Mock-experiments refer to cells without dsRNA treatments. Adh probe and 18S rRNA probes are as described before. (B) Ratios of the truncated/readthrough bands intensities in A, error bars based on two independent experiments. (C) Real time RT-PCR quantification of Pcf11 mRNA in cells treated with dsRNA relative to mock; Pcf11 mRNA levels are normalised to that of Rpl32.

7.2 Depletion of the CTD phosphatase Fcp1 enhances activity of early polyA signals.

The observation that Pcf11 depletion downregulates earlier polyA signals more than later ones could be due to inefficient recruitment of this essential processing factor at the early stage of transcription. As reviewed in the Introduction, this is likely to be linked with low Ser2 phosphorylation on the Pol II CTD. Therefore, we sought to experimentally induce changes in CTD phosphorylation. Firstly we used dsRNA to knockdown of the CTD kinase P-TEFb (Cdk9 and CycT) in S2 cell transfected with the reporters carrying polyA signals at different positions (Fig 7.2.1). The depletions resulted in reduction of the transcripts. However, the reduction was seen for all positions: P1, P2, P3 and even the readthrough (Fig 7.2.1A lanes 4-9). The most apparent reduction is observed in cells depleted of both Cdk9 and CycT (Fig 7.2.1A, lanes 10-12). These experiments indicate that P-TEFb depletion probably impairs transcription in general.

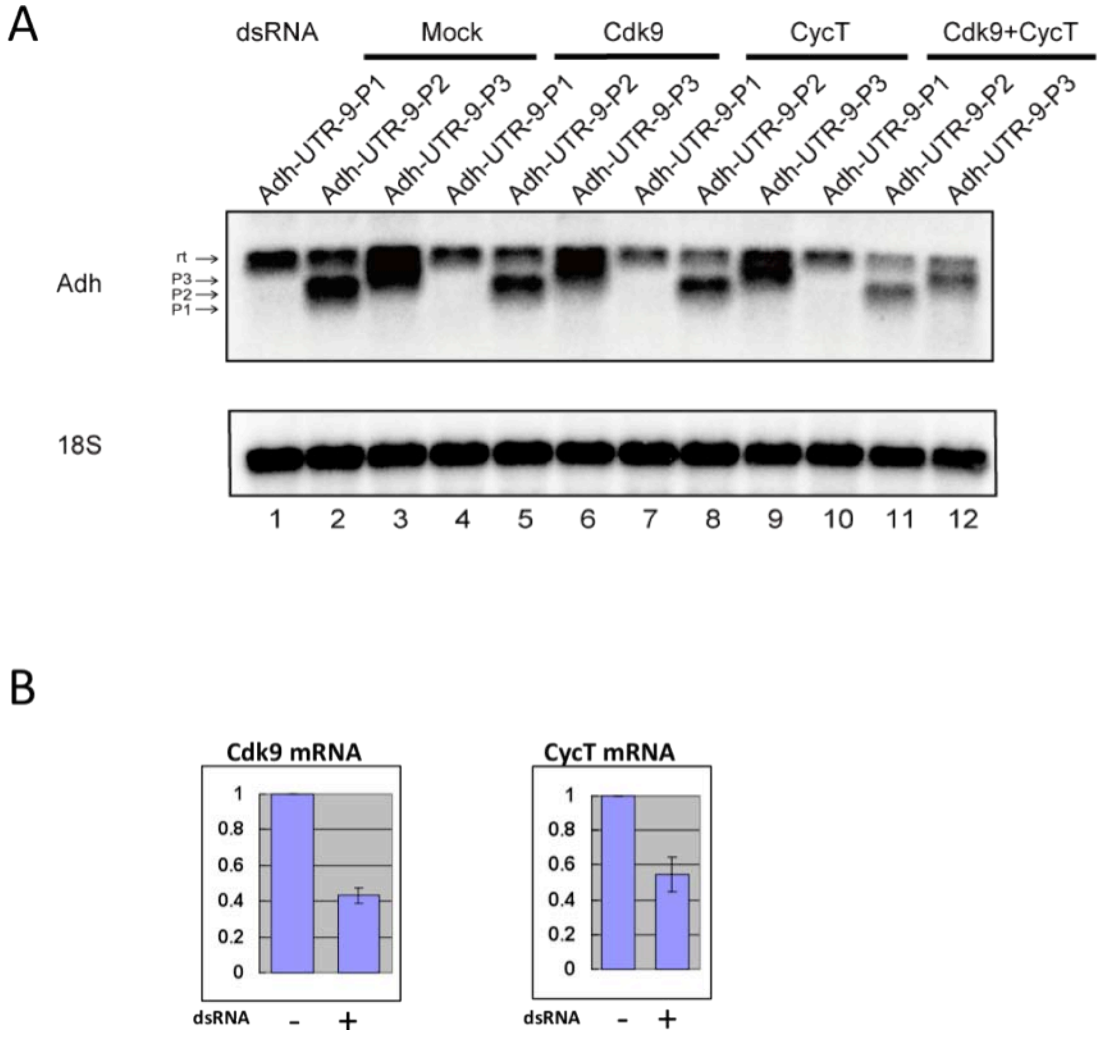


Fig 7.2.1 Depletion of P-TEFb causes a general mRNA reduction. (A) Northern blot analysis of total-RNA extracted from Cdk9 and CycT depleted S2 cells, transfected with the Adh-UTR-9-P1, Adh-UTR-9-P2 and Adh-UTR-9-P3. Mock-experiments refer to cells without dsRNA treatments (lanes 1-3). Adh and 18S rRNA probes are as described previously. (B) Real time RT-PCR quantification of Cdk9 and CycT mRNA depletion as in A.

In a second experiment aimed at changing CTD phosphorylation level, I depleted the CTD phosphatase Fcp1 by RNAi. Fcp1 has the opposite effect of P-TEFb and reduces Ser2P (Cho et al., 2001). Its homolog in Drosophila has recently been characterised as essential for Drosophila throughout developmental stages because mis-regulation of Fcp1 results in lethality (Tombacz et al., 2009). Depleting Fcp1 in S2 cells, however, did not noticeably affect the doubling time of the cells during my experiments and therefore I could test its function by RNAi. Northern blot analysis shows that Fcp1 depletion led to moderately increased polyA activity of the UTR-9 at P2 and P3 in *Adh* (Fig 7.2.2). However, contrary to expectation, depletion of Fcp1 did not recover the level of transcripts at Adh-P1.

In experiments with other reporters, mRNA level of Luc-UTR-4-P1 and Luc-UTR-4-P2 were increased upon depletion of Fcp1 whereas Luc-UTR-4-P3 was unchanged, suggesting the Fcp1 depletion affect relatively early polyA signals more than later ones (Fig 7.2.3). This is consistent with that Fcp1 depletion did not affect the readthrough transcripts in Fig 7.2.2. A possible explanation would be that CTD Ser2P level at later stage of transcription is sufficient to carry out effective 3' end processing regardless of the increase of Ser2P caused by Fcp1 depletion. This agrees with the observation that Ser2P gradually increases until 600-100 nt from TSS and then remain at a relatively constant high level until the 3' end (Kim et al., 2010; Mayer et al., 2010). Together, these data suggest that higher level of Ser2P, caused by Fcp1 depletion, leads to more efficient 3' end processing.

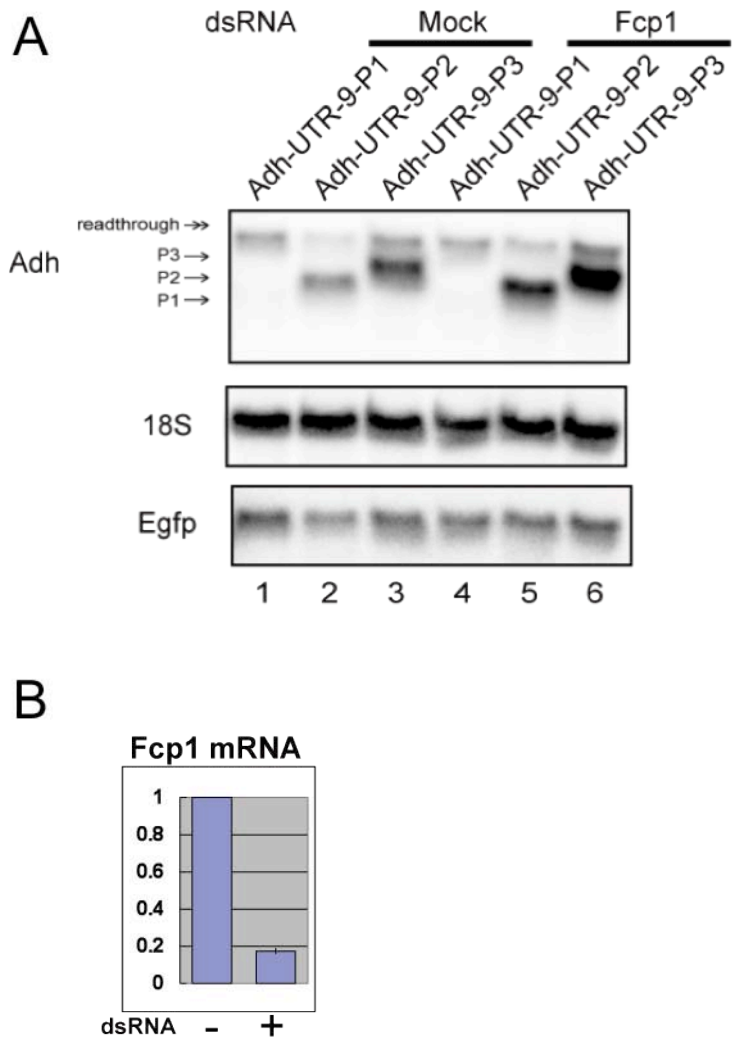


Fig 7.2.2 Fcp1 depletion enhances activity of early polyA signals. (A) Northern blot analysis of total-RNA from Fcp1-depleted S2 cells transfected with Adh-UTR-9-P1, Adh-UTR-9-P2 and Adh-UTR-9-P3. (B) Real time RT-PCR quantification of Fcp1 mRNA in cells treated with dsRNA relative to mock. Fcp1 mRNA levels are normalised to that of Rpl32.

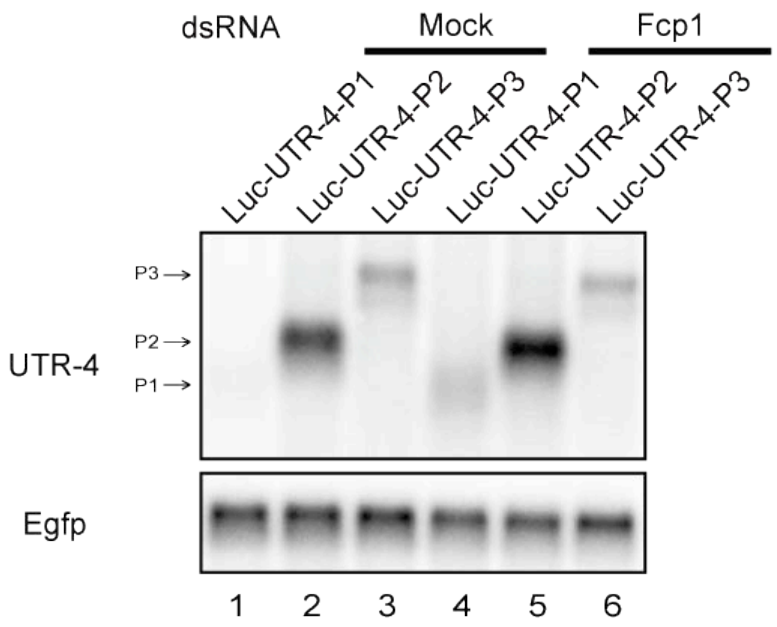


Fig 7.2.3 Fcp1 depletion enhances activity of early polyA signals in Luc-based reporters. Northern blot analysis of total-RNA from Fcp1-depleted S2 cells transfected with Luc-UTR-4-P1, Luc-UTR-4-P2 and Luc-UTR-4-P3. Same RNAi depletion procedure as in Fig 7.2.2.

In other attempts to alter CTD phosphorylation, I depleted other proteins by RNAi. One of the factors is Rtr1, which is a recently identified phosphatase of CTD Ser5P (Kim et al., 2009; Mosley et al., 2009). Another is Brd4, which is essential for P-TEFb recruitment (Hargreaves et al., 2009; Jang et al., 2005; Yang et al., 2005). Brd4 Drosophila homolog is encoded *fs(1)h* (Chang et al., 2007). Preliminary results from depleting these factors suggest no noticeable change to polyA signals at any position (see Appendix 1). Probably manipulating single factors by RNAi is insufficient to significantly affect 3' processing of transiently transfected reporters.

Due to limited time and resources, validations of RNAi were only carried out to experiments where we could detect apparent effects on Adh-P1, -P2 and -P3 transcripts. Real-time RT-PCRs shown above were results of two to four randomly selected samples. The levels of depletions were consistent between experimental repeats, but showed considerable variation between different target genes. Examples of RT-PCR validations are shown in Appendix 6. Different time of dsRNA incubation (one day, two days or three days) was tested but did not result detectable change. This could be the natural feature of the dsRNA method. To further investigate the implications from above RNAi experiments, other approaches would be required to ensure more drastic change to the level of the factors involved.

In summary, the results in this chapter suggest that 5' proximal polyA signals are silent because of low CTD Ser2P, which results in inefficient recruitment of polyA factors (such as Pcf11) at early stages of transcription elongation.

Chapter 8 Discussion

8.1 5'UTR polyA signals are frequent in the genome

As reviewed in the Introduction, the composite sequence that makes the polyA signal is the key determinant in the specification of the polyA site, and it is generally expected that such sequences should only be found at the 3' end of genes. Contrary to this view, here we have reported that bioinformatic programmes also predict the presence of polyA signals in the 5' UTR of 24% of Drosophila genes. For a subset of these sequences, we have experimentally verified their functionalities as polyA signals when they are placed at the 3' end of reporter genes: one was even more efficient than the endogenous polyA signal of *Adh,* one of the most highly expressed genes in Drosophila. The number of transcripts with putative polyA signals in their 5' UTR is probably an underestimate, as the programs we used only predicted about half of the known polyA signals in 3'UTRs.

It was documented more than 20 years ago that a functional polyA signal leads to Pol II termination (Connelly and Manley, 1988; Whitelaw and Proudfoot, 1986). While the mechanisms of termination are being gradually revealed, the recognition of functional polyA signal remains essential for termination (Richard and Manley, 2009; West et al., 2008). Therefore, finding polyA signals in 5' UTRs raise the question of why such sequences are allowed to evolve at the beginning of genes where they could potentially interfere with transcription. Several possible

implications on gene expression of having promoter proximal polyA signals are discussed below.

8.2 The polyA machinery does not produce stable mRNA at 5' proximal polyA signals

For most of the 5' UTR polyA signals we have assayed, there is currently no evidence that they are used in flies; a few 5'UTR signals might be used but can be detected only by nested PCR, suggesting that they are rarely recognised or subjected to rapid degradation. This led to the proposal that 5' UTR polyA signals are unproductive in the endogenous genes because they are too close to the 5' end. Indeed, using several different reporter genes in Drosophila and human cells, we found that the 5' UTR sequences, as well as standard polyA signals, become silent when located close to the 5' end of reporter genes. The distance at which the signals are silenced varies between reporter genes (~500 nt from TSS in *Adh* based reporters, ~200-250 nt in *lacZ* and *Luc* based reporters). This is probably because the exact phosphorylation rates of CTD Ser2 are gene specific (Ahn et al., 2004; Kim et al., 2010).

One obvious possibility is that 5' signals are skipped because the transcription complex is not yet loaded with sufficient polyA factors at the early stage of Pol II elongation (Licatalosi et al., 2002; Zhang and Gilmour, 2006). Although some polyA factors are recruited to the transcription complex as early as initiation, many key polyA factors are more concentrated at the 3' end of genes (Dantonel et al.,

1997; Glover-Cutter et al., 2007; Kim et al., 2004a). Probably the recruitment of key processing factors, such as Pcf11 (Licatalosi et al., 2002; Zhang and Gilmour, 2006), remains inefficient until Ser2 in the CTD of Pol II becomes hyperphosphorylated (Buratowski, 2009). Our observation that depletion of the processing factor Pcf11 downregulates early polyA signals more than later ones supports this model. Furthermore, depletion of the CTD phosphatase Fcp1 increased the level of shorter transcripts more than longer ones, indicating that skipping of early polyA signals depends on low Ser2 phosphorylation. However, depletion of Fcp1 only moderately enhanced the use of the most 5' proximal polyA signal in our *Luc* based reporters and did not affect the most 5' proximal polyA signal in the *Adh* based reporters. Perhaps, the 5' proximal polyA signals are not affected by Fcp1 depletion because Ser2 is not yet hyperphosphorylated when Pol II transcribes the early signal (Buratowski, 2009). In agreement with this view, Fcp1 mutants showed increased CTD Ser2P in middle and later sections of coding regions (800+ nt from TSS) but not promoter proximal regions in yeast (Cho et al., 2001). In summary, transcription elongation appears to serve as an activation mechanism that licences polyA signals.

The possibility that promoter proximal polyA signals might be silent has not been systematically assessed for cellular genes, yet, many studies have previously reported that proximity to the promoter can silence polyA signals in retroviral pre-mRNAs (Wahle, 1995). Studies with the HIV-1 provirus show that U1 snRNP binds a 5' splice site immediately downstream of the 5' LTR polyA signal and prevents its usage; the study concluded that it is the presence of the 5' splice site

rather than the physical proximity to the promoter which inhibits the early polyA signal (Ashe et al., 1995; Ashe et al., 1997). In HIV-1 the 5' LTR polyA signal is 254 nt downstream of the TSS; our results would predict that, at this distance, the polyA signals are intrinsically silent. However, the observation that non-retroviral polyA signals are active when replacing the original 5' LTR polyA signal in HIV-1 also seems to contrast with our prediction (Weichs an der Glon et al., 1991). However, HIV-1 transcription requires the viral protein Tat (Zhu et al., 1997). Tat directly interacts with P-TEFb and stimulates its CTD kinase activity whereas, for other cellular genes, P-TEFb is recruited via Brd4 (Tahirov et al., 2010; Yang et al., 2005; Zhou et al., 2000; Zhu et al., 1997). These studies suggest that the polyA signal within HIV 5' LTR is not intrinsically silent because the rapid Tat-dependent Ser2 CTD hyperphosphorylation moves forward recruitment of 3' processing factors. However, this issue is further complicated by the observation that minigene constructs driven by CMV promoter, which does not require Tat to induce transcription, could activate the HIV polyA signal located less than 100 nt from the TSS (Ashe et al., 2000). To fully address why in this case promoter proximal polyA signal can be functional, more thorough examination of the early transcription complex and the assembly of the polyA complex would be required.

In summary, against the assumption that the sequence of polyA signal alone is sufficient to define polyA signals, our data clearly suggest that in-vivo an important second determinant is the stage at which the polyA sequences emerge from the Pol II transcription complex. CTD Ser2 phosphorylation plays a particularly important

role in transforming the transcription complex to a ‘polyA ready’ stage. To trigger 3’ end processing, both the polyA signal and the ‘CTD signal’ are required (Fig 8.2.1). This mechanism is probably conserved in eukaryotes: standard polyA signals placed near the 5’ end also become silent in *S. cerevisiae* (Domenico Libri, CNRS Gif-sur-Yvette, personal communication).

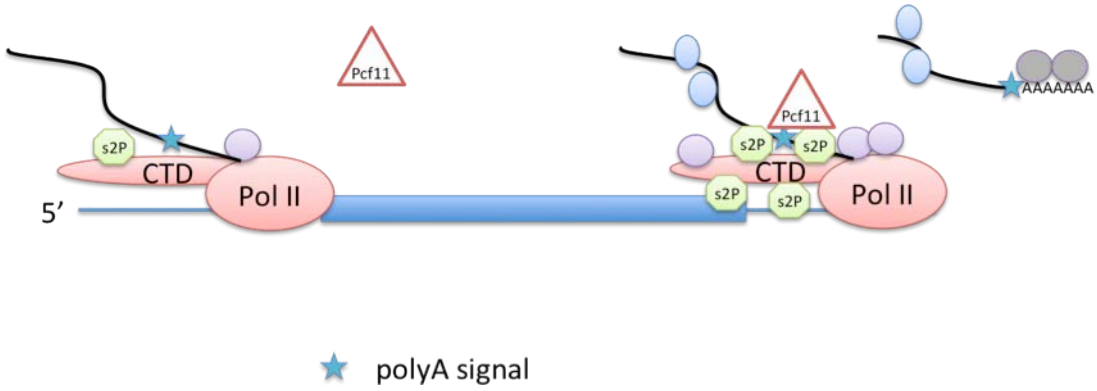


Fig 8.2.1 Model: Activation of polyA signals requires high level of CTD Ser2P. At the early stage of Pol II elongation, the CTD Ser2P level is low, hence the key polyA factor Pcf11 is not recruited; therefore the polyA signal (represented by a star) cannot be recognised and processed. Instead, at the later elongation stage, a high level of Ser2P enhances Pcf11 recruitment, allowing efficient 3' end processing as soon as the polyA signal emerges from Pol II.

8.3 Promoter proximal pausing and 5’ UTR polyA signals

It remains to be investigated whether 5’ UTR PolyA signals are involved in promoter-proximal Pol II pausing (Buratowski, 2008). Shortly after transcription elongation starts, polymerases are frequently found paused almost immediately downstream of the promoter (Price, 2008). In Drosophila, this phenomenon is regulated by two protein factors that inhibit transcription elongation: negative elongation factors (NELF) and DRB sensitivity-inducing factor (DSIF) (Wang et al., 2007; Wu et al., 2003). Pol II pausing can be overcome by the recruitment of P-TEFb, which triggers the transition from pausing to productive elongation (Peterlin and Price, 2006). The exact mechanism controlling this transition is unclear, it is possible that 5’-proximal polyA signals may enhance pausing. It has been demonstrated that the sequence of AAUAAA alone can cause Pol II pausing and inhibition of transcription (Nag et al., 2006). Also, paused polymerases at the 5’ end of genes are physically associated with CPSF and CstF, implying the possible link between pausing and the polyA complex (Glover-Cutter et al., 2007). Furthermore, paused Pol II are found at up to ~400 nt downstream of TSS on 30% of human genes (Core et al., 2008). This region is approximately the same region within which polyA signals appear to be silent in our system. Promoter proximal pausing is also found to be a general feature in Drosophila (Zeitlinger et al., 2007). Notably, the nucleotide composition (melting temperature, Tm) of the initially transcribed sequence appears to play an important role in Pol II pausing in Drosophila: the

primary region of Pol II pausing corresponded with a peak of Tm. This peak of Tm was followed by a decline, which would serve to progressively destabilize the elongation complex (Nechaev et al., 2010). Future studies should investigate whether there is a link between the presence of 5' polyA signals and paused polymerase sites.

8.4 Role of the exosome in transcription

In *S. cerevisiae*, it has been reported that the phosphorylation state of the CTD (high Ser5P and low Ser2P) at early stages of transcription prevents polyA complex-dependent termination and instead favours Nrd1-dependent termination that generates cryptic unstable transcripts, which are rapidly degraded by the nuclear exosome, Rrp6 (Gudipati et al., 2008; Vasiljeva et al., 2008). Although a similar Nrd1 induced degradation mechanism has not been identified in Drosophila and other higher eukaryotes, a study of human Pol II has shown that a fraction of the transcripts generated by Pol II paused at promoter regions are subjected to exosome degradation as knockdown of hRrp40 resulted in a relative 1.5 fold stabilisation of the transcripts in human cells (Preker et al., 2008). Although our results show that very early polyA signals (at P1 in *Adh*, *lacZ* and *Luc*) are very weakly used, it is feasible that the truncated transcripts processed at early polyA sites are at low abundance simply because they are rapidly digested by the exosome. However, dsRNA knockdown of exosome subunits did not increase the mRNA processed at early polyA signals in our experiments. On the contrary, Rrp6 depletion resulted in a

lower level of mRNA in general, suggesting that the exosome has a positive function in transcription. This observation agrees with studies in yeast that mutants of Rrp6, polyA factors and export factors all cause decreased mRNA level (Luna et al., 2005). In addition, depletion of all other known Drosophila exosome subunits described by (Graham et al., 2006) was also tested but no increase in the level of any mRNA was observed. Although the exosome is usually related to a surveillance role during elongation, the exact impact of the exosome on the transcription complex during elongation remains unclear. However, it is possible that factors other than the exosome are involved in degradation of this type of transcripts. Future study using more thorough genetic screen might provide more insights.

8.5 Outstanding problems and future perspectives

This thesis presented the discovery of large number of predicted polyA signals in Drosophila 5' UTRs. The exact number of 5' UTR polyA signals is unclear as the accuracy of the prediction programmes seemed only moderately acceptable. This could be due to three reasons: 1, the programmes were developed for human polyA signal prediction; 2, the accuracies of the programmes when predicting human polyA signals were also moderate; 3, lack of comprehensive controls in our prediction approach. However, the experimental validation gave us confidence that majority of the positive hits from the predictions are probably true. Furthermore, one of the five predicted negative hits unexpectedly showed positive function of polyA signal, implying the overall predictions might be underestimated.

This thesis also shows that polyA signals near the TSS do not seem to be productive. Relatively vague implications as to how it is regulated were provided (see section 8.2). The exact mechanism, however, requires further studies. Precise and direct measurements of CTD status and assembly of polyA complex would provide more insights on how is the switch between P1 and P2 polyA signals regulated. On the other hand, RNAi depletion of exosome showed no increase of early polyA signal's productivity. While more direct and robust method is required to confirm this result, it is still possible other RNA degradation pathways may be involved (Houseley and Tollervey, 2009).

A more biological question: Do genes with 5' UTR polyA signals share any similarities? Preliminary observations suggest these genes tend to have development-related functions. In which way the polyA signals might contribute this requires further investigations.

A possible direction would be to look at whether 5' UTR polyA signals are used in certain developmental stages. Limited testing in this thesis suggested the endogenous usage is extremely low. But the RT-PCR approach was unable to accurately capture low level of transcripts, let alone the lack of considering possible degradation pathways. On one hand, obtaining high-throughput sequencing data might provide a more global and quantitive view. On the other hand, transferring the reporter systems that already produce detectable early polyadenylated product (for example, Adh-SV40-P1) into transgenic flies for developmental-specific analysis might ease the effort of detecting the transcript.

Another interesting perspective would be to look at the influence of having the sequence of promoter proximal polyA signal on transcription, regardless of the usage of it in 3' end processing. Results in this thesis have already shown that early polyA signals, regardless of low activity, generally result in low level (in a few cases, below detectable) of steady state transcripts produced by a later polyA signal. Even when the hexamer was deleted, which led to undetectable activity of the early polyA signal, the readthrough transcript using the distal polyA signal did not show a dramatic increase. These unexpected observations suggest the polyA signal sequence might have functions other than inducing 3' end processing. Recently, deep-sequencing studies investigating Pol II promoter proximal pausing revealed potentials for development regulated mechanism (Muse et al., 2007; Nechaev et al., 2010; Zeitlinger et al., 2007). This notion may be supported by the observation that AAUAAA could facilitate Pol II pausing (Nag et al., 2006). In addition, it has been reported that nucleosome depletion is common around polyA signals whereas just downstream of polyA signals are usually nucleosome enriched areas (Spies et al., 2009). It is possible that sequence composition like a polyA signal could play a role to arrest Pol II, as the Tm track of sequence with paused Pol II appear similar to the Tm track at polyA sites (Nechaev et al., 2010). The dynamics of transcription complex when passing an early polyA signal would be worth of further investigations.

APPENDICES

Appendix 1 Depletion of Rtr1 and Brd4 do not affect relatively early polyA signal.

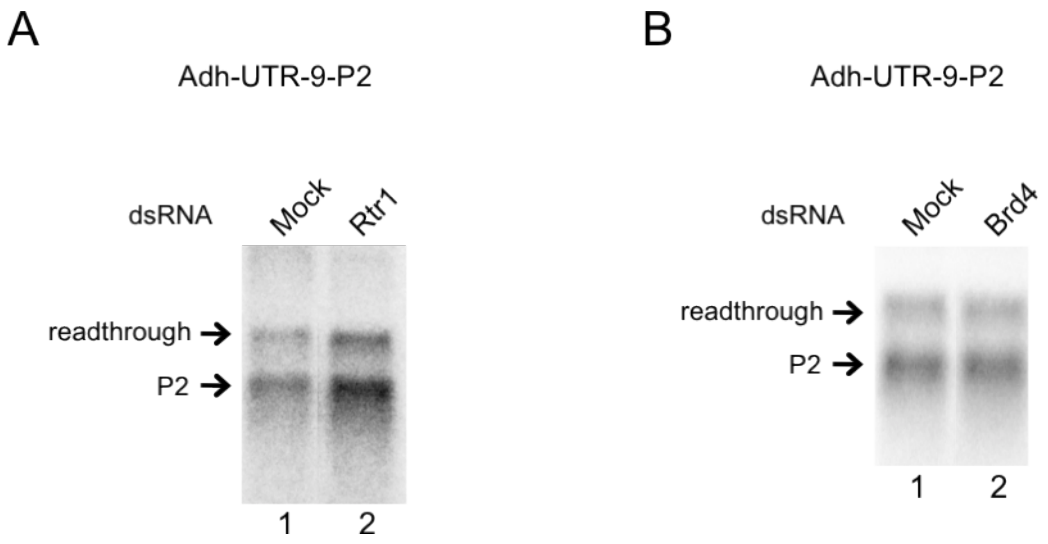


Fig A.1 Depletion of Rtr1 or Brd4 do not affect the activity of relatively early polyA signal. (A) Northern blot analysis of total RNA extracted from Rtr1 depleted S2 cells, transfected with the Adh-UTR-9-P2 reporter. Mock experiments refer to cells without dsRNA treatments. (B) Northern blot analysis of total RNA extracted from Rtr1 depleted S2 cells, transfected with the Adh-UTR-9-P2 reporter. Mock experiments refer to cells without dsRNA treatments.

Appendix 2 PolyA signal might inhibit nonsense mediated mRNA decay.

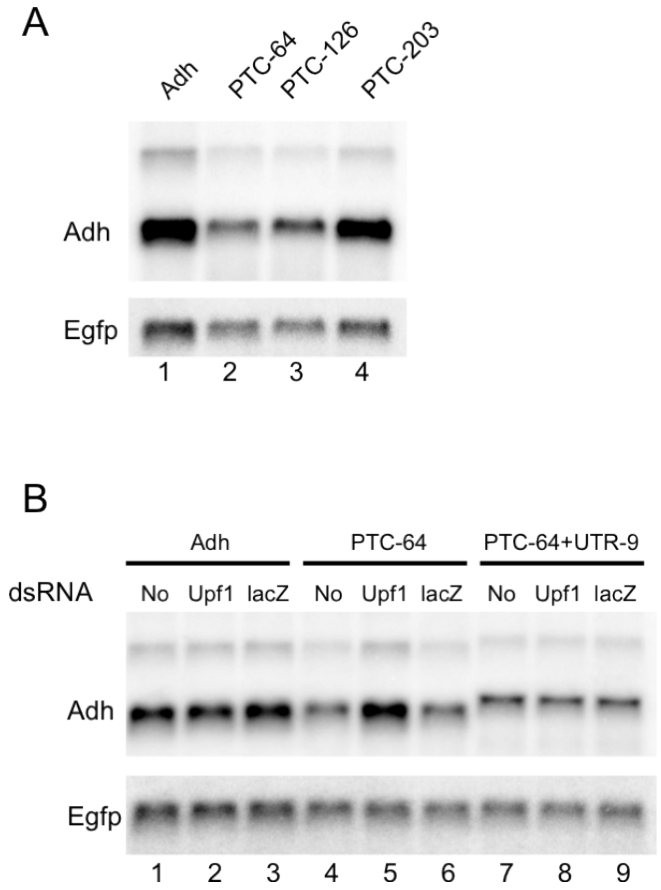


Fig A.2 Proximity to polyA signal prevents stop codon from being recognized by the nonsense mediated mRNA decay pathway. (A) Northern blot analysis of total RNA extracted from S2 cells transfected with the *Adh-Luc* reporters with pre-mature stop codons (PTC) in the *Adh*. (B) Northern blot analysis of total RNA extracted from Upf1 depleted S2 cells, transfected with corresponding *Adh-Luc* based reporters. Adh and PTC-64 are same as in A. PTC-64+UTR-9 has the sequence of UTR-9 inserted immediately downstream of the PTC. Both No dsRNA and lacZ dsRNA experiments serve as controls.

Appendix 3 Inverting Adh sequence might inhibit polyA signal activity.

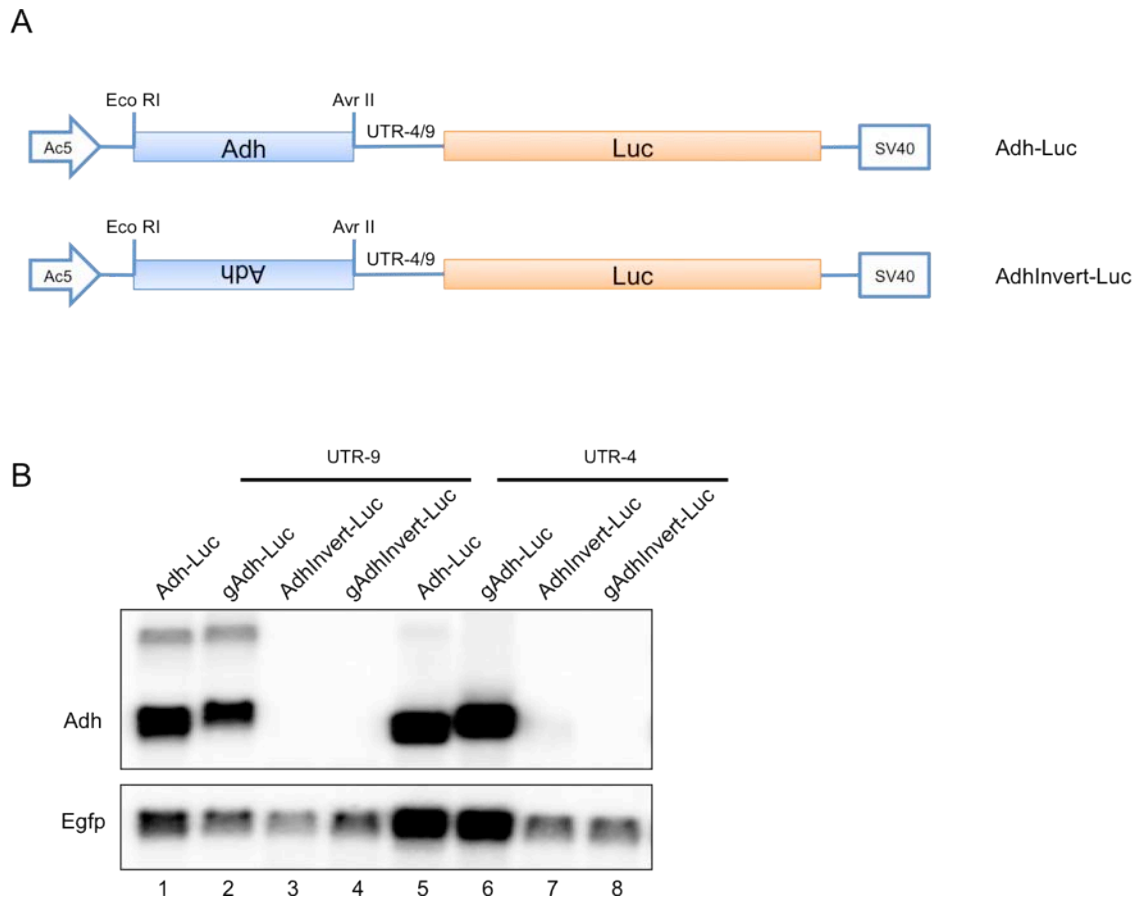


Fig A.3 Inverted *Adh* abolishes polyA signal activity. (A) Schematics of *Adh-Luc* based reporters with inverted *Adh* sequence. *Adh* sequence was inverted using primers with Eco RI site flanking 3' end and Avr II site flanking 5' end of Adh, followed by cloning back into the backbone. Both cDNA and genomic versions were made. Intergenic spacer is either UTR-4 or UTR-9. (B) Northern blot analysis of total RNA extracted S2 cells transfected with original Adh-Luc or AdhInvert-Luc reporters.

Appendix 4 Gene expression is unaffected by the presence of polyA signals in the 5' UTR.

Below are expression profiles of genes with 5' UTRs of UTR-1 to UTR-9 and Neg-1 to Neg-5. Data downloaded from Flybase in May 2010.

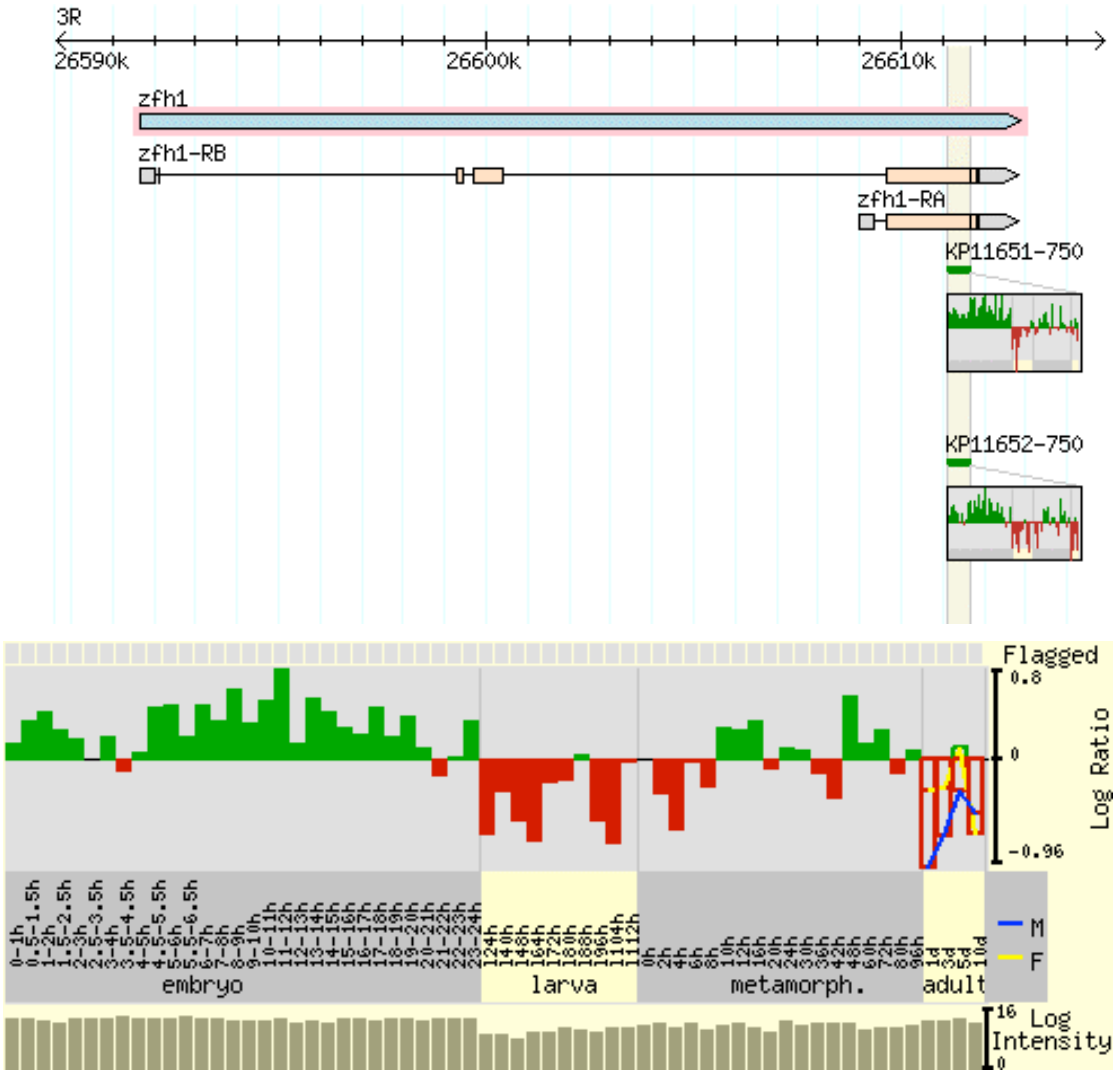


Figure A.4.1 Developmental time course for expression of CG1322 (origin of UTR-1).

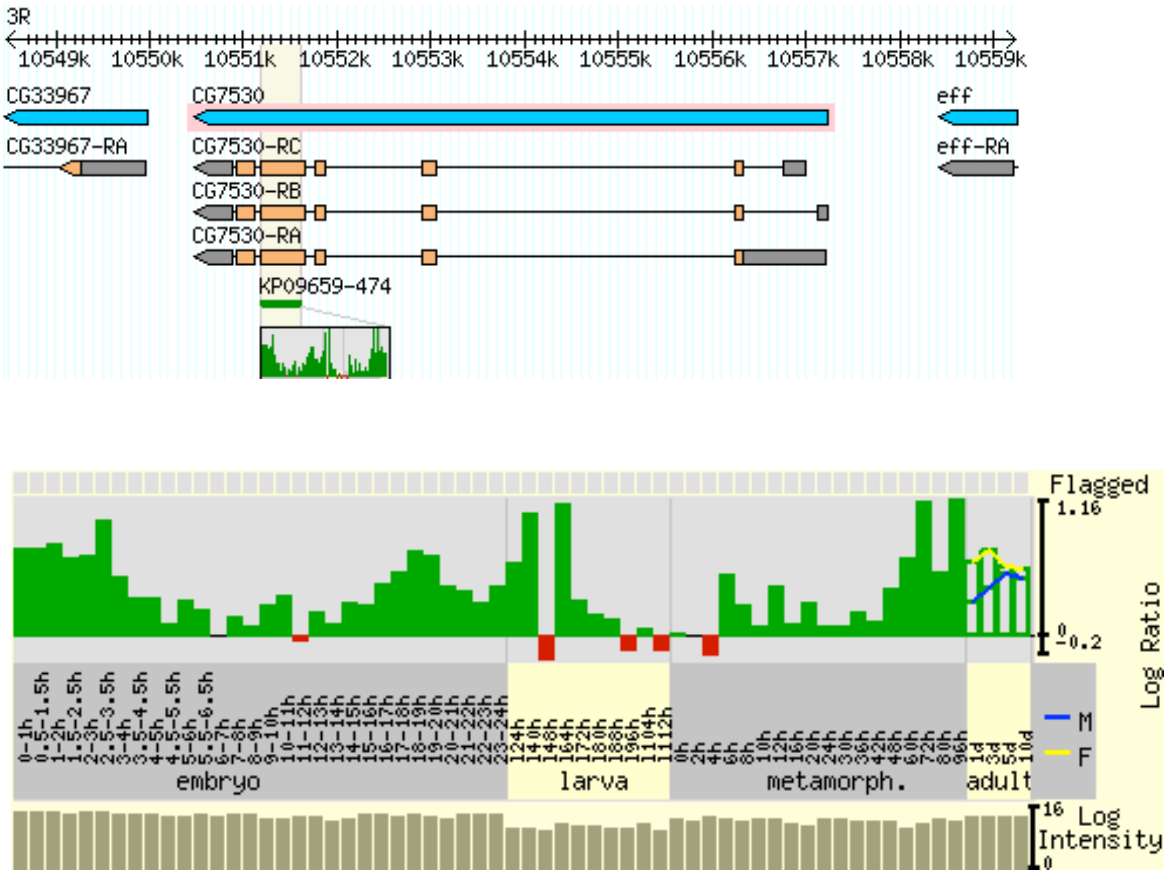


Figure A.4.2 Developmental time course for expression of CG7530 (origin of UTR-2).

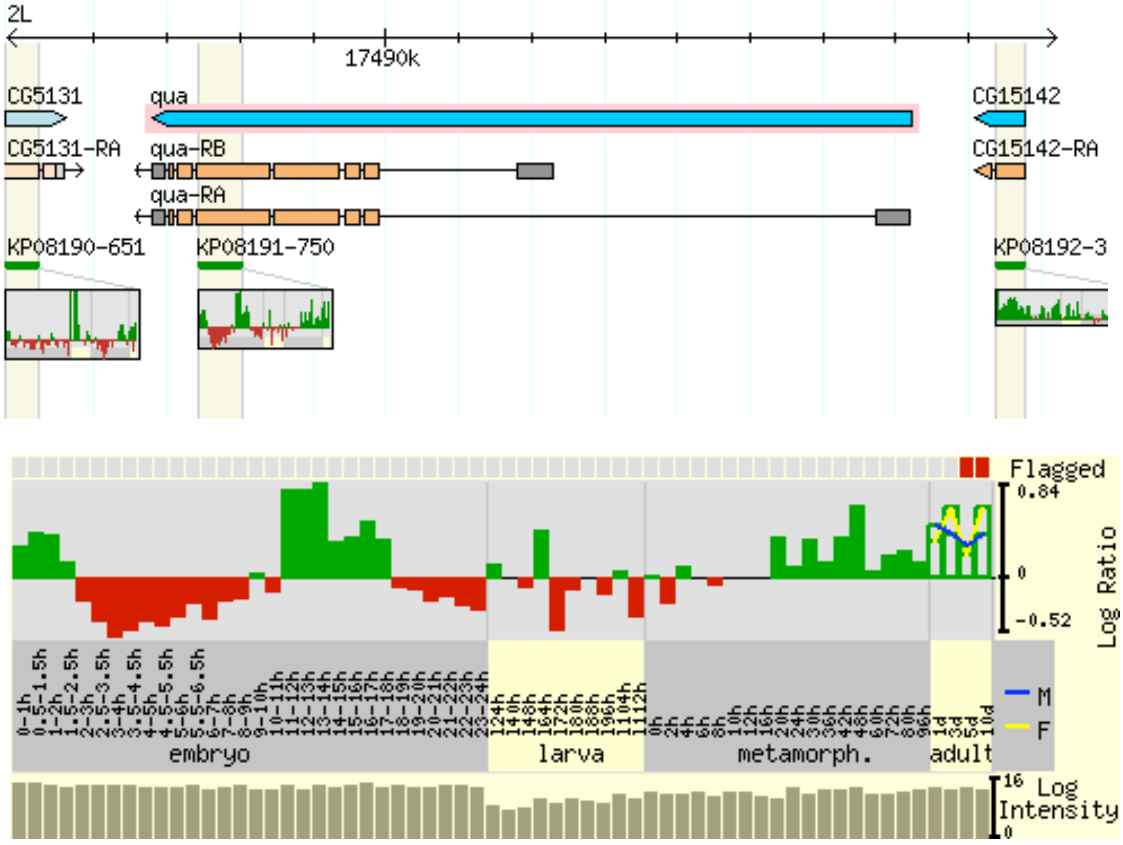


Figure A.4.3 Developmental time course for expression of CG6433 (origin of UTR-3).

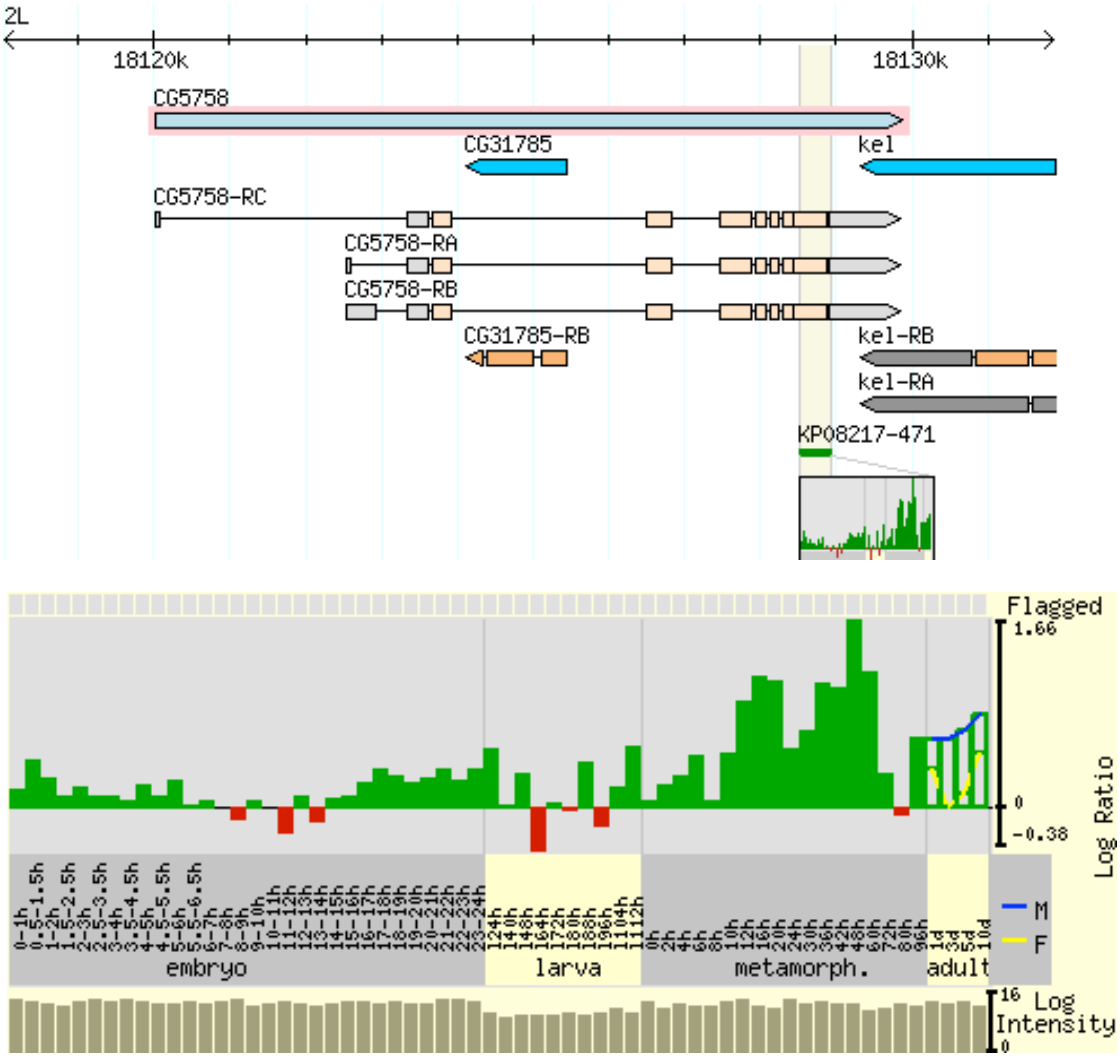


Figure A.4.4 Developmental time course for expression of CG5758 (origin of UTR-4).

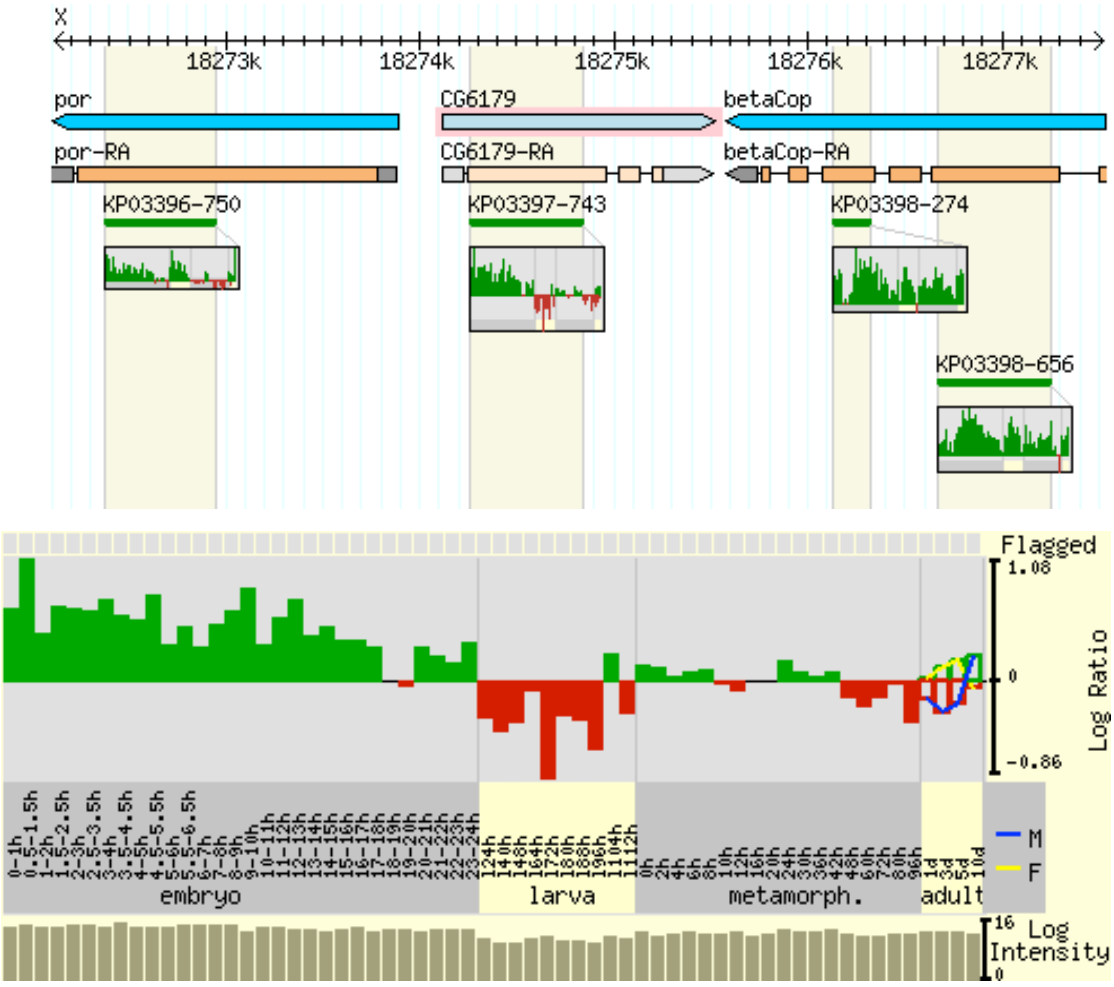


Figure A.4.5. Developmental time course for expression of CG6179 (origin of UTR-5).

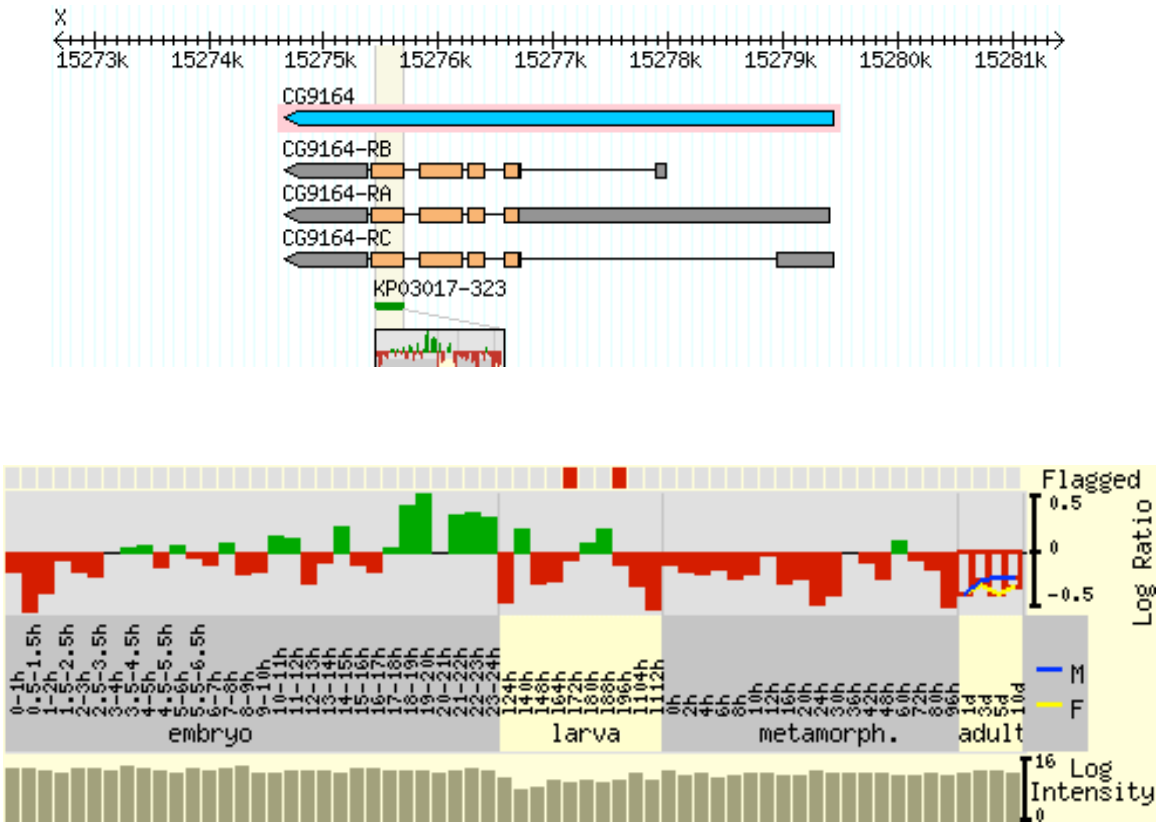


Figure A.4.6. Developmental time course for expression of CG9164 (origin of UTR-6).

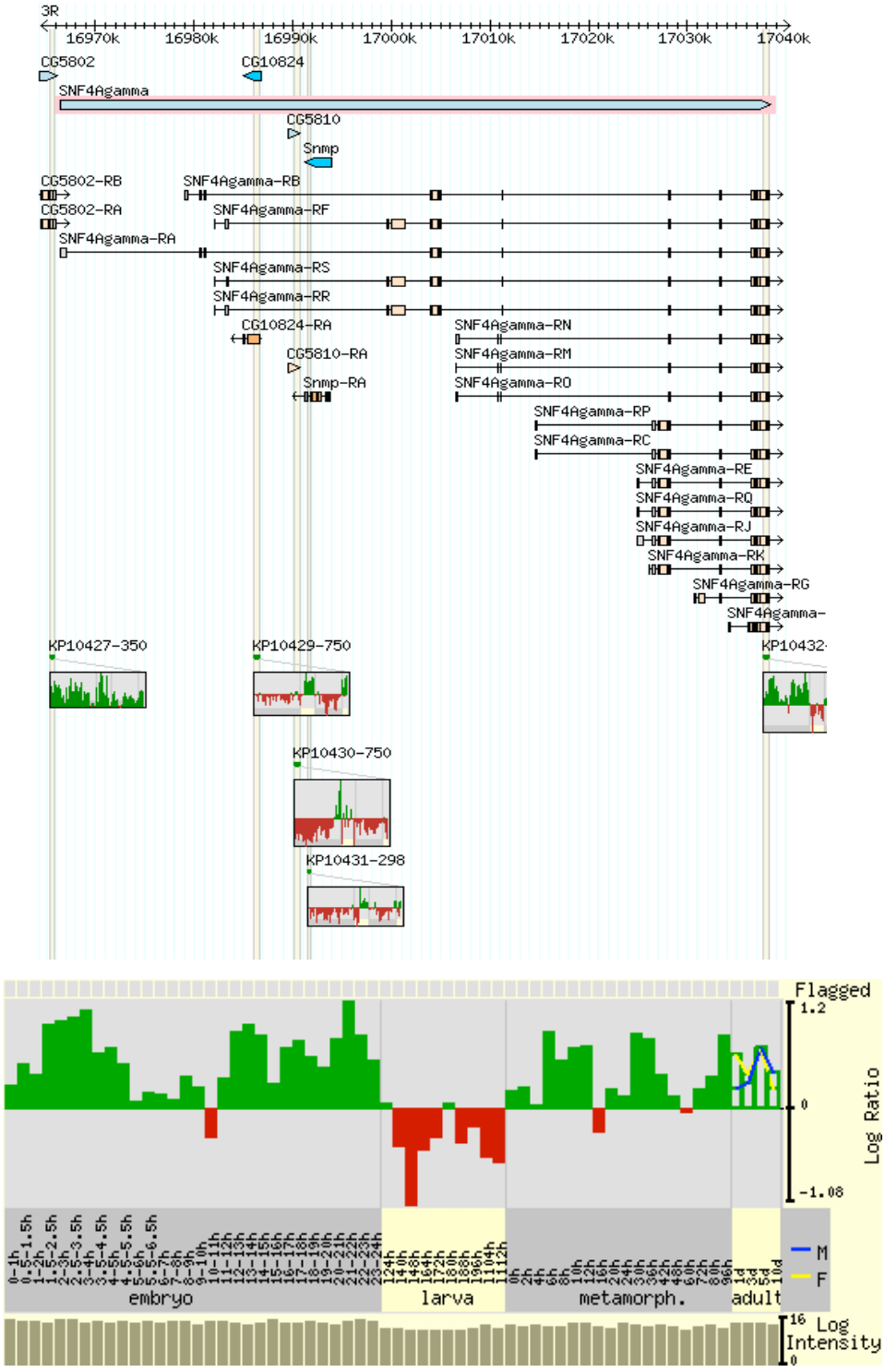


Figure A.4.7 Developmental time course for expression of CG17299 (origin of UTR-7).

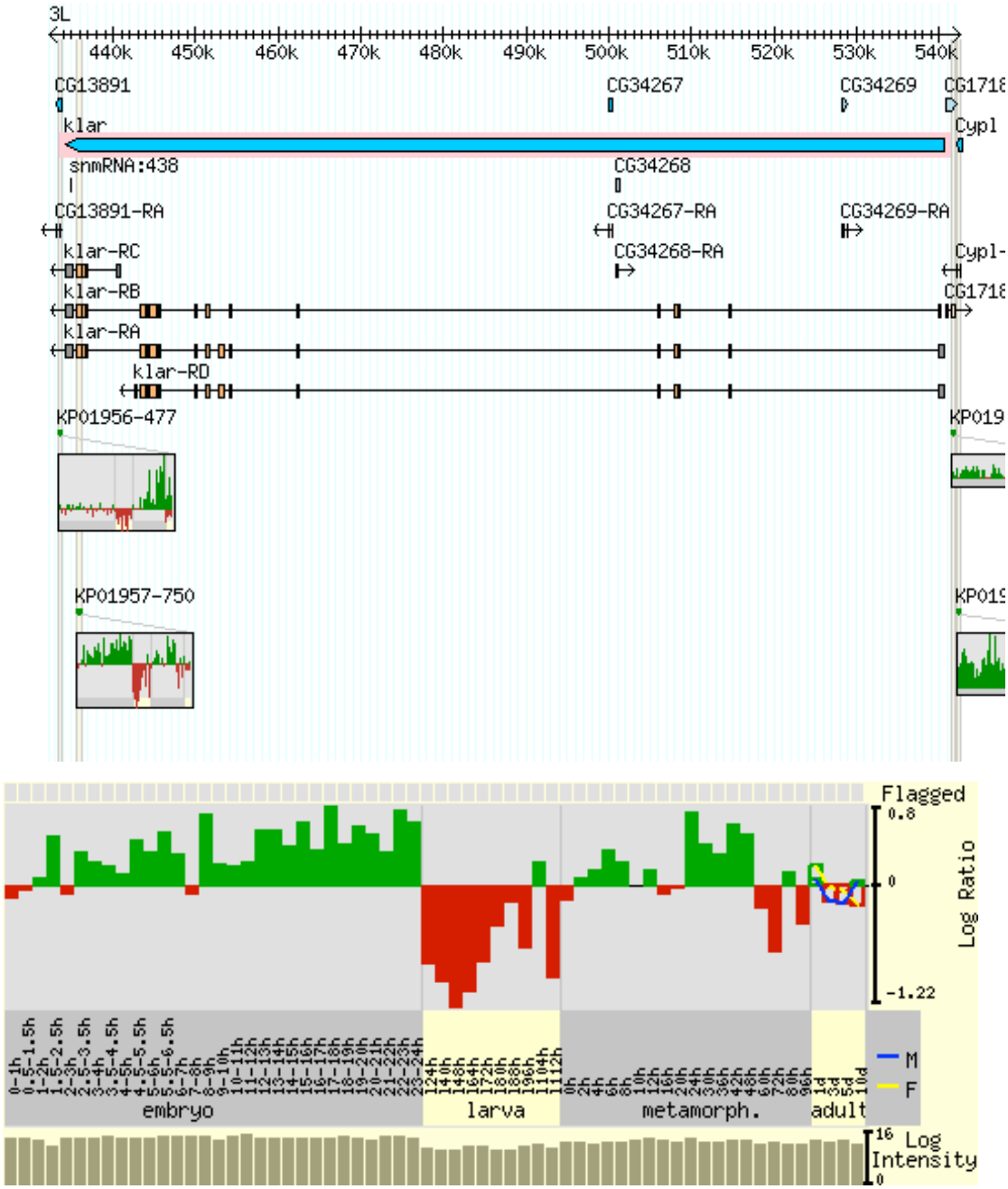


Figure A.4.8 Developmental time course for expression of CG17046 (origin of UTR-8).

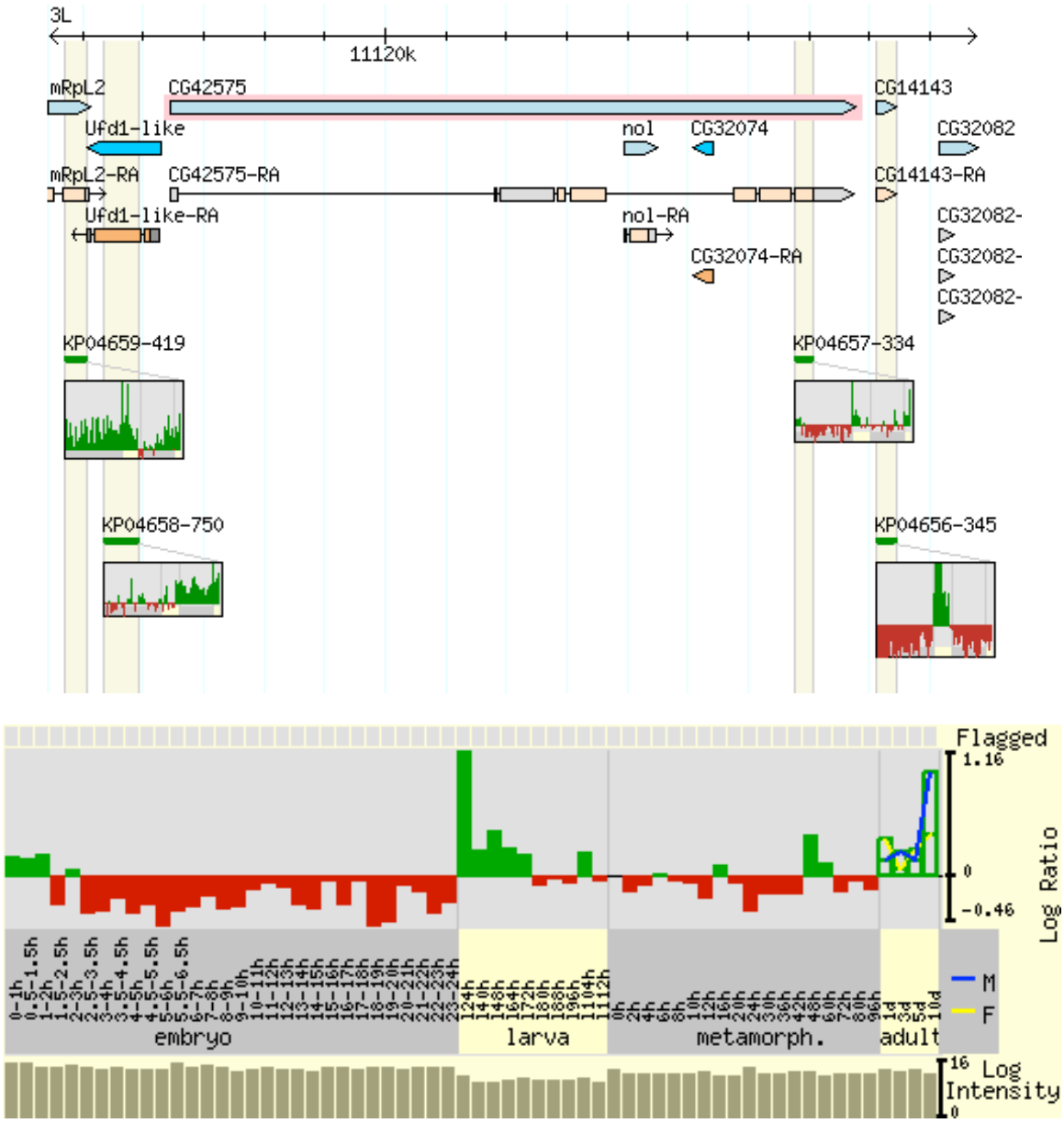


Figure A.4.9 Developmental time course for expression of CG42575 (origin of UTR-9).

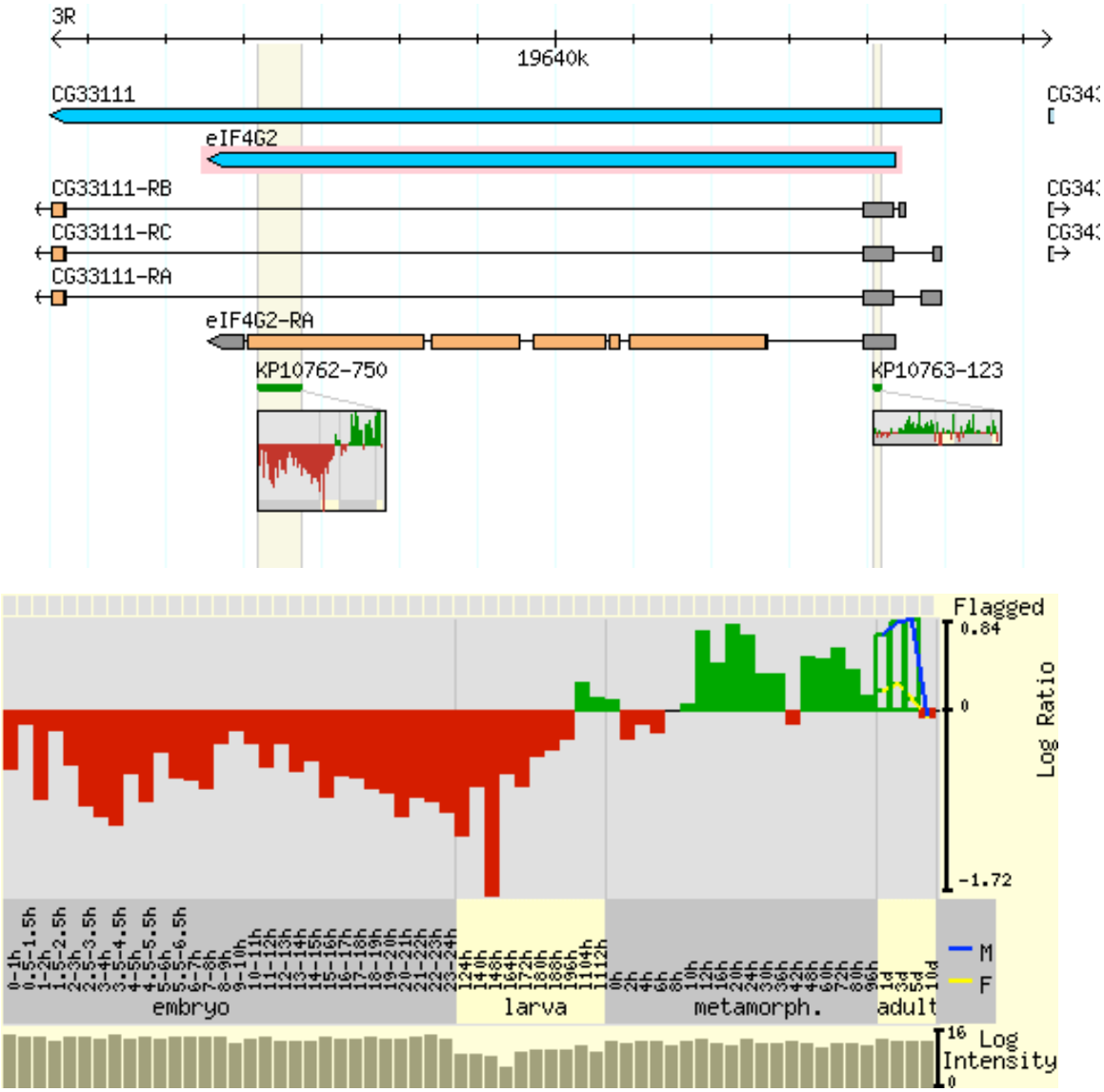


Figure A.4.10 Developmental time course for expression of CG10192 (origin of Neg-1).

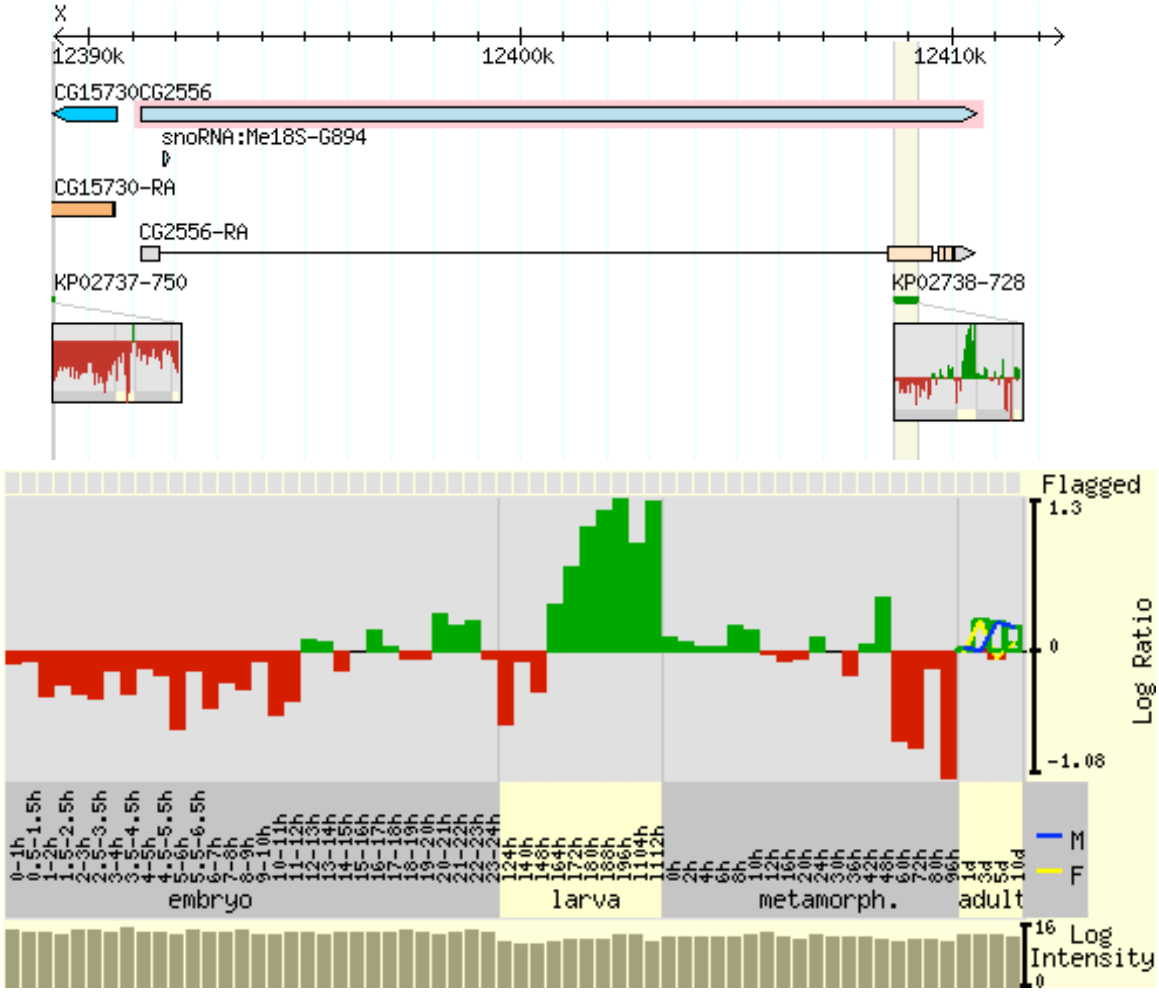


Figure A.4.11 Developmental time course for expression of CG2556 (origin of Neg-2).

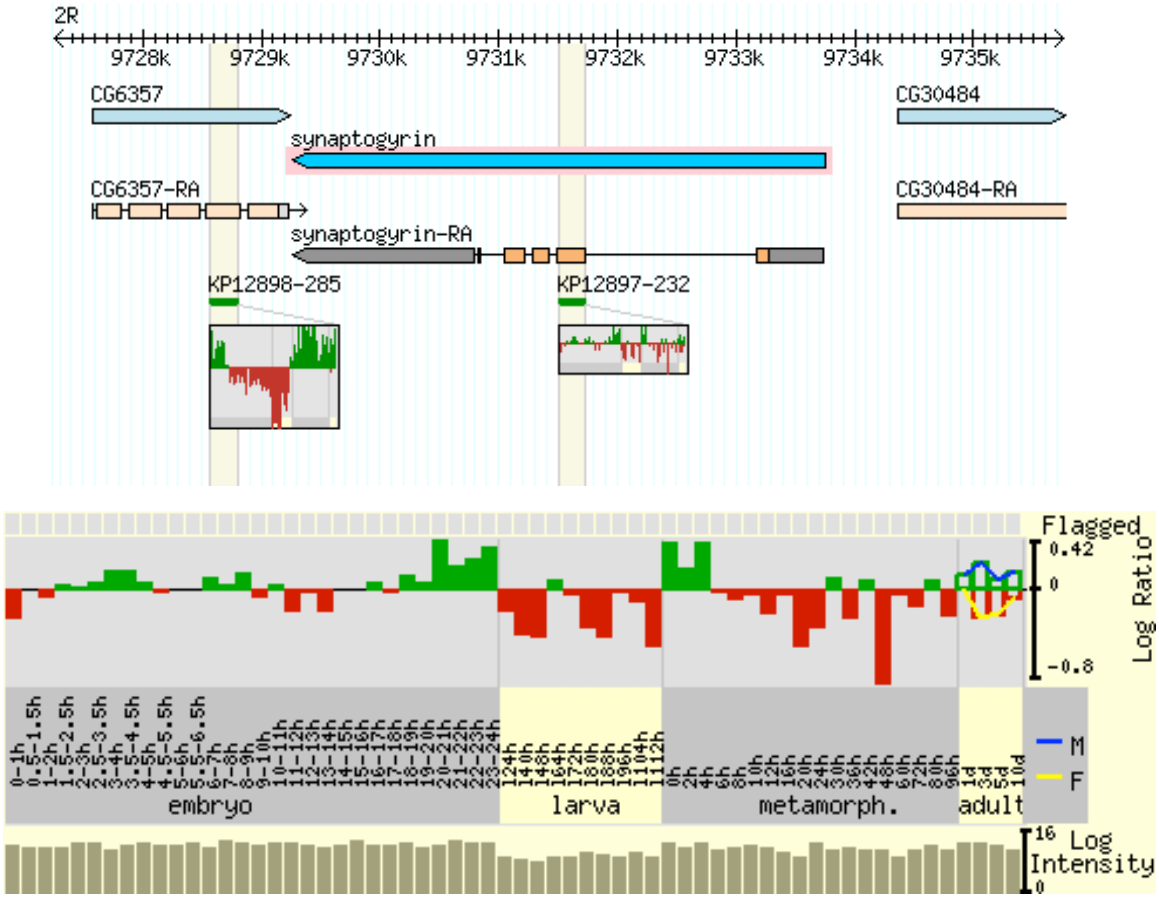


Figure A.4.12 Developmental time course for expression of CG10808 (origin of Neg-3).

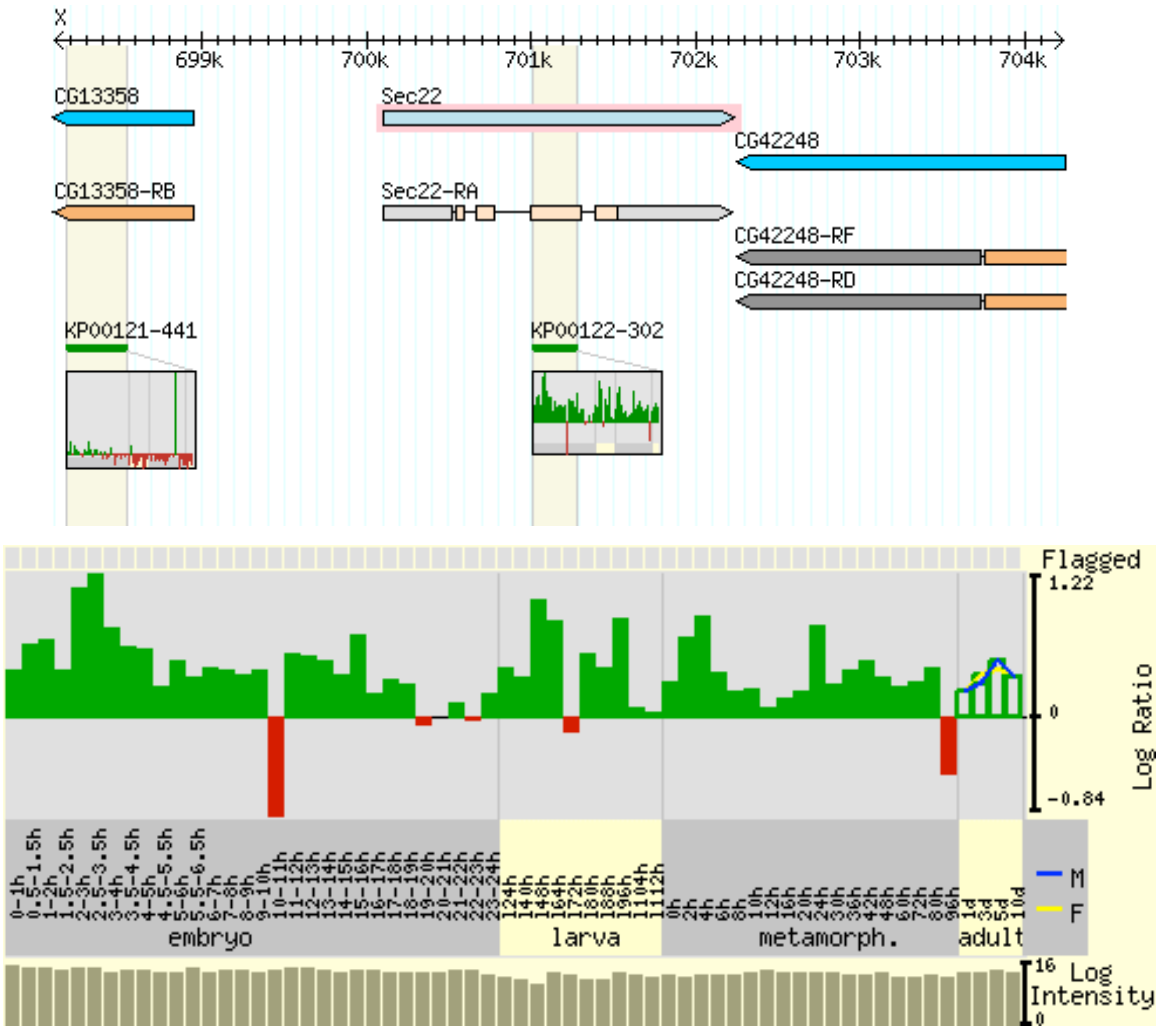


Figure A.4.13 Developmental time course for expression of CG7359 (origin of Neg-4).

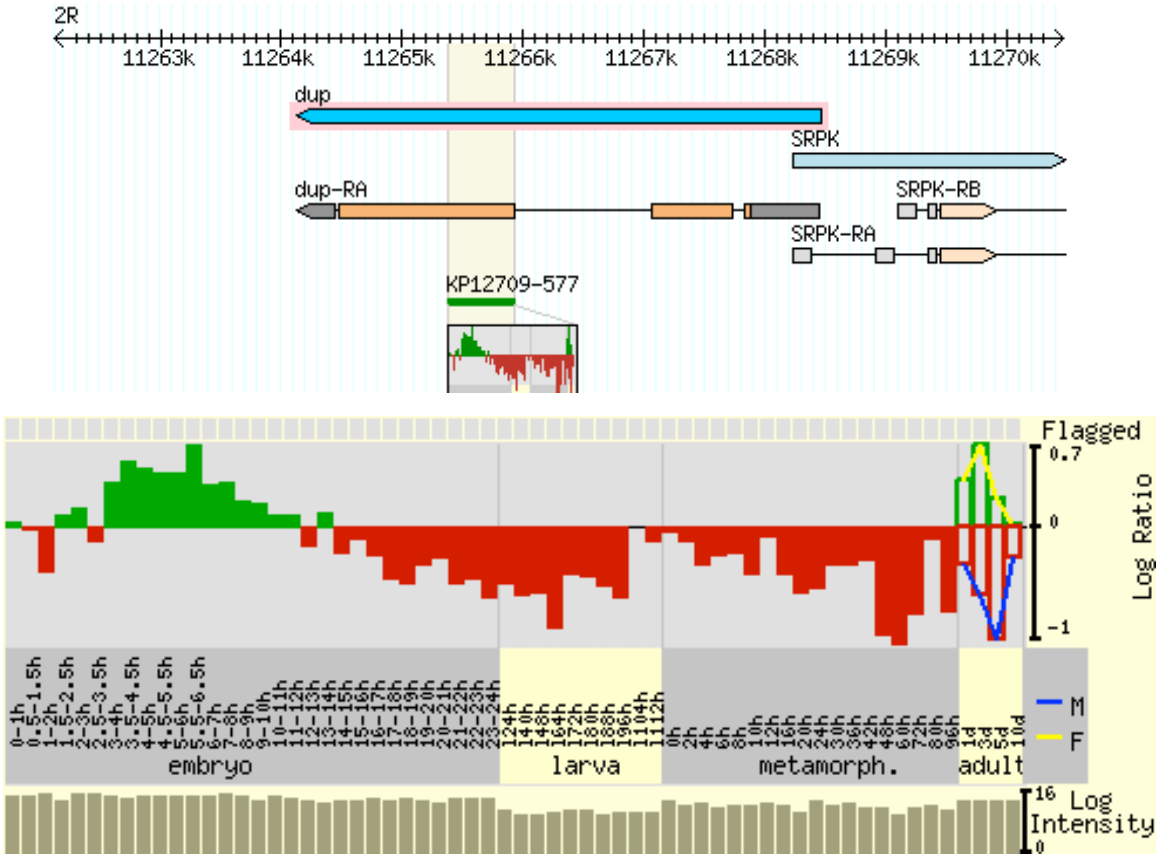


Figure A.4.14 Developmental time course for expression of CG8171 (origin of Neg-5).

Appendix 5 List of plasmids constructed in this thesis.

Name in thesis	Number in lab stock	Name in lab stock
UTR-1 Intron+	B204	pAutr-1L
UTR-2 Intron+	B205	pAutr-2L
UTR-3 Intron+	B206	pAutr-3L
UTR-4 Intron+	B207	pAutr-4L
UTR-5 Intron+	B208	pAutr-5L
UTR-6 Intron+	B209	pAutr-6L
UTR-7 Intron+	B210	pAutr-7L
UTR-8 Intron+	B211	pAutr-8L
UTR-9 Intron+	B212	pAutr-9L
UTR-10 Intron+	B213	pAutr-10L
UTR-1 Intron-	B214	pcAutr-1L
UTR-2 Intron-	B215	pcAutr-2L
UTR-3 Intron-	B216	pcAutr-3L
UTR-4 Intron-	B217	pcAutr-4L
UTR-5 Intron-	B218	pcAutr-5L
UTR-6 Intron-	B219	pcAutr-6L
UTR-7 Intron-	B220	pcAutr-7L
UTR-8 Intron-	B221	pcAutr-8L
UTR-9 Intron-	B222	pcAutr-9L
UTR-10 Intron-	B223	pcAutr-10L
gAdhInver-Luc-UTR-4	B234	pInvertAutr4L
AdhInver-Luc-UTR-4	B235	pInvertCAutr4L
gAdhInver-Luc-UTR-9	B236	pInvertAutr9L
AdhInver-Luc-UTR-9	B237	pInvertCAutr9L
Neg-1	B242	pAneg-1L
Neg-2	B243	pAneg-2L
Neg-3	B244	pAneg-3L
Neg-4	B245	pAneg-4L
Neg-5	B246	pAneg-5L
Adh-UTR-9-P1	B255	pcA64+9-wtL
Adh-UTR-9-P2	B256	pcA126+9-wtL
Adh-UTR-9-P3	B257	pcA203+9-wtL
Adh-PTC-64	B258	pcA64stop-wtL
Adh-PTC-126	B259	pcA126stop-wtL
Adh-PTC-203	B260	pcA203stop-wtL
S-UTR-9	B261	pcAs-UTR-9L
Adh-S-UTR-9-P1	B263	pcA64s-9-wtL
Adh-S-UTR-9-P2	B264	pcA126s-9-wtL

Adh-S-UTR-9-P3	B265	pcA203s-9-wtL
Adh-SV40-P1	B266	pcA64sv40-wtL
Adh-SV40-P2	B267	pcA126sv40-wtL
Adh-SV40-P3	B268	pcA203sv40-wtL
Adh-S-UTR-9-ΔTAA-ΔAATAAA-P1	B269	pcA64s-9-NSNA-wtL
Adh-S-UTR-9-ΔTAA-ΔAATAAA-P2	B270	pcA126s-9-NSNA-wtL
Adh-S-UTR-9-ΔTAA-ΔAATAAA-P3	B271	pcA203s-9-NSNA-wtL
Adh-UTR-9-ΔP2	B274	pD1-64_cA126+9-wtL
Adh-UTR-9-ΔP3	B275	pD1-64_cA203+9-wtL
Adh-SV40-ΔP2	B276	pD1-64_cA126sv40-wtL
Adh-SV40-ΔP3	B277	pD1-64_cA203sv40-wtL
LacZ-BGH-P1	B279	pKpnBGH-LacZ
LacZ-BGH-P2	B280	pLacZ-49BGH
LacZ-BGH-P3	B281	pLacZ-149BGH
Luc-UTR-4-P1	B284	pKpnutr4-Luc
Luc-UTR-4-P2	B285	pLuc64+utr4
Luc-UTR-4-P3	B286	pLuc203+utr4
hAdh-SV40-P1	B296	pCDcA64sv40-wtL
hAdh-SV40-P2	B297	pCDcA126sv40-wtL
hAdh-SV40-P3	B298	pCDcA203sv40-wtL
hAdh-SV40-ΔP2	B299	pCDD1-64cA126sv40-wtL
hAdh-SV40-ΔP3	B300	pCDD1-64cA203sv40-wtL

Table A.5 Table of plasmid constructs made and used in this thesis. First column lists the names as they appear in the thesis. Second column indicates label of plasmid in the Brogna Lab plasmid stock. Third column indicates names of plasmids as in the lab stock.

Appendix 6 Selected validation of RNAi by RT-PCR

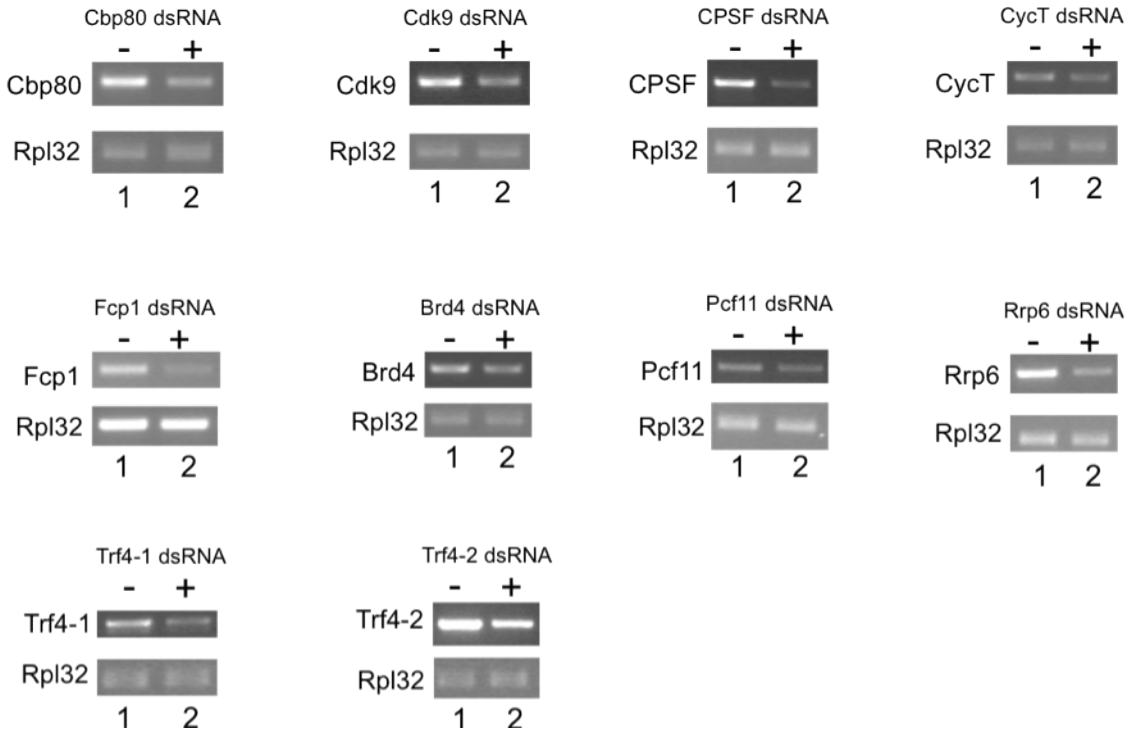


Fig A.6. RT-PCR validation of RNAi depletion. Total RNA from S2 cells treated with or without indicated dsRNA was used in reverse transcription. The following PCR using gene specific primers amplifying exonic region was individually optimised to achieve obvious comparison without reaching saturation. Each validation was repeated with at least two randomly selected RNAi experiments. Primer information is included in Materials and Methods.

References

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

Akhtar, M.S., Heidemann, M., Tietjen, J.R., Zhang, D.W., Chapman, R.D., Eick, D., and Ansari, A.Z. (2009). TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II. Mol Cell *34*, 387-393.

Andrulis, E.D., Werner, J., Nazarian, A., Erdjument-Bromage, H., Tempst, P., and Lis, J.T. (2002). The RNA processing exosome is linked to elongating RNA polymerase II in Drosophila. Nature *420*, 837-841.

Ashe, M.P., Furger, A., and Proudfoot, N.J. (2000). Stem-loop 1 of the U1 snRNP plays a critical role in the suppression of HIV-1 polyadenylation. RNA *6*, 170-177.

Ashe, M.P., Griffin, P., James, W., and Proudfoot, N.J. (1995). Poly(A) site selection in the HIV-1 provirus: inhibition of promoter-proximal polyadenylation by the downstream major splice donor site. Genes Dev *9*, 3008-3025.

Ashe, M.P., Pearson, L.H., and Proudfoot, N.J. (1997). The HIV-1 5' LTR poly(A) site is inactivated by U1 snRNP interaction with the downstream major splice donor site. EMBO J *16*, 5752-5763.

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

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

Beelman, C.A., and Parker, R. (1995). Degradation of mRNA in eukaryotes. Cell *81*, 179-183.

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

Bienroth, S., Wahle, E., Suter-Crazzolara, C., and Keller, W. (1991). Purification of the cleavage and polyadenylation factor involved in the 3'-processing of messenger RNA precursors. J Biol Chem *266*, 19768-19776.

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

Brogna, S. (1999). Nonsense mutations in the alcohol dehydrogenase gene of Drosophila melanogaster correlate with an abnormal 3' end processing of the corresponding pre-mRNA. RNA *5*, 562-573.

Brogna, S., and Ashburner, M. (1997). The Adh-related gene of Drosophila melanogaster is expressed as a functional dicistronic messenger RNA: multigenic transcription in higher organisms. EMBO J *16*, 2023-2031.

Buhler, M., Haas, W., Gygi, S.P., and Moazed, D. (2007). RNAi-Dependent and -Independent RNA Turnover Mechanisms Contribute to Heterochromatic Gene Silencing. Cell *129*, 707-721.

Buratowski, S. (2008). Transcription. Gene expression--where to start? Science *322*, 1804-1805.

Buratowski, S. (2009). Progression through the RNA Polymerase II CTD Cycle. Mol Cell *36*, 541-546.

Carroll, K.L., Pradhan, D.A., Granek, J.A., Clarke, N.D., and Corden, J.L. (2004). Identification of cis elements directing termination of yeast nonpolyadenylated snoRNA transcripts. Mol Cell Biol *24*, 6241-6252.

Chan, H.Y., Brogna, S., and O'Kane, C.J. (2001). Dribble, the Drosophila KRR1p homologue, is involved in rRNA processing. Mol Biol Cell *12*, 1409-1419.

Chang, Y.-L., King, B., Lin, S.-C., Kennison, J.A., and Huang, D.-H. (2007). A Double-Bromodomain Protein, FSH-S, Activates the Homeotic Gene Ultrabithorax through a Critical Promoter-Proximal Region. Molecular and Cellular Biology *27*, 5486-5498.

Chen, F., MacDonald, C.C., and Wilusz, J. (1995). Cleavage site determinants in the mammalian polyadenylation signal. Nucleic Acids Res *23*, 2614-2620.

Cheng, Y., Miura, R.M., and Tian, B. (2006). Prediction of mRNA polyadenylation sites by support vector machine. Bioinformatics *22*, 2320-2325.

Cho, E.J., Kobor, M.S., Kim, M., Greenblatt, J., and Buratowski, S. (2001). Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain. Genes Dev *15*, 3319-3329.

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

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

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. Science *322*, 1845-1848.

Couttet, P., Fromont-Racine, M., Steel, D., Pictet, R., and Grange, T. (1997). Messenger RNA deadenylylation precedes decapping in mammalian cells. PNAS *94*, 5628-5633.

Cramer, P., Bushnell, D.A., Fu, J., Gnatt, A.L., Maier-Davis, B., Thompson, N.E., Burgess, R.R., Edwards, A.M., David, P.R., and Kornberg, R.D. (2000). Architecture of RNA polymerase II and implications for the transcription mechanism. Science *288*, 640-649.

Cramer, P., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science *292*, 1863-1876.

Cui, M., Allen, M.A., Larsen, A., MacMorris, M., Han, M., and Blumenthal, T. (2008). Genes involved in pre-mRNA 3' end formation and transcription termination revealed by a lin-15 operon Muv suppressor screen. PNAS *105*, 16665-16670.

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

Dichtl, B., Blank, D., Sadowski, M., Hubner, W., Weiser, S., and Keller, W. (2002). Yhh1p/Cft1p directly links poly(A) site recognition and RNA polymerase II transcription termination. EMBO J *21*, 4125-4135.

Dominski, Z. (2007). Nucleases of the Metallo-&b.beta;-lactamase Family and Their Role in DNA and RNA Metabolism. Critical Reviews in Biochemistry and Molecular Biology *42*, 67 - 93.

Dominski, Z., and Marzluff, W.F. (2007). Formation of the 3' end of histone mRNA: getting closer to the end. Gene *396*, 373-390.

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

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

Garrido-Lecca, A., and Blumenthal, T. (2010). RNA polymerase II CTD phosphorylation patterns in C. elegans operons, polycistronic gene clusters with only one promoter. Mol Cell Biol.

Ghazal, G., Gagnon, J., Jacques, P.E., Landry, J.R., Robert, F., and Elela, S.A. (2009). Yeast RNase III triggers polyadenylation-independent transcription termination. Mol Cell *36*, 99-109.

Ghosh, A., Shuman, S., and Lima, C.D. (2008). The structure of Fcp1, an essential RNA polymerase II CTD phosphatase. Mol Cell *32*, 478-490.

Gilat, R., Goncharov, S., Esterman, N., and Shweiki, D. (2006). Under-representation of PolyA/PolyT tailed ESTs in Human ESTdb: an obstacle to alternative polyadenylation inference. Bioinformation *1*, 220-224.

Gilmartin, G.M., Fleming, E.S., Oetjen, J., and Graveley, B.R. (1995). CPSF recognition of an HIV-1 mRNA 3'-processing enhancer: multiple sequence contacts involved in poly(A) site definition. Genes Dev *9*, 72-83.

Gilmartin, G.M., and Nevins, J.R. (1989). An ordered pathway of assembly of components required for polyadenylation site recognition and processing. Genes Dev *3*, 2180-2190.

Gilsdorf, M., Horn, T., Arziman, Z., Pelz, O., Kiner, E., and Boutros, M. (2010). GenomeRNAi: a database for cell-based RNAi phenotypes. 2009 update. Nucleic Acids Res *38*, D448-452.

Glover-Cutter, K., Kim, S., Espinosa, J., and Bentley, D.L. (2007). RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. Nat Struct Mol Biol *advanced online publication*.

Glover-Cutter, K., Larochelle, S., Erickson, B., Zhang, C., Shokat, K., Fisher, R.P., and Bentley, D.L. (2009). TFIIH-associated Cdk7 kinase functions in phosphorylation of C-terminal domain Ser7 residues, promoter-proximal pausing, and termination by RNA polymerase II. Mol Cell Biol *29*, 5455-5464.

Gnatt, A.L., Cramer, P., Fu, J., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. Science *292*, 1876-1882.

Goodnow, C.C., Crosbie, J., Adelstein, S., Lavoie, T.B., Smith-Gill, S.J., Brink, R.A., Pritchard-Briscoe, H., Wotherspoon, J.S., Loblay, R.H., Raphael, K., *et al.* (1988). Altered immunoglobulin expression and functional silencing of self-reactive B lymphocytes in transgenic mice. Nature *334*, 676-682.

Graham, A.C., Kiss, D.L., and Andrulis, E.D. (2006). Differential Distribution of Exosome Subunits at the Nuclear Lamina and in Cytoplasmic Foci. Mol Biol Cell

17, 1399-1409.

Graveley, B.R., Fleming, E.S., and Gilmartin, G.M. (1996). RNA structure is a critical determinant of poly(A) site recognition by cleavage and polyadenylation specificity factor. Mol Cell Biol *16*, 4942-4951.

Gross, S., and Moore, C.L. (2001). Rna15 interaction with the A-rich yeast polyadenylation signal is an essential step in mRNA 3'-end formation. Mol Cell Biol *21*, 8045-8055.

Grzechnik, P., and Kufel, J. (2008). Polyadenylation Linked to Transcription Termination Directs the Processing of snoRNA Precursors in Yeast. Molecular Cell *32*, 247-258.

Gudipati, R.K., Villa, T., Boulay, J., and Libri, D. (2008). Phosphorylation of the RNA polymerase II C-terminal domain dictates transcription termination choice. Nat Struct Mol Biol *15*, 786-794.

Han, K. (1996). An efficient DDAB-mediated transfection of Drosophila S2 cells. Nucleic Acids Res *24*, 4362-4363.

Hargreaves, D.C., Horng, T., and Medzhitov, R. (2009). Control of inducible gene expression by signal-dependent transcriptional elongation. Cell *138*, 129-145.

Hilleren, P., McCarthy, T., Rosbash, M., Parker, R., and Jensen, T.H. (2001). Quality control of mRNA 3'-end processing is linked to the nuclear exosome. Nature *413*, 538-542.

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

Holton, T.A., and Graham, M.W. (1991). A simple and efficient method for direct cloning of PCR products using ddT-tailed vectors. Nucleic Acids Res *19*, 1156-.

Houseley, J., LaCava, J., and Tollervey, D. (2006). RNA-quality control by the exosome. Nat Rev Mol Cell Biol *7*, 529-539.

Houseley, J., and Tollervey, D. (2009). The many pathways of RNA degradation. Cell *136*, 763-776.

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

Ishigaki, Y., Li, X., Serin, G., and Maquat, L.E. (2001). Evidence for a Pioneer Round of mRNA Translation: mRNAs Subject to Nonsense-Mediated Decay in

Mammalian Cells Are Bound by CBP80 and CBP20. Cell *106*, 607-617.

Jang, M.K., Mochizuki, K., Zhou, M., Jeong, H.S., Brady, J.N., and Ozato, K. (2005). The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. Mol Cell *19*, 523-534.

Jensen, T.H., Patricio, K., McCarthy, T., and Rosbash, M. (2001). A block to mRNA nuclear export in S. cerevisiae leads to hyperadenylation of transcripts that accumulate at the site of transcription. Mol Cell *7*, 887-898.

Jurica, M.S., and Moore, M.J. (2003). Pre-mRNA Splicing: Awash in a Sea of Proteins. Molecular Cell *12*, 5-14.

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

Kim, H., Erickson, B., Luo, W., Seward, D., Graber, J.H., Pollock, D.D., Megee, P.C., and Bentley, D.L. (2010). Gene-specific RNA polymerase II phosphorylation and the CTD code. Nat Struct Mol Biol *17*, 1279-1286.

Kim, M., Ahn, S.H., Krogan, N.J., Greenblatt, J.F., and Buratowski, S. (2004a). Transitions in RNA polymerase II elongation complexes at the 3' ends of genes. EMBO J *23*, 354-364.

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

Kim, M., Suh, H., Cho, E.J., and Buratowski, S. (2009). Phosphorylation of the yeast Rpb1 C-terminal domain at serines 2, 5, and 7. J Biol Chem *284*, 26421-26426.

Komarnitsky, P., Cho, E.J., and Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. Genes Dev *14*, 2452-2460.

Kyburz, A., Friedlein, A., Langen, H., and Keller, W. (2006). Direct Interactions between Subunits of CPSF and the U2 snRNP Contribute to the Coupling of Pre-mRNA 3' End Processing and Splicing. Molecular Cell *23*, 195-205.

LaCava, J., Houseley, J., Saveanu, C., Petfalski, E., Thompson, E., Jacquier, A., and Tollervey, D. (2005). RNA Degradation by the Exosome Is Promoted by a Nuclear Polyadenylation Complex. Cell *121*, 713-724.

Lee, J.Y., Ji, Z., and Tian, B. (2008). Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. Nucleic Acids Research, gkn540.

Legendre, M., and Gautheret, D. (2003). Sequence determinants in human polyadenylation site selection. BMC genomics *4*, 7.

Libri, D., Dower, K., Boulay, J., Thomsen, R., Rosbash, M., and Jensen, T.H. (2002). Interactions between mRNA Export Commitment, 3'-End Quality Control, and Nuclear Degradation. Molecular and Cellular Biology *22*, 8254-8266.

Licatalosi, D.D., Geiger, G., Minet, M., Schroeder, S., Cilli, K., McNeil, J.B., and Bentley, D.L. (2002). Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II. Mol Cell *9*, 1101-1111.

Logan, J., Falck-Pedersen, E., Darnell, J.E., Jr., and Shenk, T. (1987). A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. Proc Natl Acad Sci U S A *84*, 8306-8310.

Lopez, F., Granjeaud, S., Ara, T., Ghattas, B., and Gautheret, D. (2006). The disparate nature of "intergenic" polyadenylation sites. RNA *12*, 1794-1801.

Luna, R., Jimeno, S., Marin, M., Huertas, P., Garcia-Rubio, M., and Aguilera, A. (2005). Interdependence between transcription and mRNP processing and export, and its impact on genetic stability. Mol Cell *18*, 711-722.

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

Mandel, C.R., Bai, Y., and Tong, L. (2008). Protein factors in pre-mRNA 3'-end processing. Cell Mol Life Sci *65*, 1099-1122.

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

Mao, X., Green, J.M., Safer, B., Lindsten, T., Frederickson, R.M., Miyamoto, S., Sonenberg, N., and Thompson, C.B. (1992). Regulation of translation initiation factor gene expression during human T cell activation. J Biol Chem *267*, 20444-20450.

Mayer, A., Lidschreiber, M., Siebert, M., Leike, K., Soding, J., and Cramer, P. (2010). Uniform transitions of the general RNA polymerase II transcription complex. Nat Struct Mol Biol.

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

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

Meinhart, A., and Cramer, P. (2004). Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors. Nature *430*, 223-226.

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

Millevoi, S., and Vagner, S. (2009). Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. Nucleic Acids Research, gkp1176.

Miyamoto, S., Chiorini, J.A., Urcelay, E., and Safer, B. (1996). Regulation of gene expression for translation initiation factor eIF-2 alpha: importance of the 3' untranslated region. Biochem J *315 (Pt 3)*, 791-798.

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

Moreira, A., Takagaki, Y., Brackenridge, S., Wollerton, M., Manley, J.L., and Proudfoot, N.J. (1998). The upstream sequence element of the C2 complement poly(A) signal activates mRNA 3' end formation by two distinct mechanisms. Genes Dev *12*, 2522-2534.

Mosley, A.L., Pattenden, S.G., Carey, M., Venkatesh, S., Gilmore, J.M., Florens, L., Workman, J.L., and Washburn, M.P. (2009). Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation. Mol Cell *34*, 168-178.

Moucadel, V., Lopez, F., Ara, T., Benech, P., and Gautheret, D. (2007). Beyond the 3' end: experimental validation of extended transcript isoforms. Nucleic Acids Res *35*, 1947-1957.

Murthy, K.G., and Manley, J.L. (1992). Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus. J Biol Chem *267*, 14804-14811.

Murthy, K.G., and Manley, J.L. (1995). The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation. Genes & Development *9*, 2672-2683.

Muse, G.W., Gilchrist, D.A., Nechaev, S., Shah, R., Parker, J.S., Grissom, S.F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. Nat Genet *39*, 1507-1511.

Myer, V.E., and Young, R.A. (1998). RNA polymerase II holoenzymes and subcomplexes. J Biol Chem *273*, 27757-27760.

Nag, A., Narsinh, K., Kazerouninia, A., and Martinson, H.G. (2006). 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*, 1534-1544.

Nakamura, R., Takeuchi, R., Takata, K.-i., Shimanouchi, K., Abe, Y., Kanai, Y., Ruike, T., Ihara, A., and Sakaguchi, K. (2008). TRF4 Is Involved in Polyadenylation of snRNAs in Drosophila melanogaster. Molecular and Cellular Biology *28*, 6620-6631.

Nechaev, S., Fargo, D.C., dos Santos, G., Liu, L., Gao, Y., and Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science *327*, 335-338.

Ni, Z., Saunders, A., Fuda, N.J., Yao, J., Suarez, J.-R., Webb, W.W., and Lis, J.T. (2008). P-TEFb Is Critical for the Maturation of RNA Polymerase II into Productive Elongation In Vivo. Molecular and Cellular Biology *28*, 1161-1170.

Ni, Z., Schwartz, B.E., Werner, J., Suarez, J.R., and Lis, J.T. (2004). Coordination of transcription, RNA processing, and surveillance by P-TEFb kinase on heat shock genes. Mol Cell *13*, 55-65.

Niwa, M., Rose, S.D., and Berget, S.M. (1990). In vitro polyadenylation is stimulated by the presence of an upstream intron. Genes Dev *4*, 1552-1559.

Nunes, N.M., Li, W., Tian, B., and Furger, A. (2010). A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence. EMBO J *29*, 1523-1536.

Orozco, I.J., Kim, S.J., and Martinson, H.G. (2002). The poly(A) signal, without the assistance of any downstream element, directs RNA polymerase II to pause in vivo and then to release stochastically from the template. J Biol Chem *277*, 42899-42911.

Osheim, Y.N., Proudfoot, N.J., and Beyer, A.L. (1999). EM Visualization of Transcription by RNA Polymerase II: Downstream Termination Requires a Poly(A) Signal but Not Transcript Cleavage. Molecular Cell *3*, 379-387.

Perales, R., and Bentley, D. (2009). "Cotranscriptionality": the transcription elongation complex as a nexus for nuclear transactions. Mol Cell *36*, 178-191.

Peterlin, B.M., and Price, D.H. (2006). Controlling the elongation phase of transcription with P-TEFb. Mol Cell *23*, 297-305.

Phatnani, H.P., and Greenleaf, A.L. (2006). Phosphorylation and functions of the RNA polymerase II CTD. Genes Dev *20*, 2922-2936.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. Science *322*, 1851-1854.

Preker, P.J., Lingner, J., Minvielle-Sebastia, L., and Keller, W. (1995). The FIP1 gene encodes a component of a yeast pre-mRNA polyadenylation factor that directly interacts with poly(A) polymerase. Cell *81*, 379-389.

Price, D.H. (2008). Poised Polymerases: On Your Mark...Get Set...Go! Molecular Cell *30*, 7-10.

Proudfoot, N. (2004). New perspectives on connecting messenger RNA 3' end formation to transcription. Current Opinion in Cell Biology *16*, 272-278.

Proudfoot, N.J., Furger, A., and Dye, M.J. (2002). Integrating mRNA Processing with Transcription. Cell *108*, 501-512.

Ramanathan, P., Guo, J., Whitehead, R.N., and Brogna, S. (2008). The intergenic spacer of the Drosophila Adh-Adhr dicistronic mRNA stimulates internal translation initiation. RNA Biol *5*, 149-156.

Rhoads, R.E. (2009). eIF4E: new family members, new binding partners, new roles. J Biol Chem *284*, 16711-16715.

Richard, P., and Manley, J.L. (2009). Transcription termination by nuclear RNA polymerases. Genes Dev *23*, 1247-1269.

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

Rigo, F., and Martinson, H.G. (2008). Functional coupling of last-intron splicing and 3'-end processing to transcription in vitro: the poly(A) signal couples to splicing before committing to cleavage. Mol Cell Biol *28*, 849-862.

Rigo, F., and Martinson, H.G. (2009). Polyadenylation releases mRNA from RNA polymerase II in a process that is licensed by splicing. RNA, -.

Rogers, J., Fasel, N., and Wall, R. (1986). A novel RNA in which the 5' end is generated by cleavage at the poly(A) site of immunoglobulin heavy-chain secreted

mRNA. Mol Cell Biol *6*, 4749-4752.

Rondon, A.G., Mischo, H.E., Kawauchi, J., and Proudfoot, N.J. (2009). Fail-safe transcriptional termination for protein-coding genes in S. cerevisiae. Mol Cell *36*, 88-98.

Rosonina, E., Kaneko, S., and Manley, J.L. (2006). Terminating the transcript: breaking up is hard to do. Genes Dev *20*, 1050-1056.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular cloning : a laboratory manual, 2nd edn (Cold Spring Harbor, Cold Spring Harbor Laboratory Press).

Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A., and Burge, C.B. (2008). Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites. Science *320*, 1643-1647.

Schmid, M., and Jensen, T.H. (2008). The exosome: a multipurpose RNA-decay machine. Trends Biochem Sci *33*, 501-510.

Shatkin, A.J., and Manley, J.L. (2000). The ends of the affair: capping and polyadenylation. Nat Struct Biol *7*, 838-842.

Shi, Y., Di Giammartino, D.C., Taylor, D., Sarkeshik, A., Rice, W.J., Yates, J.R., 3rd, Frank, J., and Manley, J.L. (2009). Molecular Architecture of the Human Pre-mRNA 3' Processing Complex. Mol Cell *33*, 365-376.

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

Sonenberg, N. (2008). eIF4E, the mRNA cap-binding protein: from basic discovery to translational research. Biochem Cell Biol *86*, 178-183.

Spies, N., Nielsen, C.B., Padgett, R.A., and Burge, C.B. (2009). Biased Chromatin Signatures around Polyadenylation Sites and Exons. Mol Cell *36*, 245-254.

Steinmetz, E.J., Conrad, N.K., Brow, D.A., and Corden, J.L. (2001). RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. Nature *413*, 327-331.

Tabaska, J.E., and Zhang, M.Q. (1999). Detection of polyadenylation signals in human DNA sequences. Gene *231*, 77-86.

Tahirov, T.H., Babayeva, N.D., Varzavand, K., Cooper, J.J., Sedore, S.C., and Price, D.H. (2010). Crystal structure of HIV-1 Tat complexed with human P-TEFb. Nature *465*, 747-751.

Takagaki, Y., and Manley, J.L. (1994). A polyadenylation factor subunit is the human homologue of the Drosophila suppressor of forked protein. Nature *372*, 471-474.

Takagaki, Y., and Manley, J.L. (1997). RNA recognition by the human polyadenylation factor CstF. Mol Cell Biol *17*, 3907-3914.

Takagaki, Y., Ryner, L.C., and Manley, J.L. (1988). Separation and characterization of a poly(A) polymerase and a cleavage/specificity factor required for pre-mRNA polyadenylation. Cell *52*, 731-742.

Takagaki, Y., Ryner, L.C., and Manley, J.L. (1989). Four factors are required for 3'-end cleavage of pre-mRNAs. Genes Dev *3*, 1711-1724.

Takagaki, Y., Seipelt, R.L., Peterson, M.L., and Manley, J.L. (1996). The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. Cell *87*, 941-952.

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

Tian, B., Pan, Z., and Lee, J.Y. (2007). Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. Genome Res *17*, 156-165.

Tombacz, I., Schauer, T., Juhasz, I., Komonyi, O., and Boros, I. (2009). The RNA Pol II CTD phosphatase Fcp1 is essential for normal development in Drosophila melanogaster. Gene *446*, 58-67.

Vanacova, S., and Stef, R. (2007). The exosome and RNA quality control in the nucleus. EMBO reports *8*, 651-657.

Vanacova, S., Wolf, J., Martin, G., Blank, D., Dettwiler, S., Friedlein, A., Langen, H., Keith, G., and Keller, W. (2005). A new yeast poly(A) polymerase complex involved in RNA quality control. PLoS Biol *3*, e189.

Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S., and Meinhart, A. (2008). The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. Nat Struct Mol Biol *15*, 795-804.

Venkataraman, K., Brown, K.M., and Gilmartin, G.M. (2005). Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. Genes Dev *19*, 1315-1327.

Wahle, E. (1995). 3'-end cleavage and polyadenylation of mRNA precursors. Biochim Biophys Acta *1261*, 183-194.

Wahle, E., Lustig, A., Jeno, P., and Maurer, P. (1993). Mammalian poly(A)-binding protein II. Physical properties and binding to polynucleotides. J Biol Chem *268*, 2937-2945.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008a). Alternative isoform regulation in human tissue transcriptomes. Nature.

Wang, S.W., Stevenson, A.L., Kearsey, S.E., Watt, S., and Bahler, J. (2008b). Global role for polyadenylation-assisted nuclear RNA degradation in posttranscriptional gene silencing. Mol Cell Biol *28*, 656-665.

Wang, X., Lee, C., Gilmour, D.S., and Gergen, J.P. (2007). Transcription elongation controls cell fate specification in the Drosophila embryo. Genes Dev *21*, 1031-1036.

Weichs an der Glon, C., Ashe, M., Eggermont, J., and Proudfoot, N.J. (1993). Tat-dependent occlusion of the HIV poly(A) site. EMBO J *12*, 2119-2128.

Weichs an der Glon, C., Monks, J., and Proudfoot, N.J. (1991). Occlusion of the HIV poly(A) site. Genes Dev *5*, 244-253.

West, S., Gromak, N., Norbury, C.J., and Proudfoot, N.J. (2006). Adenylation and Exosome-Mediated Degradation of Cotranscriptionally Cleaved Pre-Messenger RNA in Human Cells. Molecular Cell *21*, 437-443.

West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature *432*, 522-525.

West, S., and Proudfoot, N.J. (2008). Human Pcf11 enhances degradation of RNA polymerase II-associated nascent RNA and transcriptional termination. Nucleic Acids Res *36*, 905-914.

West, S., and Proudfoot, N.J. (2009). Transcriptional Termination Enhances Protein Expression in Human Cells. Mol Cell *33*, 354-364.

West, S., Proudfoot, N.J., and Dye, M.J. (2008). Molecular dissection of mammalian RNA polymerase II transcriptional termination. Mol Cell *29*, 600-610.

Whitelaw, E., and Proudfoot, N. (1986). Alpha-thalassaemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to 3' end processing in the human alpha 2 globin gene. EMBO J *5*, 2915-2922.

Wu, C.H., Yamaguchi, Y., Benjamin, L.R., Horvat-Gordon, M., Washinsky, J., Enerly, E., Larsson, J., Lambertsson, A., Handa, H., and Gilmour, D. (2003). NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila.

Genes Dev *17*, 1402-1414.

Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.C., Dufour, M.E., Boulay, J., Regnault, B., Devaux, F., Namane, A., Seraphin, B., *et al.* (2005). Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell *121*, 725-737.

Yang, Q., Gilmartin, G.M., and Doublie, S. (2010). Structural basis of UGUA recognition by the Nudix protein CFIm25 and implications for a regulatory role in mRNA 3' processing. Proc Natl Acad Sci U S A *107*, 10062-10067.

Yang, Z., Yik, J.H., Chen, R., He, N., Jang, M.K., Ozato, K., and Zhou, Q. (2005). Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. Mol Cell *19*, 535-545.

Zarudnaya, M.I., Kolomiets, I.M., Potyahaylo, A.L., and Hovorun, D.M. (2003). Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. Nucleic Acids Research *31*, 1375-1386.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.-W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat Genet *39*, 1512-1516.

Zhang, Z., Fu, J., and Gilmour, D.S. (2005). CTD-dependent dismantling of the RNA polymerase II elongation complex by the pre-mRNA 3'-end processing factor, Pcf11. Genes Dev *19*, 1572-1580.

Zhang, Z., and Gilmour, D.S. (2006). Pcf11 is a termination factor in Drosophila that dismantles the elongation complex by bridging the CTD of RNA polymerase II to the nascent transcript. Mol Cell *21*, 65-74.

Zhou, M., Halanski, M.A., Radonovich, M.F., Kashanchi, F., Peng, J., Price, D.H., and Brady, J.N. (2000). Tat modifies the activity of CDK9 to phosphorylate serine 5 of the RNA polymerase II carboxyl-terminal domain during human immunodeficiency virus type 1 transcription. Mol Cell Biol *20*, 5077-5086.

Zhu, Y., Pe'ery, T., Peng, J., Ramanathan, Y., Marshall, N., Marshall, T., Amendt, B., Mathews, M.B., and Price, D.H. (1997). Transcription elongation factor P-TEFb is required for HIV-1 tat transactivation in vitro. Genes Dev *11*, 2622-2632.