Landscape, biogenesis and function of 3' UTR isoforms in *Drosophila*

by

Piero Sanfilippo

A Dissertation

Presented to the Faculty of the Louis V. Gerstner, Jr.

Graduate School of Biomedical Sciences,

Memorial Sloan Kettering Cancer Center

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

New York, NY

May, 2017

Eric C. Lai, PhD	Date
Dissertation Mentor	

To my grandmother

Anna Grzywacz,

whose tale of endurance and survival

will always be source of inspiration.

ABSTRACT

The sequencing of the human genome revealed that it encodes a similar number of protein-coding genes as the much simpler nematode *C. elegans*. This fact suggested that diversification of non-coding regulatory elements and alternative aspects of the transcriptome might underlie biological complexity of higher eukaryotes. The recent discovery of widespread regulated expression of mRNA isoforms that differ in the length of the 3' untranslated region (3' UTR) has added an additional layer of complexity to our understanding of how transcriptome diversity influences biology. To probe this phenomenon, I applied a combination of molecular, cellular and computational biology tools to the *Drosophila* model system. I have characterized the landscape of mRNA 3' UTR isoforms in several fruit fly species, delineated a role for a family of RNA binding proteins in the biogenesis of neural specific 3' UTR isoforms, and investigated the *in vivo* consequences of exceptionally long 3' UTRs on protein expression.

In the first section, I describe my efforts to capture the diversity and probe evolutionary conservation of 3' UTR isoform expression in *Drosophila*. Towards this aim, I developed a protocol to specifically sequence the 3' ends of mRNAs to compile a comprehensive atlas of 3' UTR isoforms. My atlas provides evidence that the accumulation of alternate length 3' UTR isoforms in *Drosophila* is much broader than earlier recognized and under significant tissue specific regulation. Evolutionary comparison in three species of *Drosophila* shows conservation tissue specific 3' end isoform expression. Furthermore, analysis of conservation and divergence of putative regulatory cis-elements uncovers strategies to preserve or change the tissue specific expression of 3' UTR isoforms.

In the second part of my thesis, I investigate the role of the Elav family of RNA binding proteins (RBPs) in regulating 3' UTR length in the nervous system of *D. melanogaster*. I probe the consequences of loss and gain of Elav on the 3' UTR landscape, and provide evidence that this factor by itself is not necessary for the expression of the vast majority of long neural 3' UTR isoforms. Instead, I propose that regulation of these isoforms might be a general property of the Elav family of RBPs mediated by regulation of polyadenylation site (pA) recognition.

In the final section, I investigate how long and differentially processed 3' UTRs can act as potent modulators of protein expression *in vivo* using the *elav* 3' UTR as a model. Genetic analysis shows that the 3' UTR of *elav* is involved in suppressing a previously unrecognized ubiquitous component of Elav expression. Furthermore, I provide evidence that ubiquitous non-neuronal Elav is under active microRNA pathway repression and is specifically modulated by action of miR-279/996 acting on the *elav* 3' UTR.

My work uncovers widespread and conserved expression of tissue specific mRNA 3' UTR isoforms in *Drosophila*. The biological consequence of this 3' end diversity remains unclear but my functional studies of complex 3' UTRs suggests that interesting biology awaits to be uncovered. The generation of a comprehensive atlas of 3' UTR isoforms in *Drosophila* provides an important tool to formulate both biological and mechanistic hypotheses that will further our understanding of the role of 3' UTR diversity in biology.

VITAE

Piero was born in Italy and moved permanently to the United States in 2005 to pursue his education.

ACKNOWLEDGMENTS

I wanted to thank all the members of the Lai lab, past and present. Foremost, I thank my mentor Dr. Eric Lai for giving me the opportunity to investigate multiple aspects of 3' UTR biology in *Drosophila* and the freedom to learn to be both a wet and dry scientist. I would like to especially thank Dr. Pedro Miura, Dr. Peter Smibert and Sol Shenker for discussions and insights on the study of APA. Special thanks to Dr. Sonali Majumdar, Alexandra Panzarino, Dr. Jiayu Wen and Dr. Peter Smibert for contributing experiments to my thesis. I wanted to thank Dr. Sonali Majumdar for being an incredible mentor in my transition to post-doctoral training and for her insightful feedback on this manuscript and process. I would like to thank my classmates, who are a tremendous group of smart and talented people, and made the GSK core course an invaluable formative experience. I would like to thank my committee members, Dr. Robert Benezra and Dr. Iestyn Whitehouse, who have provided me with important scientific insight throughout the years and foremost been incredibly supportive through the course of my PhD. Thanks to Dr. Stew Shuman for acting as Chair to my thesis defense committee and to Dr. Bin Tian for acting as external examiner and to both for insightful discussion during the thesis defense and for improvements to this manuscript. Thanks to the GSK graduate school for support and to everyone in the GSK office who have made this a pleasant journey. Finally, I would like to thank Dr. Ken Marians, who has provided me with incredible mentorship and support throughout the years and has shaped a terrific graduate program.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

3' RACE - rapid amplification of cDNA 3' ends

3' UTR - 3' untranslated region

3'-seq - 3' end sequencing

3X-MT - 3X RNA recognition motif mutant

5' UTR - 5' untranslated region

A - adenine

APA - alternative cleavage and polyadenylation

ARE - AU rich elements

cDNA - complementary DNA

CLIP - cross-linked immuno precipitation

CNS - central nervous system

CTD - carboxy terminal domain

DSE - downstream element

EST - expressed sequence tags

hnRNA - heterogeneous nuclear RNA

IHC - immunohistochemistry

kDa - kilodalton

L1 - 1st instar larva

LOF - loss of function

mRNA - messenger RNA

MT - mutant

mya - million years ago

pA - polyadenylation site

PAS- polyadenylation signal

pol II - RNA polymerase II

RIP - RNA immuno precipitation

RNA-seq - RNA sequencing

RPM - reads per million

RRM - RNA recognition motif

RT-PCR - reverse transcription-polymerase chain reaction

U- uridine

WT - wild type

CHAPTER 1: Introduction

A historical perspective

Polyadenylated RNAs in eukaryotes are characterized by discrete ends that are determined through the recognition of signals on nascent RNA, which leads to cleavage and polyadenylation of the transcript. The past decade has witnessed a great expansion in our understanding of the complexity of 3' end formation, given the realization that eukaryotic organisms express the majority of their genes as isoforms that differ at the 3' end in a regulated manner. However, the groundwork for our current understanding of this regulatory mechanism has been laid out much earlier, starting with the discovery in the 1960s of polyadenylated (poly(A)) RNA.

Discovery of poly(A) RNA

The first insight on the existence of polyadenylated RNA came from the discovery in 1960 of an enzyme from calf thymus nuclei that was able to catalyze the formation of poly(A) (Edmonds and Abrams, 1960). Subsequent to this discovery, several poly(A) polymerases were identified (Edmonds and Winters, 1976), however the cellular role of these enzymes was only recognized with the discovery of poly(A) messenger RNA (mRNA). The discovery of poly(A) mRNA started with the detection of long stretches of poly(A) from RNA in actively translated polysomes, through resistance of RNA digestion with RNases that cut at C, G or U nucleotides (Edmonds et al., 1971; Lim and Canellakis, 1970). The discovery of poly(A) on mRNA accelerated the pace of discoveries in the mRNA field as it could be leveraged as a tool to isolate the message from total RNA by oligo(dT) affinity chromatography (Aviv and Leder, 1972) and for priming for complementary DNA (cDNA) synthesis (Bank et al., 1972; Proudfoot et al., 1976; Verma et al., 1972). Early on only incredibly abundant mRNAs could be isolated for *in vitro* studies, hence the nature of the earliest characterized mRNAs such as *alpha-globin* (Mathews et al., 1971) and *immunoglobulin* (Brownlee et al., 1973), highly abundant cellular mRNAs, which were the focus of hundreds of studies until more sensitive technologies were developed.

Delineating 3' end formation

The deployment of oligo(dT) to generate cDNA ensured that at least the sequence at the 3' end of the mRNA could be synthesized and sequenced. The earliest report of such an effort comes from the sequencing of the 3' end of 6 eukaryotic mRNAs, which resulted in the identification of an AAUAAA hexamer positioned 14-20 nucleotides from the start of the poly(A) tail (Fig. 1.1) (Proudfoot and Brownlee, 1976). From this earliest observation, Proudfoot and colleagues hypothesized that the AAUAAA hexamer might be a conserved eukaryotic signal for RNA processing, a prediction that was proven as AAUAAA was confirmed to be the canonical polyadenylation signal (PAS) that mediates 3' end formation in all high eukaryotes analyzed to date (Wickens and Stephenson, 1984) (Beaudoing et al., 2000; Tian et al., 2005). Noted at the time was the large distance between the stop codon and the PAS, suggesting that non coding sequence separated the stop codon from the poly(A) tail (Proudfoot, 1976; Proudfoot and Longley, 1976) (Crawford et al., 1977). This non coding region was early defined and still known today as the 3' untranslated region (3' UTR). Subsequent studies, fostered by improving technology, confirmed the importance of the AAUAAA PAS for promoting 3' end cleavage and polyadenylation through mutational studies. For example, study of the viral SV40 PAS showed requirement of AAUAAA for polyadenylation and formation of a stable transcript (Fitzgerald and Shenk, 1981). These studies were followed by the discovery of additional upstream and downstream sequence elements that all together acted to define the site of cleavage and polyadenylation (Conway and Wickens, 1985; Gil and Proudfoot, 1984; Gil and Proudfoot, 1987; Hart et al., 1985). Concomitantly, a system for the *in vitro* polyadenylation of RNA was developed (Manley, 1983) leading to the discovery of the protein components that make up the cleavage and polyadenylation machinery as we know it today (Shi et al., 2009; Yang and Doublié, 2011). All these efforts centered on understanding the mechanism that generates the end of expressed genes through cleavage and polyadenylation. However, shortly after the discovery of the first PAS, evidence emerged of the recognition of multiple pAs, with the consequent realization that a gene can be expressed as isoforms that differ at the 3' end.

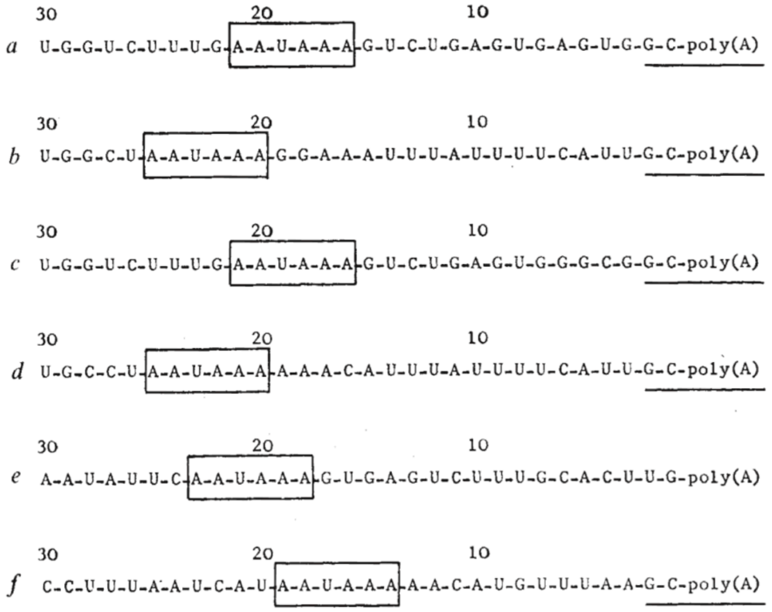

30 20 10
a U-G-G-U-C-U-U-U-G-A-A-U-A-A-A-G-U-C-U-G-A-G-U-G-A-G-U-G-G-C-poly(A)
30 20 10
b U-G-G-C-U-A-A-U-A-A-A-G-G-A-A-A-U-U-U-A-U-U-U-U-C-A-U-U-G-C-poly(A)
30 20 10
c U-G-G-U-C-U-U-U-G-A-A-U-A-A-A-G-U-C-U-G-A-G-U-G-G-G-C-G-G-C-poly(A)
30 20 10
d U-G-C-C-U-A-A-U-A-A-A-A-A-A-C-A-U-U-U-A-U-U-U-U-C-A-U-U-G-C-poly(A)
30 20 10
e A-A-U-A-U-U-C-A-A-U-A-A-A-G-U-G-A-G-U-C-U-U-U-G-C-A-C-U-U-G-poly(A)
30 20 10
f C-C-U-U-U-A-A-U-C-A-U-A-A-U-A-A-A-A-A-C-A-U-G-U-U-U-A-A-G-C-poly(A)

Figure 1.1 Discovery of the canonical AAUAAA PAS

Initial discovery of the AAUAAA hexamer adjacent to the poly(A) tail on cDNA from 6 eukaryotic mRNAs. The mRNAs in question are: a - rabbit *alpha-globin*, b - rabbit *beta-globin*, c - human *alpha-globin*, d - *beta-globin*, e - mouse *immunoglobulin light chain*, f - chicken *ovalbumin*. Boxed is the AAUAAA hexamer recognized to be common amongst all 6 mRNAs and immediately hypothesized to be a signal for RNA processing. Reprinted with permission (Proudfoot and Brownlee, 1976)

From one to many

It took only four years after the discovery of the canonical AAUAAA PAS (Proudfoot and Brownlee, 1976) to identify the first case of alternative cleavage and polyadenylation (APA). Early and colleagues were the first one to propose regulated APA as a mechanism controlling the switch from membrane bound to secreted form of the immunoglobulin *IgM* by recognition of a more proximal pA (Alt et al., 1980; Early et al., 1980). Strong evidence in support to this model came more than a decade later, with experiments showing that an increase in the core cleavage and polyadenylation factor Cstf-64 can control the switch from membrane bound to secreted forms of *IgM* by increasing recognition of the weaker proximal intronic pA (Takagaki et al., 1996). Similarly, the tissue specific switch in expression between Calcitonin and CGRP was shown to be mediated through the differential recognition of an intronic pA(Amara et al., 1984; Amara et al., 1982). Recognition of more than one pA in 3' UTRs was also discovered early on, for example with the observation of multiple 3' UTR isoforms of the *DHF reductase* gene (Setzer et al., 1982). By the end of the 1990s approximately 100 genes were described to undergo APA, many of which expressed isoforms in a regulated manner depending on biological context (Edwalds-Gilbert et al., 1997). EST data analysis (Beaudoing et al., 2000; Gautheret et al., 1998; Yan and Marr, 2005) and most recently deep sequencing of polyadenylated RNA 3' ends (Derti et al., 2012; Jan et al., 2011; Lianoglou et al., 2013; Ozsolak et al., 2010; Ulitsky et al., 2012) have shown that the majority of eukaryotic genes express isoforms with differential 3' ends in a regulated manner. In the past decade efforts have gone into deepening our understanding of the complexity and biological context of 3' end isoforms expression landscapes as well as the regulatory mechanism involved in the regulated expression of these isoforms. What follows is a short review of our current understanding of cleavage and polyadenylation and the regulation mechanisms underlying the expression of different 3' end species.

mRNA 3' end formation

3' end formation is a co-transcriptional event

As with other RNA processing events such as nascent RNA splicing (Kornblihtt, 2005), cleavage and polyadenylation is a co-transcriptional event. This conclusion arises from several studies that showed both necessity of the carboxy-terminal domain (CTD) of RNA polymerase II (RNA pol II) for pA recognition as well as interaction of core cleavage and polyadenylation machinery (pA machinery) components with transcription factors and activators at the initiation complex at the promoter. Early evidence of coupling with transcription came from *in* vitro observations (Mifflin and Kellems, 1991) followed by experimental evidence showing that deletion of the CTD of RNA pol II in cells leads to loss of pA recognition on an SV40 pA and that the core pA machinery components CPSF and CstF bind the CTD and co-purify with the RNA pol II holoenzyme complex (McCracken et al., 1997). Concomitantly, CPSF was shown to bind the transcription factor TFIID and in an *in* vitro reconstituted system TFIID was shown to mediate the recruitment of CPSF to the pre-initiation complex where upon onset of elongation, CPSF dissociates from TFIID and becomes engaged with the CTD of the elongating RNA pol II (Dantonel et al., 1997). Later studies have also shown that Ser2 phosphorylation of the CTD is specifically coupled with cleavage and polyadenylation, increasing in proximity of the 3' end of the transcriptional unit and related to RNA pol II pausing around pAs (Davidson et al., 2014). These and other results have been the rationale for subsequent studies into the role of transcription factors and transcription elongation in the regulation of APA. More recently, binding of U1 snRNP to nascent RNA has been proposed to inhibit recognition of the majority of pA sites found in introns. Knock down of U1 snRNP was shown to lead to ectopic recognition of pAs in introns close to the transcription start site (Kaida et al., 2010), suggesting that the previously observed excess of U1 snRNP might be implicated in silencing intronic pAs genome-wide ensuring proper recognition of distal pAs. This mechanism, termed telescripting might also play a role in regulating recognition of alternative intronic and proximal pAs. For example, mild suppression of U1 snRNP in PC12 cells recapitulates aspects of mRNA shortening observed during neuronal activation (Kaida et al., 2010). It is possible that transcriptional activation, which follows neuronal activation, might create a shortage of available U1 snRNP, resulting in the unmasking of alternative proximal pAs (Kaida et al., 2010). Finally, it appears that cleavage and polyadenylation can influence the last step of transcription, termination and dissociation of RNA pol II from the DNA template, an aspect of transcription still poorly understood and under active investigation (Proudfoot, 2016).

Polyadenylation signals

From the earliest discovery of AAUAAA as the canonical PAS in metazoans (Proudfoot and Brownlee, 1976), a multitude of studies have provided evidence for additional PAS variants as well as multiple cis-elements that synergize with the PAS to mediate cleavage and polyadenylation. Studies of pAs of many other organisms, including plants and yeast, have shown some generalities to the structure of pA sites, characterized by A/AU-rich PAS and U/GU-rich elements flanking the PAS (Nunes et al., 2010; Tian and Graber, 2012). In metazoans, AAUAAA and AUUAAA are the top used PASs, shown to be of much higher efficiency than other variants in studies of point mutants of the SV40 pA signal (Sheets et al., 1990; Wilusz et al., 1989). The SV40 pA signal has played an important role in our understanding of 3' end formation, starting with early deletion analyses that confirmed AAUAAA as the PAS (Fitzgerald and Shenk, 1981). Other studies have shown differential PAS variant usage in different tissues, for example noting that AAUAAA is much less prevalent upstream of 3' ends of mRNA generated in testis (MacDonald and Redondo, 2002), suggesting that PAS identity plays a regulatory function in tissue specific signal recognition. In addition to the PAS, it was soon evident that pAs had characteristic GU/U rich elements downstream to the cleavage site (DSE) (Conway and Wickens, 1985; Gil and Proudfoot, 1984; Gil and Proudfoot, 1987; Hart et al., 1985; Salisbury et al., 2006). More recent analysis of DSEs in human genes pAs confirmed presence of U-rich elements in 80% of cases (Zarudnaya et al., 2003). Further studies have implicated additional cis-elements found upstream to the PAS, such as UGUA, which can at times act as a PAS and recruit CFI, a component of the pA machinery (Moreira et al., 1995; Venkataraman et al., 2005; Zhao et al., 1999). Motifs such as these and others are likely to serve as platforms for recruiting RBPs and other factors that can regulate the efficiency of 3' end formation.

Cleavage and polyadenylation machinery

The protein factors involved in the recognition of the PAS and accessory elements began to emerge with the developmental of an *in vitro* system that recapitulated cleavage and polyadenylation (Manley, 1983). Differently from splicing, 3' end processing is mediated through recognition of RNA signals by protein factors alone, which recognize signals on nascent RNA to mediate cleavage and polyadenylation (Colgan and Manley, 1997; Shi and Manley, 2015). Three proteins form the core pA machinery: CPSF, CstF and CFI. The cleavage and polyadenylation specificity factor (CPSF), a multi subunit complex, is involved in both the recognition of the PAS hexamer as well as the catalysis of the cleavage reaction (Bienroth et al., 1991; Chan et al., 2014; Ryan et al., 2004; Takagaki et al., 1988). The second member of the core of the machinery, the cleavage and stimulation factor (CstF) is involved in recognizing the U/GU rich DSE (Takagaki et al., 1990). Finally, CFI recognizes UGUA sequences which are found most often upstream of the core PAS (Gilmartin and Nevins, 1991). CstF and CFI help stabilize CPSF, increasing the efficiency of signal recognition (Rüegsegger et al., 1996). These core factors recruit a plethora of additional proteins, some of which are considered part of the pA machinery (symplekin, poly(A) polymerase, CFII) (Fig. 1.2) as well as other RBPs, in a co-transcriptional manner (Hirose and Manley, 2000). It is likely that with additional studies we will expand our concept of poly(A) machinery and related regulatory factors, as studies in yeast have identified many more proteins involved in mediating 3' end formation, which are likely conserved (Zhao et al., 1999). Recent studies in human cells have identified additional proteins such as Wdr33 and Rbbp6 as part of the core pA machinery (Shi et al., 2009). These as well as all other core pA machinery components are deeply conserved amongst yeast, fly and human, underscoring the important role of accurately defining the 3' end of transcripts (Table 1.1) (Mount and Salz, 2000; Shi et al., 2009).

D. melanogaster	***H. sapiens* Protein**	***S. cerevisiae***	**Reference**
General			
hrg (CG9854)	Poly(A) polymerase	Pap1	*Mount and Salz, 2000*
PAbp (CG5119)	Poly(A) binding protein	Pab1	*Mount and Salz, 2000*
Pabp2 (CG2163)	Poly(A) binding protein II	Sgn1	*Mount and Salz, 2000*
CG4612	Poly(A) binding protein	Pab1	*Mount and Salz, 2000*
CPSF			
Cpsf160 (CG10110)	CPSF-160 kD	Cft1	*Mount and Salz, 2000*
Cpsf100 (CG1957)	CPSF-100 kD	Ysh1	*Mount and Salz, 2000*
Cpsf73 (CG7698)	CPSF-73 kD	Ysh1	*Mount and Salz, 2000*
IntS11 (CG1972)	CPSF-73–kD variant	Ysh1	*Mount and Salz, 2000*
IntS9 (CG5222)	Close CPSF-100/73	Ysh1	*Mount and Salz, 2000*
Clp (CG3642)	CPSF-30 kD	Yth1	*Mount and Salz, 2000*
Fip1 (CG1078)	Fip1	Fip1	
Wdr33 (CG1109)	Wdr33	Pfs2	*Shi et al, 2009*
CstF			
Su(f) (CG17170)	CstF-77 kD	Rna14	*Mount and Salz, 2000*
CstF-64 (CG7697)	CstF-64 kD and CSTF2T	Rna15	*Mount and Salz, 2000*
CstF-50 (CG2261)	CstF-50 kD	unknown	*Mount and Salz, 2000*
Sym (CG2097)	Symplekin	Pta1	*Mount and Salz, 2000*
Cleavage factor I			
CG3689	25-kD (CPSF5)	unknown	*Mount and Salz, 2000*
CG7185	68-kD/59-kD (CPSF6/7)	unknown	*Mount and Salz, 2000*
Cleavage factor II			
Pcf11 (CG10228)	Pcf11	Pcf11	*Mount and Salz, 2000*
cbc (CG5970)	Clp1	Clp1	
Additional pA factors			
snama (CG3231)	Rbbp6	Mpe1	*Shi et al, 2009*
Pp1-87B (CG5650)	Pp1 alpha	Glc7	*Shi et al, 2009*
Pp1-13C (CG9156)	Pp1 beta	Glc7	*Shi et al, 2009*

Table 1.1 Orthologs of proteins of the pA machinery in fly, human and yeast.

Orthologous genes of *H. sapiens* proteins of the core pA machinery in *D. melanogaster* and *S. cerevisiae*. Evidence for orthology is provided in the study as referenced or obtained through orthoDB and FlyBase.

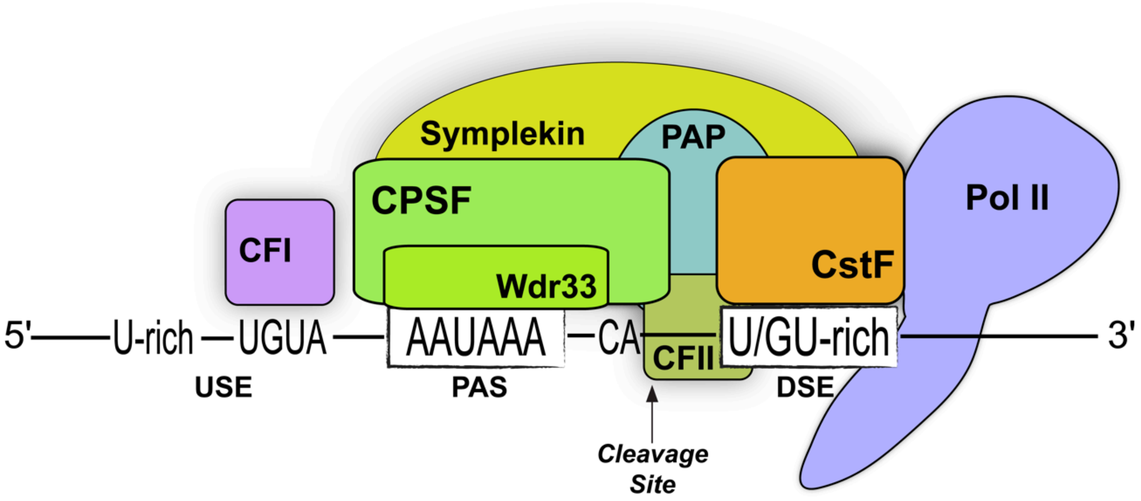


Figure 1.2 The cleavage and polyadenylation machinery at the poly(A) site. A multi-protein complex cleaves the mRNA 3' terminus and adds an untemplated poly(A) tail. RNA polymerase II recruits pA machinery factors and enhances the cleavage reaction, and is thus considered part of the pA machinery. The cleavage and polyadenylation specificity factor (CPSF) complex recognizes the PAS, in part through Wdr33 (Schönemann et al., 2014), and catalyzes endonucleolytic cleavage. The Cleavage stimulation factor (CstF) complex aids recognizes the downstream sequence element (DSE). Cleavage factor I (CFI) recognizes UGUA motif often present upstream to the PAS. Finally, poly(A) polymerase (PAP) catalyzes poly(A) formation. Other factors such as symplekin and CFII are recruited by these core components. Additional auxiliary factors have been omitted for clarity, as have other physical connections that the pA machinery has with the transcription complex and splicing factors. USE, upstream sequence element. Reprinted with permission with modifications (Miura et al., 2014).

Methods to identify 3' ends of polyadenylated RNA

One of the reasons for the increased interest in understanding the scope and regulation of 3' end formation in the past decade lies in the development of modern sequencing technologies, starting with sequencing of expressed sequence tags (ESTs) and ending with the massive parallel sequencing of the 3' ends of transcripts. Different tools can be adopted to query 3' end formation, starting with Northern blotting technology, which was involved in many of the early observations of APA (Edwalds-Gilbert et al., 1997) to reverse transcription-polymerase chain reaction (RT-PCR) and rapid amplification of cDNA ends (3' RACE) (Scotto-Lavino et al., 2006). The recent advent of next generation sequencing with the development of RNA sequencing (RNA-seq) and specialized protocols to sequence 3' ends, have greatly advanced our ability to probe the complexity of 3' end formation (Elkon et al., 2013; Hoque et al., 2013; Jan et al., 2011; Pelechano et al., 2012; Shepard et al., 2011). These methods have demonstrated that 3' end formation is a highly regulated and complex RNA processing event (Derti et al., 2012; Hoque et al., 2013; Jan et al., 2011; Lianoglou et al., 2013; Mangone et al., 2010; Ozsolak et al., 2010; Smibert et al., 2012); these and future tools will ensure that we will gain a better picture of the regulatory mechanisms that underlie the formation of complex landscapes of 3' end isoforms expression (Fig. 1.3).

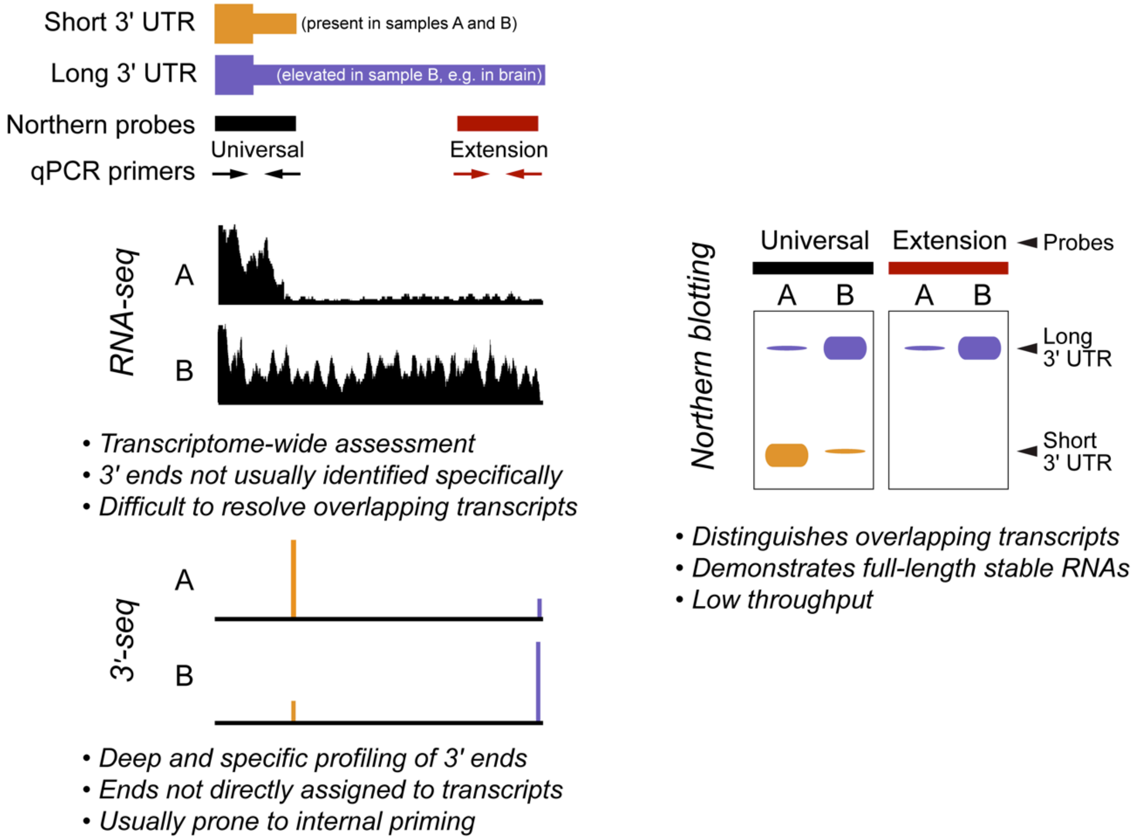

Short 3' UTR
(present in samples A and B)
Long 3' UTR
(elevated in sample B, e.g. in brain)
Northern probes
Universal
Extension
qPCR primers
RNA-seq
A
B
• Transcriptome-wide assessment
• 3' ends not usually identified specifically
• Difficult to resolve overlapping transcripts
3'-seq
A
B
• Deep and specific profiling of 3' ends
• Ends not directly assigned to transcripts
• Usually prone to internal priming
Northern blotting
Universal
Extension
Probes
A
B
A
B
Long 3' UTR
Short 3' UTR
• Distinguishes overlapping transcripts
• Demonstrates full-length stable RNAs
• Low throughput

Figure 1.3 Experimental approaches to study 3' end formation.
The schematics at top left depict a model gene that expresses alternative 3' end isoforms. Isoforms can be assayed at the level of individual genes using Northern blotting or RT-PCR. The right panel illustrates model data for Northern measurements of APA isoforms. The middle and bottom left panels illustrate simulation of genome-wide approaches. RNA-seq data provides information of expression of the whole transcription unit, while 3' end specific methods such as 3'-seq capture and sequence the 3' termini of poly(A) RNA. Reprinted and modified with permission (Miura et al., 2014).

Alternative Cleavage and Polyadenylation

APA is the rule, not the exception

The past decade has witnessed an explosion in our understanding of transcript 3' end diversity. Our understanding went from approximately ~100 or so genes expressing multiple 3' end isoforms at the end of the 1990s (Edwalds-Gilbert et al., 1997) to the realization that more than 50%, and often more than 70% of all genes in eukaryotes express multiple 3' end isoforms via APA (Shi, 2012).

3' end isoform accumulation is highly regulated

The earliest genome-wide analyses of 3' end diversity using EST data immediately hinted at significant tissue specific regulation of APA isoforms expression. Tissue coming from male gonads as well as ones of neural origin, tended to have different patterns of 3' end formation compared to other somatic tissues (MacDonald and Redondo, 2002; Tian et al., 2005; Zhang et al., 2005). This has proven true with the adoption of RNA-seq and 3' end sequencing technology, which has shown that neural tissues tend to express more long 3' UTR isoforms of mRNA (Lianoglou et al., 2013; Miura et al., 2013; Smibert et al., 2012) as well as shorter 3' UTRs in testis (Li et al., 2016; Smibert et al., 2012). The expression of long neural isoforms has so far proven to be conserved from fly to human, however the expression of the shortest 3' UTRs in testis is not always seen, as in the case of zebrafish, where ovary is instead found to express such isoform length profile (Ulitsky et al., 2012). The biological role that these tissue specific global patterns of 3' end expression play remains mysterious and is ripe subject for investigation.

Beyond tissue specific patterns of 3' end expression, the past few years have seen several studies illustrating shifts in the expression of 3' isoforms in specific biological contexts. For example, early studies that used microarray technology showed changes in the expression of different length 3' UTR isoforms upon T cell activation (Sandberg et al., 2008), cell differentiation (Ji et al., 2009) and neuronal activation (Flavell et al., 2008). These studies have been joined by many others (Hollerer et al., 2016; Hu et al., 2017; Jia et al., 2017; Lianoglou et al., 2013) which are starting to show that modulation of 3' UTR isoform length might be a general part of a gene expression program. Finally, the expression of 3' end isoforms has been seen to change in pathological conditions, such as cancer (Erson-Bensan and Can, 2016; Masamha et al., 2014; Mayr and Bartel, 2009; Miles et al., 2016), where genes tend to express shorter 3' UTR isoforms. The implication of different APA isoforms in disease is quickly growing with the discovery of changes of 3' end isoforms in heart disease (Creemers et al., 2016), in OPMD muscular dystrophy (Jenal et al., 2012) and the identification of specific 3' UTR length isoforms correlated with pathological expression of α-synuclein protein in Parkinson's disease (Rhinn et al., 2012).

Regulatory mechanisms of alternative 3' end isoform expression

The expression of 3' end isoforms generated through APA can be modulated through different mechanisms, the majority of which involves the differential recognition of pAs. This mode of regulation was observed early on with the realization that modulation of the levels of CstF-64 mediates the differential recognition of an intronic pA (Takagaki et al., 1996). The recognition that RBPs can influence pA recognition came early with the discovery that U1A can inhibit cleavage and polyadenylation of its cognate mRNA through binding upstream of the PAS (Boelens et al., 1993) and that *Drosophila* Elav mediates neural specific recognition of intronic pA sites (Soller, 2003). These two modes of APA regulation have emerged as dominant ones, with several protein factors involved in modulating the recognition of pA sites through either differential availability of the pA machinery or through binding in proximity to the pA, in both an activating or repressive manner. In addition to these models of regulation, the promoter has been shown to modulate the recognition of PAS either through differential recruitment of pA machinery components and RBPs or through putative modulation of pol II elongation rate.

Role of the pA machinery

Recent genome-wide perturbation studies coupled with cross-linking immuno precipitation (CLIP) and 3' end sequencing of components of the pA machinery have revealed a complex role that these factors play in regulating APA (Hwang et al., 2016; Li et al., 2015; Martin et al., 2012; Yao et al., 2013). Loss of CstF-64 was confirmed to lead to lengthening of 3' UTRs, however the effect was significant only upon depletion of both CstF-64 and the paralog tau-CstF-64 (Yao et al., 2013). Other pA factors, such as components of CFII and the novel component RBBP6 have also been shown to lead to lengthening of 3' UTRs upon their depletion (Li et al., 2015). Of note is the proposed model by which RBBP6 modulates 3' UTR length, by modulation of pA site recognition but possibly also through stabilization of AU-rich elements in 3' UTRs (Di Giammartino et al., 2014). Conversely, other pA factors have been shown to lead to 3' UTR shortening upon their depletion, such as CFI(Hardy and Norbury, 2016; Li et al., 2015) and PABPN1 (Jenal et al., 2012). These studies implicate the pA machinery as an important regulator of APA, as its core and accessory components could be regulated to achieve changes in APA, as seen during IgM maturation (Takagaki et al., 1996), cell differentiation (Ji and Tian, 2009) and in patient tumor samples (Masamha et al., 2014). However, the signals and regulatory mechanisms that lead to modulation of levels of pA machinery components are currently not well understood.

Role of RBPs

RBPs are emerging as major regulators of 3' end isoform expression (Erson-Bensan, 2016; Shi and Manley, 2015; Zheng and Tian, 2014). CLIP studies have shown that binding of RBPs around pAs influences recognition in a position dependent manner. This is analogous to observations of positional binding of RBPs around alternative spliced exons (Licatalosi et al., 2008; Rot et al., 2017).

Studies have shown that close binding to pA sites tends to be repressive while further binding downstream or upstream of a pA site promotes recognition of the site, as shown for NOVA (Licatalosi et al., 2008), Mbnl (Batra et al., 2014), FUS (Masuda et al., 2016) and TDP-43 (Rot et al., 2017). The mechanism that leads to changes in APA isoform expression is often inferred to involve modulation of pA signal recognition, although careful experimental validation of this model is often missing. Some factors have been characterized more carefully using biochemical methods. For example a 55 kDa protein was found to bind upstream to the PAS (Brackenridge and Proudfoot, 2000) and inhibit cleavage. Similarly, Hu RBP binding upstream or downstream pA sites inhibits both cleavage and polyadenylation in HeLa extracts (Zhu et al., 2007). This could in part occur through inhibition of CstF-64 binding, as Hu is found to interact with CstF-64 and Hu over-expression appears to inhibit binding of CstF-64 *in vitro* (Zhu et al., 2007). Similarly hnRNPH was found to compete with CstF-64 binding at DSEs resulting in the choice of more downstream pAs (Nazim et al., 2017). On the other hand, RBPs such as αCP, have been found to enhance isoform expression through putative specific binding a short distance upstream of specific pAs (Ji et al., 2013). These and other studies point to a complex network or RBPs which depending on binding position, cellular RBP levels as well as levels of pA machinery components, shape specific landscapes of 3' UTR isoforms expression.

Role of transcriptional control

Given the co-transcriptional nature of many RNA processing events, transcriptional regulation has been shown to regulate different aspects of alternative transcript processing. Alternative cleavage and polyadenylation, as shown for splicing (Cramer et al., 1999; La Mata et al., 2003; Naftelberg et al., 2015), can also be regulated at the transcriptional level. Modulation of transcription elongation has been shown to affect the recognition of pAs. For example, analysis of *Drosophila* carrying a mutation in RNA pol II, which causes a lower intrinsic elongation rate, leads to an increase in the selection of a proximal pA of the gene *polo* (Pinto et al., 2011). Given that the spacing between pAs can influence their recognition (Galli et al., 1987; Galli et al., 1988), changes in RNA pol II elongation rate could have significant effects on pA recognition more broadly. Other evidence that modulation of elongation rate might play a role in the differential recognition of pAs come from studies of elongation factors, such as ELL2, which has been shown to mediate recruitment of CstF-64 to RNA pol II during elongation. This interaction can mediate the recognition of more proximal pAs, such as the famous intronic pA of the immunoglobulin heavy chain locus (Martincic et al., 2009). Interestingly the levels of this elongation factor appear to increase in plasma cells and could hence be directly involved in regulating pA site recognition in this system. The promoter sequence has also been shown to regulate recognition of alternative pAs (Ji et al., 2011), as swapping different promoters upstream of a reporter system lead to differential pA recognition (Ji et al., 2011), an effect similar to what observed for the selective inclusion of cassette exons (Kornblihtt, 2005). Specific transcription activators that could mediate such regulation have been identified (Nagaike et al., 2011) and RBPs, such as Elav have been proposed to be directly recruited to a special class of promoters characterized by paused RNA pol II to mediate neural lengthening of 3' UTRs (Oktaba et al., 2015). Other mechanisms involving transcription could possibly arise. For example, recent reports have shown that m6A methylation correlates with the expression of short 3' end isoforms (Molinie et al., 2016), and given that RNA methylation can occur in the nucleus (Liu et al., 2014), co-transcriptional m6A methylation could be yet another mechanism that modulates pA recognition.

Biological consequences of alternative 3' end transcript expression

The major consequence of APA is the expression of mRNA isoforms with different 3' UTRs. This has immediate consequences, as the 3' UTR acts as the hub for post-transcriptional regulation and changes in its sequence will have an impact on the information content and regulatory capacity of the 3' UTR (Matoulkova et al., 2014) (Fig. 1.4).

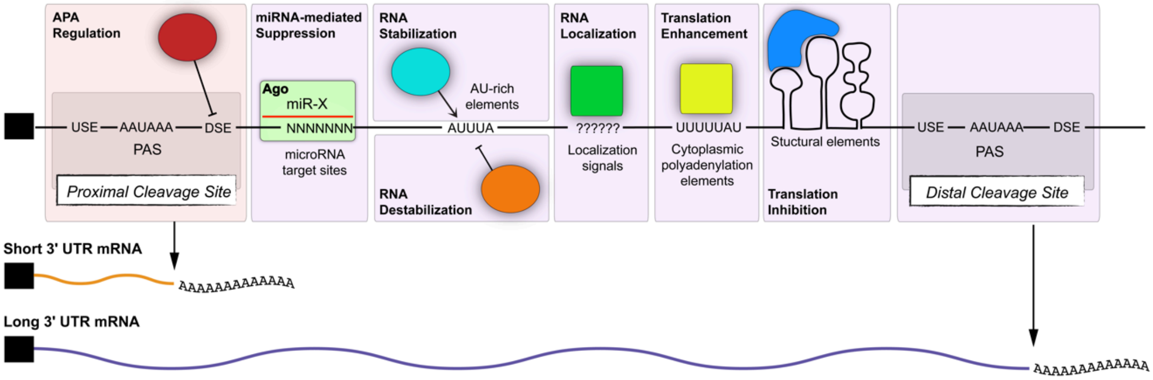


Figure 1.4 The 3' UTR is a hub for post-transcriptional regulation
The sequence of the 3' UTR encodes a variety of regulatory cis-elements which orchestrate the binding of different RBPs. These interactions mediate various aspects of post-transcriptional regulation, such as transcript stability, translation and localization, as well as 3' end processing. Differential choice of pAs will then change the information content of a 3' UTR. Reprinted and modified with permission (Miura et al., 2014).

Normal biology

Given that the 3' UTR can affect stability, translation and localization of mRNAs, APA has been shown to affect all of these processes. The role of APA on modulation of transcript stability and translation has been widely documented. Early reports tended to consider extension of 3' UTRs as a repressive event, given the enrichment of microRNA (miRNA) binding sites in extended regions of 3' UTRs (Sandberg et al., 2008) and evidence of lower translational efficiency of longer isoforms (Mayr and Bartel, 2009). However, recent studies present a much more complex picture. Examples abound of positive and negative effects of both long and short 3' UTR isoforms on protein expression, such as the higher translation of the long isoform of α-synuclein (Rhinn et al., 2012) or the higher translation upon mTOR activation of transcripts with short 3' UTRs (Chang et al., 2015). Genome-wide probing into the stability of transcripts, suggests that, at least in murine 3T3 cells, 3' UTR length has only a slight effect on transcript stability (Spies et al., 2013). Finally, given that functional conserved miRNAs tend to be positioned at the end of 3' UTRs (Grimson et al., 2007), studies have shown that conserved miRNA binding sites are enriched immediately upstream of proximal pAs, hence 3' UTR shortening might lead to potentiation of miRNA targeting of short 3' UTR isoforms (Hoffman et al., 2016). These reports highlight the complexity of post-transcriptional regulation of mRNA to which APA introduces an additional level of regulation, likely to be specific on cellular and physiological context.

Specific subcellular localization of mRNAs is rapidly emerging as a general aspect of mRNA biology, with potential significant impacts on the regulation and biological output of gene expression, as exemplified by the discovery of hundreds of transcripts localized to specific subcellular compartments (Lécuyer et al., 2007). The localization of transcripts appears to be in part regulated by alternative RNA processing, as mRNA isoforms that differ at the 3' end have been shown to differentially localize in the cell. For example, the long 3' UTR isoform of BDNF has been shown to localize to dendrites where it is locally translated (An et al., 2008b). Specific subcellular localization of APA isoforms has also been found, with examples of 3' end isoforms which are retained in the nucleus and fail to be exported to the cytoplasm (Avendaño-Vázquez et al., 2012) and the general observation in myoblasts that longer 3' UTR isoforms tended to accumulate in the insoluble cell fraction (Wang et al., 2012). Finally, events which result in a change in the C-terminus ends of a protein through recognition of pAs in introns, have been shown to have wide biological effects, such as the early reports of a role in mediating conversion between forms of membrane and soluble IgM (Alt et al., 1980). These and other mechanistic consequences appear to particularly play a role in the nervous system (Miura et al., 2014). Future studies are bound to uncover many more functional consequences of the expression of different APA isoforms in the nervous system and beyond.

Aberrant biology

An important role of 3' UTR biology in many pathologies has emerged in the past two decades (Conne et al., 2000). Mutations in 3' UTR elements that mediate loss of gene expression were uncovered early on, with the discovery of a form of α-thalassemia which is caused by a mutation in the AAUAAA PAS of the *α-globin* gene, causing a decrease in expression levels and disease (Higgs et al., 1983). The role of broad changes in patterns of APA isoforms expression are not yet understood, however such changes have been observed in many pathologies, such as cancer
(Erson-Bensan and Can, 2016; Masamha et al., 2014; Mayr and Bartel, 2009), heart disease (Creemers et al., 2016), muscular dystrophy (Jenal et al., 2012) and others and have become active field of clinical investigation (Hollerer et al., 2014). Given the broad impact that 3' UTR mediated post-transcriptional regulation has on the life of mRNA, global and specific changes in the expression of 3' UTR isoforms and mutations that create or impair binding of regulators will certainly continue to emerge as both contributing and causative agents of disease.

CHAPTER 2: Landscape and evolution of terminal 3' UTR isoforms in *Drosophila*

Summary

The *Drosophila* transcriptome is one of the best annotated eukaryotic transcriptomes, however it still lacks detailed annotation of 3' end isoforms. Given recent reports of strong regulation of specific 3' UTR isoforms, underlying the importance of this process, we set to generate a comprehensive atlas of the end of polyadenylated RNA in *Drosophila*. We report here the most comprehensive annotation of polyadenylated ends of RNA in *Drosophila* to date and provide evidence of expression of multiple 3' end isoforms for two thirds of all genes. Analysis of tissue specific patterns of 3' UTR isoforms show much broader tissue specific regulation, with hundreds of genes having the propensity to express longer 3' UTR isoforms in head and shorter ones in testis. We detect conservation of head specific expression of longer 3' UTR isoforms in two other species and uncover evidence of faster rate of evolution of testis specific events. Finally, we present a deep analysis of pAs conservation and provide evidence that cis-element changes play an important part in the evolution of patterns of 3' UTR isoform expression in *Drosophila*.

Introduction

The adoption of the fruit fly *Drosophila melanogaster* as a model organism has proven incredibly successful, (Bellen et al., 2010; Letsou and Bohmann, 2005), leading to many important breakthroughs in biology, starting from the discovery of Thomas Hunt Morgan that genes are inherited through chromosomes (Morgan, 1910). In addition, *Drosophila* turned out to be relevant to understanding human disease, as 50% of the disease causing genes in humans are conserved in the fly (Gonzalez, 2013; Rubin et al., 2000). More recently, *Drosophila* has been instrumental in the rise of modern genomics. The fly genome was the first large metazoan genome to be sequenced using whole-genome shotgun sequencing, with the publication in 2000 of the 120 Mb euchromatic genome of *D.* melanogaster (Adams et al., 2000; Myers et al., 2000). This achievement set the ground for the use of WGS to sequence larger genomes, including the human genome (Venter et al., 2001).

Since the sequencing and assembly of the first *Drosophila* genome, 5 more releases have been generated, substantially improving the current genome with the most recent Release 6, which assembles in heterochromatic repetitive regions (Hoskins et al., 2015). In addition to the *D. melanogaster* genome, 11 additional genomes of related fly species were published in 2007 (Clark et al., 2007), opening *Drosophila* as a model for the study of evolutionary genomics. Given the early availability of these tremendous resources, significant effort has been placed to thoroughly annotate the *Drosophila* genome, making it one of the best annotated eukaryotic genomes. With every improvement in technology came a revision of the *Drosophila* transcriptome, starting with sequencing of EST libraries (Rubin et al., 2000), gene microarrays (Hild et al., 2003) and full-genome tiling arrays (Manak et al., 2006; Oliver, 2006), next generation mRNA sequencing through coordinated efforts of the modENCODE project (Brown et al., 2014; Graveley et al., 2011), total RNA sequencing (Duff et al., 2015; Westholm et al., 2014) and small RNA sequencing (Berezikov et al., 2011; Wen et al., 2014). These combined efforts have uncovered a tremendous complexity of the *Drosophila* transcriptome, exemplified in genes such as *Dscam1* and 46 others, that in *Drosophila* express more than 1000 isoforms each (Brown et al., 2014). The employment of RNA-seq has considerably improved our understanding of the transcribed regions of the genome as well as the dynamics of alternative splicing.

Despite the improvements in the *Drosophila* transcriptome through analysis of RNA-seq datasets, these datasets fail to provide an accurate picture of the diversity at the extremes of the transcript, where the use of alternative transcription start sites and alternative polyadenylation sites (pAs) diversify 5' UTRs and 3' UTRs respectively (Steijger et al., 2013). To define the ends of transcripts, more specialized techniques need to be adopted (Sun et al., 2012; Takahashi et al., 2012). The 5' definition of transcripts has been explored to a greater degree, with the deployment of CAGE sequencing as a part of the modENCODE project (Brown et al., 2014; Graveley et al., 2011), however the 3' ends of transcripts in *Drosophila* remain yet to be characterized genome-wide. This in spite of recent reports of wide alternative usage of pAs in *Drosophila* which leads to the expression of alternative 3' end isoforms in a tissue specific manner (Hilgers et al., 2011b; Smibert et al., 2012). The use of protocols to capture the expression of 3' end isoforms in other eukaryotes has uncovered extensive diversity, with the overall conclusion that the majority of genes in higher metazoans are expressed as isoforms that differ at the 3' end (Shi, 2012). However, at this time only a quarter of all genes have gene models indicative of heterogeneous 3' ends in *Drosophila*.

The major impact of differential recognition of pAs is the expression of mRNA isoforms with different length 3' UTRs. This occurs mainly through the recognition of pAs scattered throughout the 3' UTR through alternative cleavage and polyadenylation (APA) (Edwalds-Gilbert et al., 1997; Proudfoot, 2011; Tian and Manley, 2017). Given that 3' UTRs are the main hub for post-transcriptional regulation, APA has a broad impact on the function of the affected transcripts (Tian and Manley, 2017). Since the expression of different 3' UTR isoforms appears to be highly regulated with cellular (Sandberg et al., 2008), tissue (Shi, 2012) and developmental state (Ji et al., 2009), as well as in disease (Mayr and Bartel, 2009), APA has the potential to be a major biological regulator. This appears to be especially true in the nervous system, where global lengthening of 3' UTRs has been observed from *Drosophila* to humans (Hilgers et al., 2011b; Miura et al., 2013; Smibert et al., 2012; Zhang et al., 2005) as well as in male gonads, where instead concerted shortening of 3' UTRs appears to be conserved (Li et al., 2016; McMahon et al., 2006; Smibert et al., 2012).

To gain a comprehensive view of the complexity of 3' UTR isoforms expression in *Drosophila* we set to build a genome-wide atlas of alternative cleavage and polyadenylation. We report here the most comprehensive annotation of polyadenylated ends of RNA in *Drosophila* to date. We find evidence of APA for two thirds of all genes and report tissue specific expression of 3' UTR isoforms for hundreds of genes. Comparison of tissue specific patterns of 3' UTR expression between species shows conservation of expression of long 3' UTR isoforms in head and suggests a faster rate of evolution for testis specific events. Finally, we present a deep analysis of pAs conservation between species and provide evidence that cis-element changes play an important part in the evolution of patterns of 3' UTR isoform expression.

Results

Genome-wide annotation of polyadenylated 3' ends of RNA in Drosophila

To obtain a comprehensive atlas of 3' UTR isoforms expressed in *Drosophila* we sequenced 3' ends of transcripts using a 3'-seq protocol we developed (Chapter 5). Briefly, polyadenylated ends were enriched from fragmented total RNA using an oligo(dT) strategy to bind at the junction between the poly(A) tail and the end of the RNA transcript. These RNA fragments were cloned through a number of steps into a library for high throughput sequencing (Fig 2.1). To obtain a broad sampling of expressed genes, we chose a set of samples that would enable us to maximize the number of expressed genes through the sequencing of a multitude of tissues, developmental time-points and cell lines, similarly to what previously done for the more extensive modENCODE (Brown et al., 2014; modENCODE Consortium et al., 2010). However, the modENCODE effort did not focus on annotating 3' end transcript isoforms, leaving this aspect of the transcriptome poorly annotated (Brown et al., 2014; Matthews et al., 2015). To build our atlas we made libraries of a set of tissues (head, ovary, testis, carcass and whole fly), 10 different cell lines and 6 time-points spanning embryonic development. To take full advantage of data available through modENCODE we sequenced the same total RNA samples used to generate the modENCODE cell line data, allowing for direct comparison with the existing RNA-seq datasets (Cherbas et al., 2011). Altogether, these samples allowed us to capture more than 97% of all *Drosophila* genes, with 83% of all genes expressed at a minimum of 1 RPM in at least one library, providing us with near comprehensive evidence of the 3' end formation landscape of the *Drosophila* transcriptome.

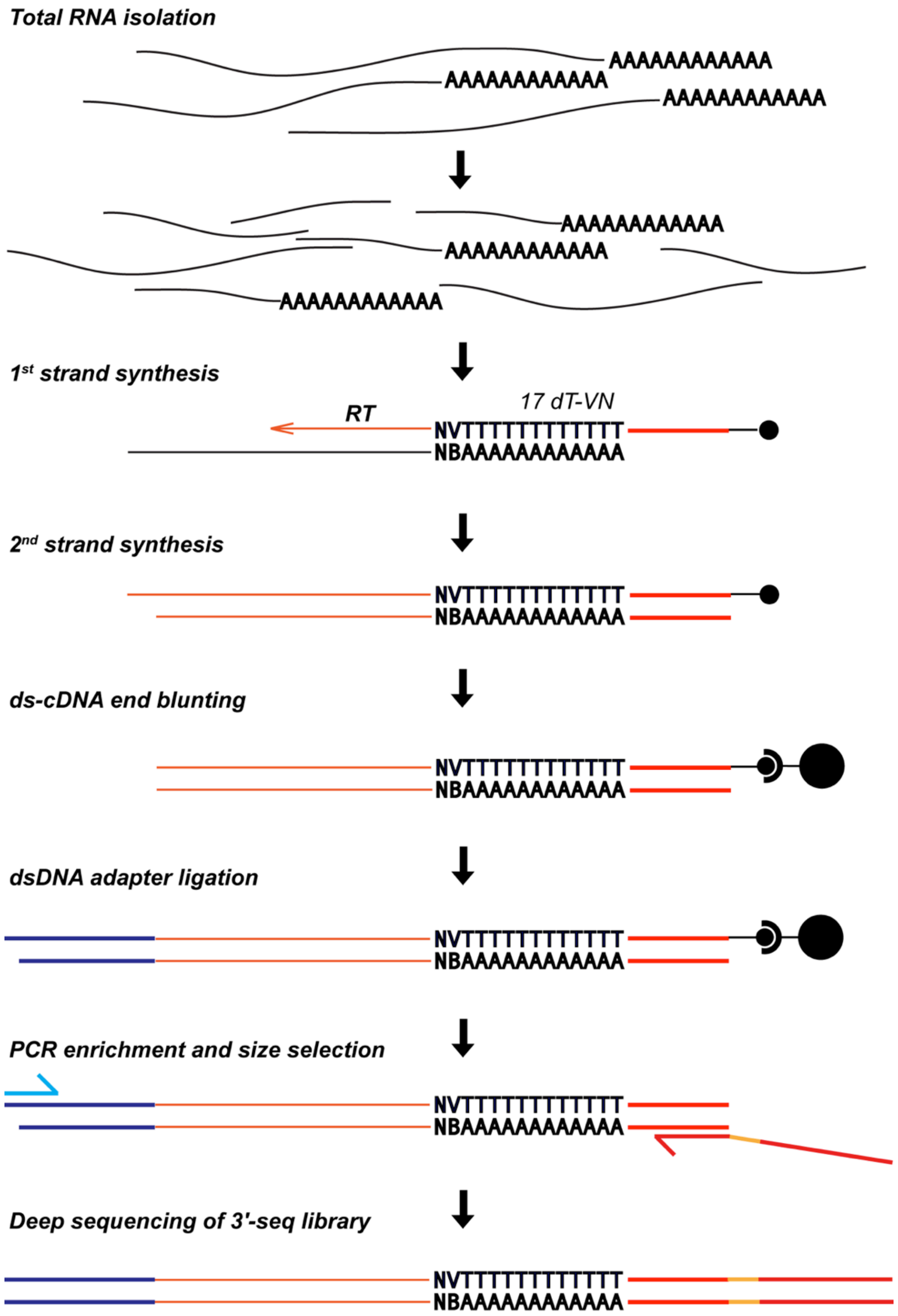


Figure 2.1 Schematic representation of 3'-seq.
The protocol represented in this schematic is outlined in detail in Chapter 5.

Following mapping, filtering and clustering of the data, as described in detail in Chapter 5, we obtain around 200 million mapping reads (Table 2.1), and identify ~62,000 sites of cleavage and polyadenylation with evidence of at least 3 independent reads. To assess the quantitative nature of 3'-seq, we compared gene expression measurements obtained using either 3'-seq or RNA-seq of matched samples, and observed a strong linear correlation (r = 0.91-0.94) between the two methods, indicating that 3'-seq is an effective method to quantify gene expression of 3' end isoforms (Fig. 2.2A-C). This dataset constitutes the most comprehensive atlas of annotations of 3' ends of polyadenylated RNAs in *D. melanogaster* to date.

D. melanogaster

3'-seq library	Mapped reads
Head Female	19038383
Ovary	19976914
Testis	11308288
Carcass Female	9146088
Carcass Male	22582073
Whole Body Female	1894540
Whole Body Male	14024775
Embryo 0-45min	7058231
Embryo 45-90min	59109
Embryo 90min-6h	3517300
Embryo 6-12h	5233924
Embryo 12-18h	3464045
Embryo 18-24h	7720275
BG3	922868
D16-c3	3194323
D4	14558434
D32	3227072
D20	5659091
mbn2	3238239
L1	15351857
GM2	6754734
OSS	6505160
S2R+	9924908

Table 2.1 Mapping statistics for 3'-seq samples from *D. melanogaster*.
Total 3'-seq reads mapping onto the *D. melanogaster* genome (dm6) after mapping the raw FASTQ files as described in Chapter 5.

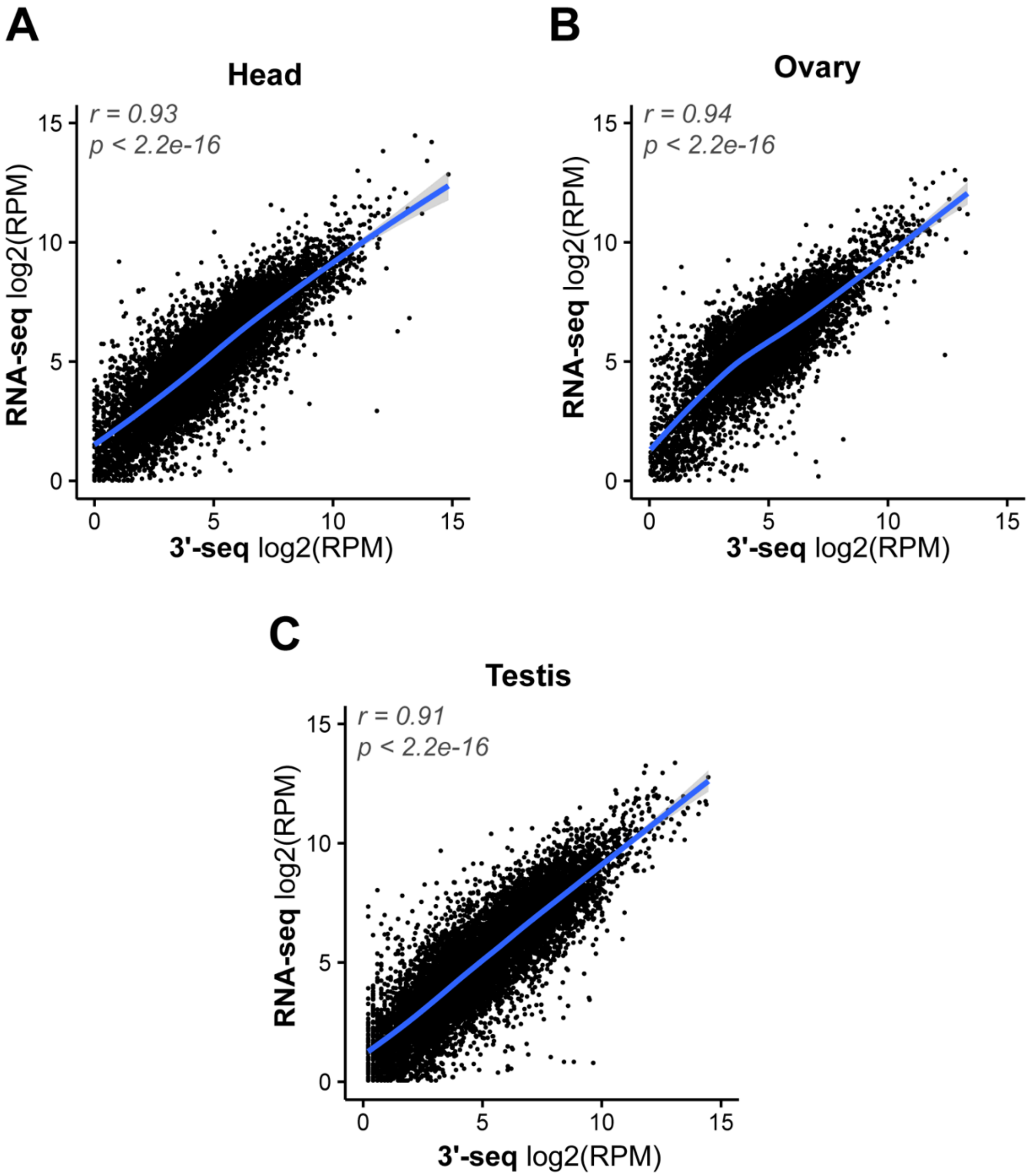


Figure 2.2 3'-seq is a quantitative measure of gene expression.
Correlation of gene expression measurements derived either using RNA-seq or 3'-seq of matched samples. (A) Female heads. (B) Ovary. (C) Testis. Pearson correlation coefficient and p values are shown on the top of the graph. Blue line represent loess fit curve. (Gene expression quantification performed by Jiayu Wen).

To verify that our 3'-seq method can capture 3' ends of RNA, we looked at the distribution of the raw reads as well as the processed 3' end calls onto gene models extracted from the most recent *D. melanogaster* annotation (r6.12). Confirming that we can detect 3' ends, we observe that more than 78% of mapped reads and 67% of clusters fall onto 3' UTRs or 5 kb downstream of existing gene models (Fig 2.3A,B). To further validate our method, we looked at the sequence composition around the pAs that were identified in our atlas to verify the presence of an A-rich PAS upstream and a U-rich region downstream of the cleavage site, a characteristic signature of pAs (Tian and Graber, 2012). As expected, we observe enrichment of As upstream and U-rich sequence downstream to the cleavage site (Fig 2.3C). The observed nucleotide composition around the putative cleavage sites is highly indicative of the presence of functional PAS. To confirm the presence of PAS upstream of the 3' ends annotated in our atlas, we looked for the presence of the most common hexamers upstream to the putative cleavage sites. We observe high enrichment of known PAS variants (Fig 2.3D), similar to those identified in previous studies that characterized the predominant PAS in *Drosophila* through EST data analysis. The same top 4 PAS variants, AAUAAA>AUUAAA>AAUAUA>UAUAAA were identified (Graber et al., 1999; Retelska et al., 2006). These results support that our data represents putative 3' ends of polyadenylated RNA.

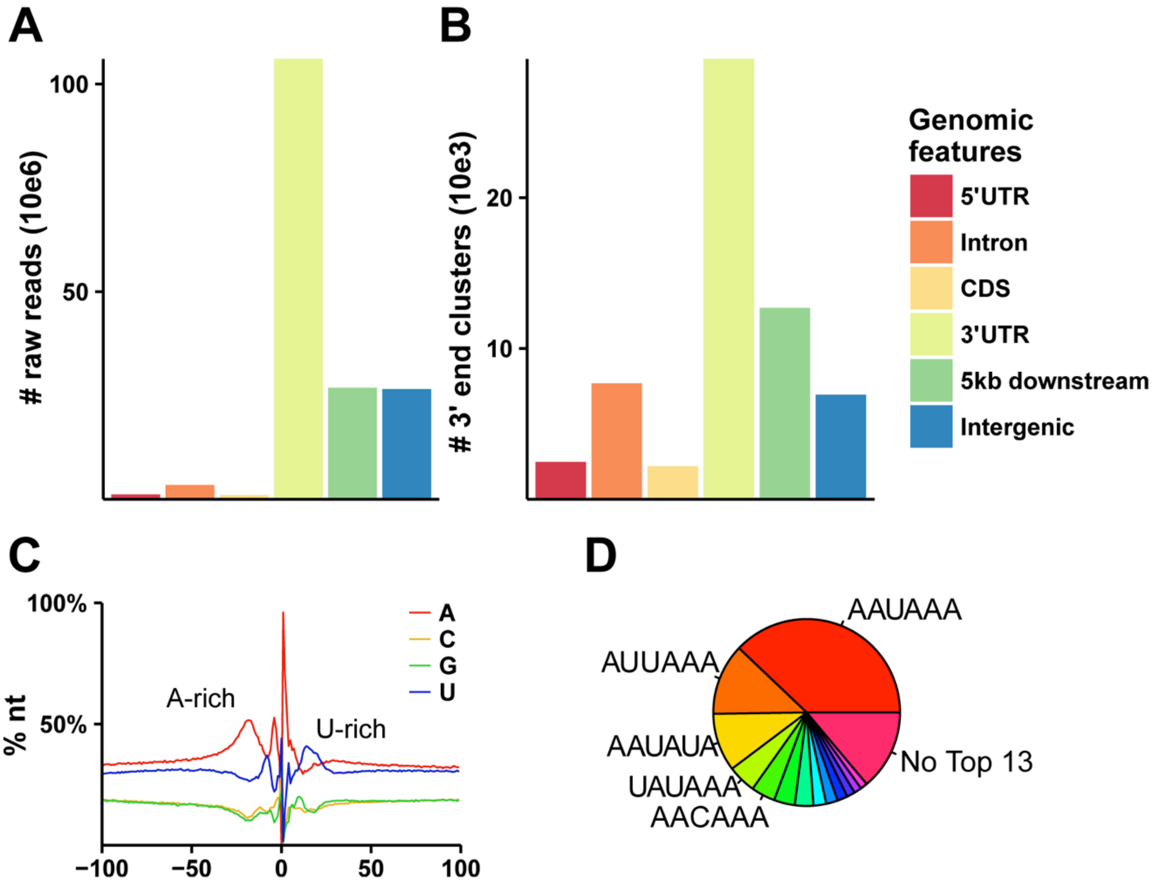


Figure 2.3 3'-seq captures 3' ends of poly(A) RNA.
(A) Distribution of mapped 3'-seq reads onto existing genomic features in the *D. melanogaster* annotation (r6.12). (B) Distribution of 3'-seq based calls of 3' ends in our atlas onto existing genomic features in the *D. melanogaster* annotation (r6.12). (C) 3'-seq derived 3' ends are aligned around the predicted cleavage site and the nucleotide distribution around putative cleavage site is plotted (centered at 0). % of each of 4 nucleotides in a window of 200 bp centered at the cleavage site are shown. (D) De-novo derivation of most represented hexamers in a 50 nt window upstream of 3'-seq derived 3' end positions.

Despite extensive annotation of the *D. melanogaster* transcriptome through efforts such as the modENCODE project (Brown et al., 2014; modENCODE Consortium et al., 2010), we observe a substantial number of pAs occurring in a 5 kb window downstream to existing annotation (Fig. 2.3B), suggesting that the most distal 3' end of multiple loci might still remain under-annotated. For example the gene *chb* and *CG6178* show extension of the model by ~850 nt, as evidenced by both 3'-seq and RNA-seq signals in samples from heads (Fig. 2.4A). In order to assign these putative 3' ends to upstream transcriptional units, we leveraged RNA-seq as additional evidence to aid in attributing the ends. Briefly, we de-novo predicted transcriptional units by using isoSCM (Shenker et al., 2015), a tool we previously developed to predict 3' end isoforms through analysis of RNA-seq datasets. IsoSCM builds de-novo models based on certain coverage parameters. In this case we required that the sequence between the end of the annotation and the 3'-seq cluster have continuous RNA-seq evidence with gaps of less than 100 bp. To our surprise, we identify ~7000 ends that can be attributed within 5 kb of existing gene models. It is important to note that the majority of the events that we identify represent predicted minor isoforms, for example the novel long 3' UTR extension of the gene *msi* is representative of less than 1% of total *msi* transcript in head (Fig. 2.4B). Despite this, we chose to annotate these isoforms as they could be representative of a subset of transcripts that are specific to small cell populations, and would hence be under-reported in broad efforts, such as modENCODE.

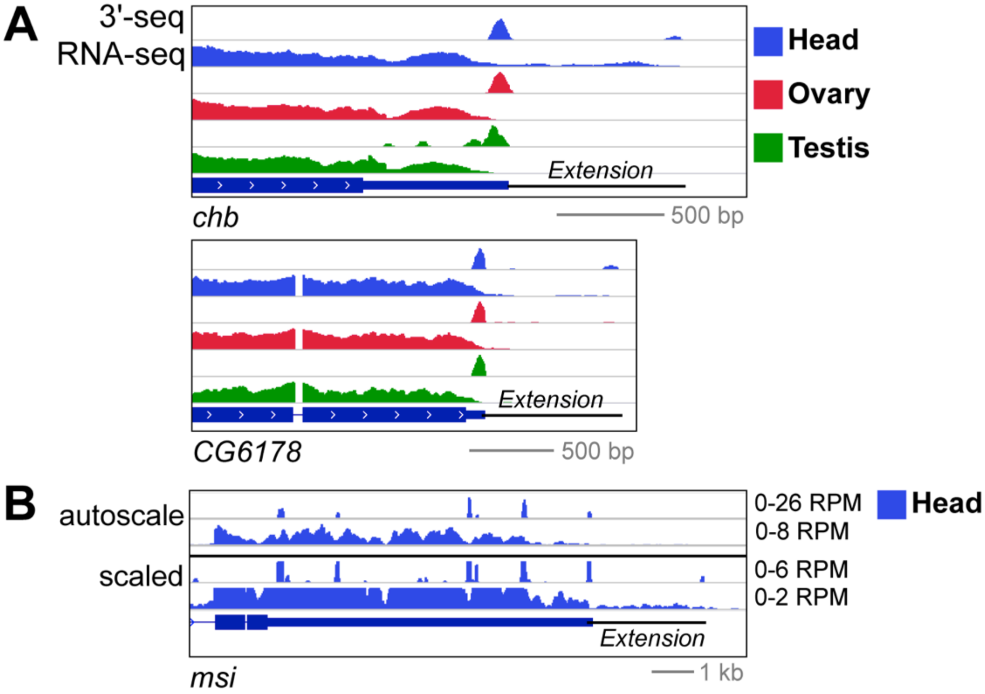


Figure 2.4 Examples of 3' end extension of gene models in the current r6.12 annotation.

RNA-seq and 3'-seq tracks are shown using auto-scaling parameters (A) and (B) top panel, while in the lower panel the scale was set manually to be able to visualize the low level extension of *msi*. 3'-seq evidence in a 5kb window downstream of existing annotation (r6.12), shown in blue, as well as continuous RNA-seq were used to extend gene models as shown. Samples are shown in color as shown in the legend. Scale for the genomic range is shown in grey. Images were taken using the IGV genome browser.

To annotate 3' ends, we assigned our 3' end annotation to existing extended gene models. Altogether, we expand the number of genes that undergo APA from the current 25% in FlyBase r6.12 to about 65% (Fig. 2.5A), similarly to what observed in other eukaryotes (Shi, 2012). We also find that 25% of all genes express more than 4 3' end isoforms (Fig. 2.5B), as shown for one of the most extreme cases, *Hrb27C* which has biochemical evidence for 47 different ends(Fig. 2.5C), underscoring the complexity of 3' end isoform expression found in *Drosophila*. Our previous analysis of 3' UTR diversity in *Drosophila* confirms the complexity of 3' end expression of *Hrb27C* which has at least 7 different 3' end isoforms as detected by conventional Northern blotting (Smibert et al., 2012). We note that not all genes that express long 3' UTRs (> 4kb), express large numbers of discreet isoforms, as in the case of *Hrb27C*. For example, we annotate only three 3' UTR isoforms for *beat-VII*, despite a long 4 kb 3' UTR (Fig. 2.5D). The low level events, such as the majority of annotated ends in *Hrb27C*, retain features of bona fide PAS. For example, 65% of ends that represent less than 5% of total gene expression present one of the top 3 PAS upstream of the cleavage site, suggesting that these are biochemically valid events generated through recognition of strong PAS. However, to provide a more practical measure of APA in *Drosophila*, we also provide annotation of 3' ends that comprise at least 5% of total expression of the locus, to focus the annotation on isoforms that are more highly expressed and could more likely be empirically validated. Altogether, our study vastly expands the complexity of terminal 3' UTR isoform expression in *Drosophila*.

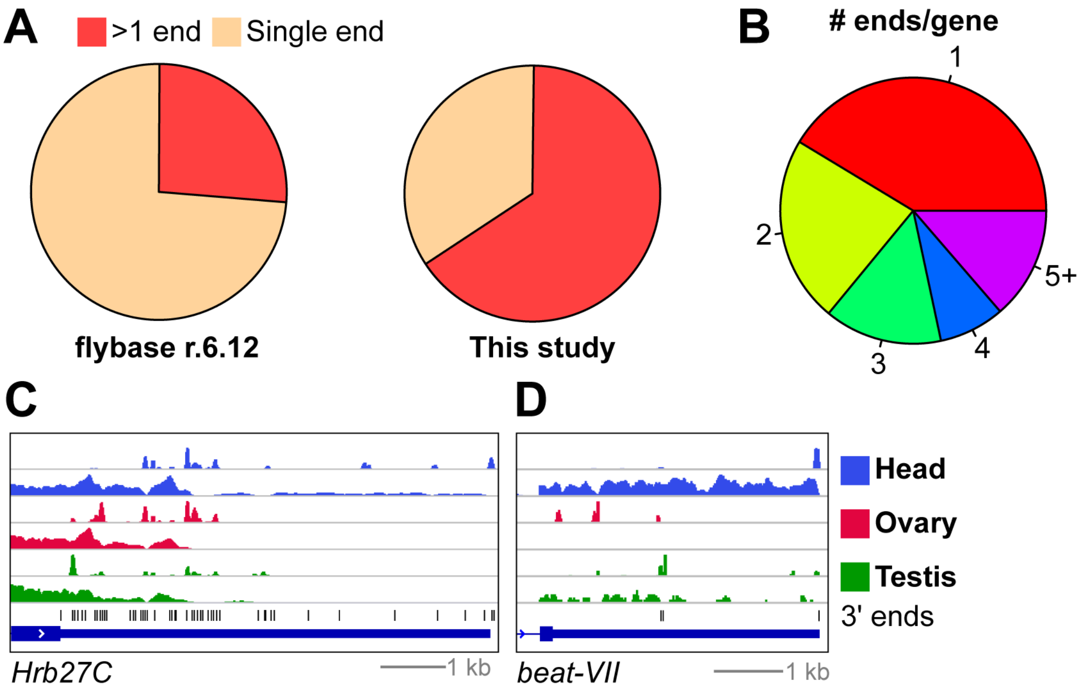


Figure 2.5 The majority of genes in *D. melanogaster* express multiple 3' UTR isoforms.

(A) Percentage of genes with either one annotated end or more than one for the current *D. melanogaster* annotation (r6.12) or our revised 3'-seq based atlas. (B) Percentage of genes with different numbers of annotated 3' end isoforms in our 3'-seq based atlas. (C) Example of *Hrb27C*, a gene with an extreme 47 ends, as annotated in our atlas. (D) Example of a gene with a long 4 kb 3' UTR that does not exhibit a large amount of differ 3' UTR isoforms. Scale for the genomic range is shown in grey. Images were taken using the IGV genome browser.

Tissue specific dynamics of 3' UTR isoform expression in Drosophila

Given that samples from head, ovary and testis provide a broad sampling of transcript (Brown et al., 2014) as well as 3' end isoform diversity (Smibert et al., 2012) we focus the remaining analysis on these three tissues. We have previously shown that a subset of ~400 *Drosophila* genes express isoforms with long 3' UTRs in the nervous system while ~100 genes express specific short 3' UTR variants in testis (Smibert et al., 2012). However, these dynamics of 3' UTR isoform expression were assayed by analyzing stranded RNA-seq datasets, which can only act as a readout of 3' end isoform expression changes that occur at large distance between recognized PAS as well as represent high degree change in isoform expression between the compared samples (Shenker et al., 2015). In order to assess if we could expand the scope of tissue specific 3' end isoform expression, we calculated the weighted length of genes expressed in pairs of tissue using 3'-seq isoform quantification. Weighted 3' UTR length is obtained taking the average of all 3' UTR isoforms length per gene weighted by the contribution of each isoform expression. When comparing the weighed length of 3' UTRs in testis or ovary to the 3' UTRs of the same genes expressed in head, we find expression of longer isoforms for hundreds of genes, substantially expanding the number of genes that express longer 3' UTR isoforms in this tissue (Fig 2.6A,B). Similarly, we find that the majority of genes that undergo a change in the expression of the different length 3' UTR isoforms in testis, express more shorter isoforms in this tissue, when compared to either head or ovary (fig. 2.6A,C). An example of these dynamics is shown for *fs(1)K10, Gαo* and *H*, which express more short 3' UTR isoforms in testis, longer 3' UTR isoforms in ovary and the longest ones in head (Fig. 2.6D). These results vastly expand the number of genes that change 3' UTR isoform expression between these three tissues, making the regulation of 3' UTR isoform expression a highly regulated event in *Drosophila*.

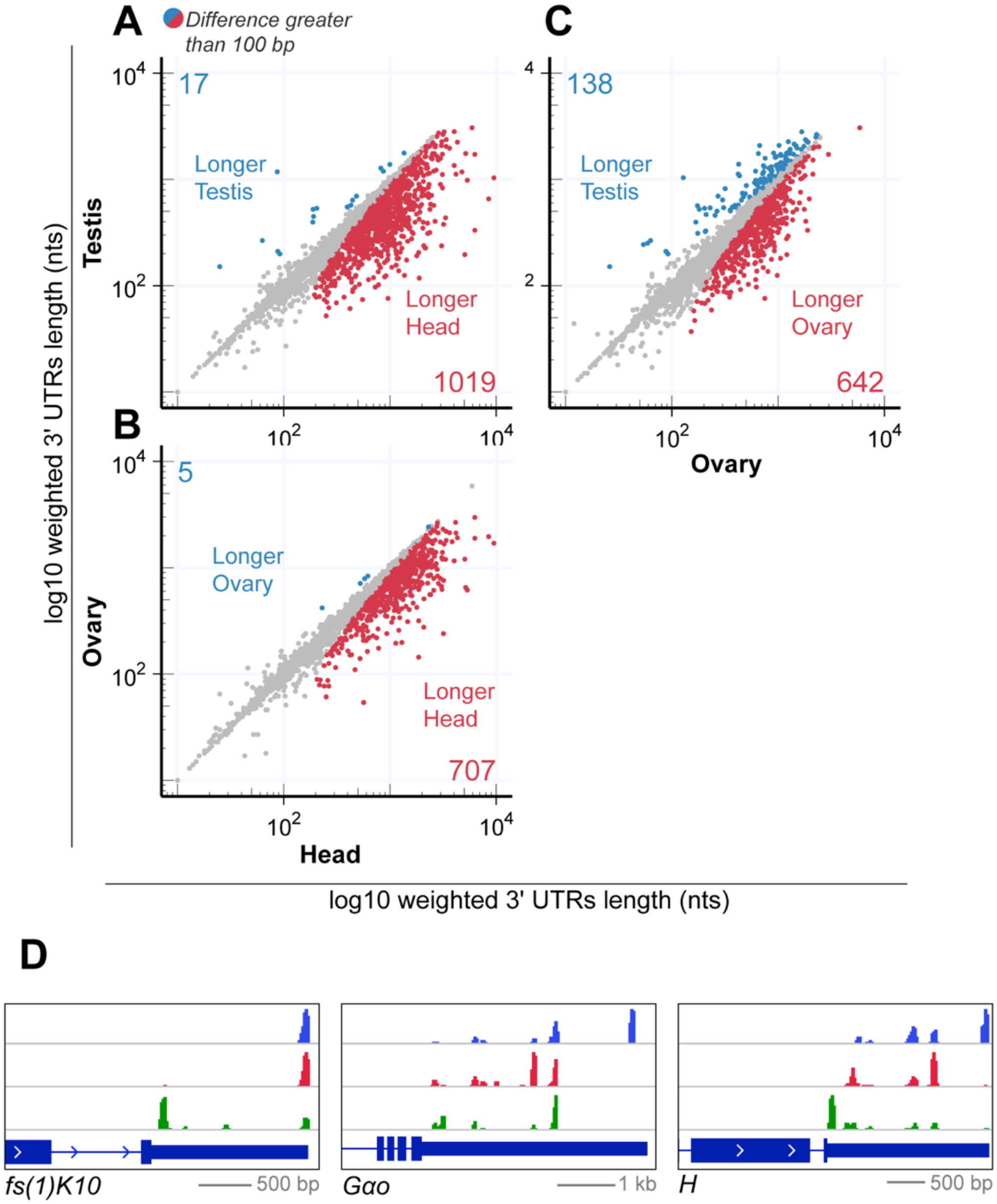


Figure 2.6 Hundreds of genes have the propensity to express long 3' UTRs in head and short 3' UTRs in testis.

(A-C) Weighted 3' UTR length comparison between tissues. Weighted 3' UTR length is obtained taking the average of all 3' UTR isoforms length per gene weighted by the contribution of each isoform expression. Genes are expressed at a minimum of 5 RPM in all samples. The genes which weighted 3' UTR length differs by 100 bp or more between samples are shown in color, red - longer weighted length in the sample on the x axis, blue - longer weighted length in the sample on the y axis. (A) Head vs. testis. (B) Head vs. ovary. (C) Ovary vs. testis. (D) Examples of 3'-seq evidence for genes that express different length 3' UTR isoforms in head, ovary and testis. Scale for the genomic range is shown in grey. Images were taken using the IGV genome browser.

In our previous study, we had identified ovary as a tissue with intermediate length 3' UTR isoforms, a similar 3' UTR length profile to most somatic tissues available at the time (Smibert et al., 2012). However, during manual browsing of RNA-seq datasets, we noticed that the ovary appeared to express discreet short length 3' UTR isoforms, often shorter than those observed in the somatic carcass sample, as well as in cell lines such as S2 and L1 (Fig 2.7A). Furthermore, we noticed the same short 3' UTRs in very early 0-2h embryos, which appeared to lengthen very rapidly following 2h of embryogenesis. This was also true for some of the genes that were characterized to express neural specific 3' UTR long isoforms, such as *AGO1* (Hilgers et al., 2011b; Smibert et al., 2012). However, *AGO1* appears to express more short 3' UTR isoforms only in ovary and 0-2h embryos, while all other samples express more longer 3' UTR isoforms (Fig. 2.7B), suggesting expression of longer 3' UTR isoforms of *AGO1* is not neural specific. Instead, these observations suggest that a subset of genes express 3' UTR isoforms specific to the ovary and early embryogenesis. Since the ovary is mostly comprised of developing eggs, the bulk of which would come from late stage developing eggs, we hypothesized that gene expression profile in ovary might be driven by gene expression of early embryogenesis. Indeed, when we compare the 3'-seq between ovary and 0-2h embryos, we obtain very high correlation (r_s = 0.86, Fig.2.7C), confirming that ovary might be a readout of 0h embryos, which are loaded with maternal transcript and are transcriptionally silent (Tadros and Lipshitz, 2009).

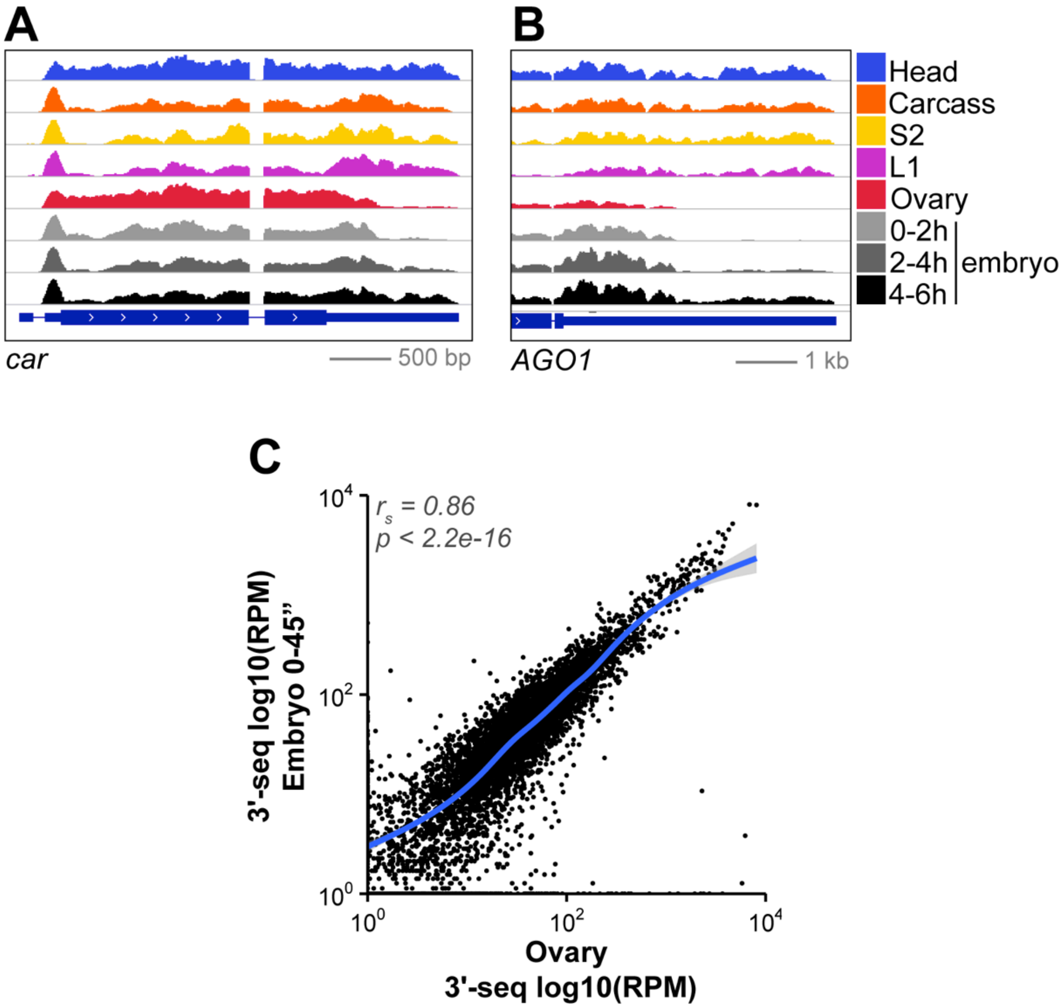
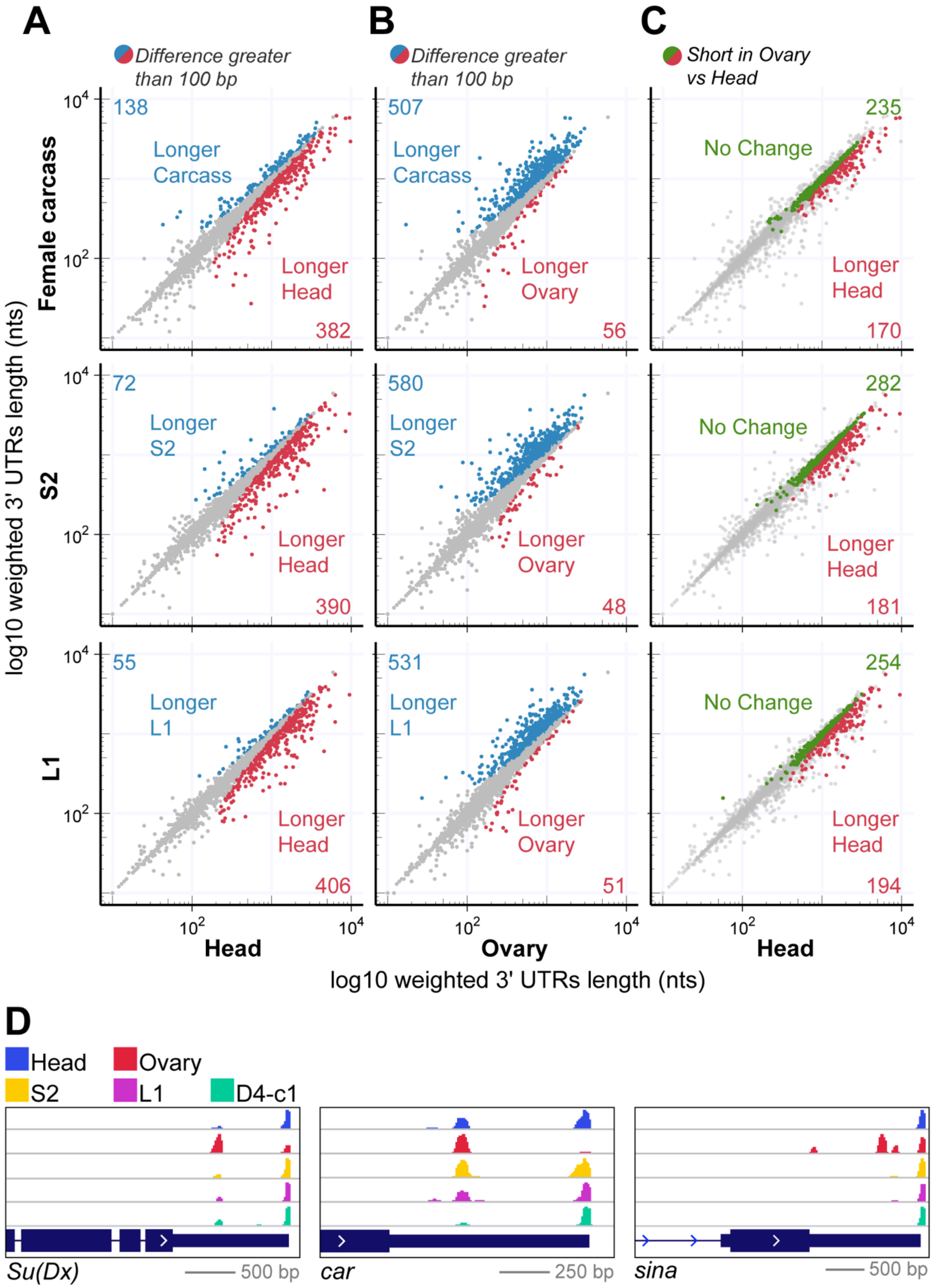


Figure 2.7 A subset of genes express a pattern of 3' UTR expression specific to the ovary.

(A-B) Example of genes that express a pattern of 3' UTR isoform expression specific to the ovary and early 0-2h embryo. Tracks represent total RNA-seq from the modENCODE project. Samples are shown in different color and their identity is according to the legend on the left of the top panel. Scale for the genomic range is shown in grey. Images were taken using the IGV genome browser. (C) Comparison of normalized 3'-seq between ovary and very early 0-45 min embryos. Test of correlation with Spearman rank order rho and p value shown on top of the graph. Blue line represent loess fit curve.

To identify ovary/early embryo specific events, we looked to see if we could identify genes that express shorter 3' UTRs in ovary compared to the somatic carcass sample, which is devoid of ovary and head, as well as cell line samples. As expected from our previous analysis, hundreds of genes express longer 3' UTR isoforms in head compared to either carcass or cell lines (Fig. 2.8A), confirming head specific expression of longer 3' UTR isoforms. However, when we compare weighted 3' UTR length of these samples to ovary, we identify hundreds of targets expressing longer 3' UTRs compared to ovary (Fig. 2.8B). These results confirm a trend to express shorter 3' UTR isoforms in the ovary compared to somatic cells, as hinted by RNA-seq browsing. Next we re-evaluated the set of genes that showed longer 3' UTR isoform expression in head compared to ovary. Given that some ovary 3' UTR isoforms seem to be specific to the ovary, such as *AGO1* (Fig. 2,7B), a subset of genes that showed expression of long 3' UTR isoforms in head could instead be genes that express a more ubiquitous set of 3' UTR isoforms and only express shorter ones in ovary. We find that of the genes that are shortened in ovary compared to head (Fig. 2.6B), only half express shorter 3' UTR isoforms in carcass or the cell lines compared to head (Fig. 2.8C). Examples of this are shown for *Su(dx)*, *car* and *sina* which express more short isoforms specifically in ovary (Fig 2.8D). These results confirm the expression of a specific pattern of shorter 3' UTR isoforms in the ovary and early embryo, suggesting that female gonads have a specific program of 3' UTR isoforms expression as seen for head and testis.



A
B
C
Difference greater than 100 bp
Difference greater than 100 bp
Short in Ovary vs Head
Female carcass
S2
L1
log10 weighted 3' UTRs length (nts)
10^4
10^2
138
Longer Carcass
Longer Head
382
507
Longer Carcass
Longer Ovary
56
235
No Change
Longer Head
170
72
Longer S2
Longer Head
390
580
Longer S2
Longer Ovary
48
282
No Change
Longer Head
181
55
Longer L1
Longer Head
406
531
Longer L1
Longer Ovary
51
254
No Change
Longer Head
194
Head
Ovary
Head
log10 weighted 3' UTRs length (nts)
D
Head
Ovary
S2
L1
D4-c1
Su(Dx)
500 bp
car
250 bp
sina
500 bp

Figure 2.8 3' UTR expression dynamics specific to the ovary.
(A-C) Weighted 3' UTR length comparison between samples. Weighted 3' UTR length is obtained taking the average of all 3' UTR isoforms length per gene weighted by the contribution of each isoform expression. Genes are expressed at a minimum of 5 RPM in all samples. (A,B) The genes which weighted 3' UTR length differs by 100 bp or more between samples are shown in red when longer weighted length is in the sample on the x axis, blue when longer weighted length is in the sample on the y axis. (C) Genes that were longer in head when compared to ovary are overlaid on top of the comparison of female carcass, S2 or L1 cells to head. In green are genes that were longer in head compared to ovary but express a similar pattern of 3' UTR isoforms when comparing the three samples to head. In red are genes that express longer isoforms in head when compared to ovary and all three samples as well. (A,C) Comparison of female carcass, S2 or L1 cells to head. (B) Comparison of female carcass, S2 or L1 cells to ovary. (D) Examples of 3'-seq of genes that express shorter 3' UTR isoforms specifically in ovary, compared to head or cell line samples. Samples are identified as shown in the legend. Scale for the genomic range is shown in grey. Images were taken using the IGV genome browser.

Multiple studies have shown that the efficiency of cleavage and polyadenylation is strongly affected by the strength of the underlying PAS (Prescott and Falck-Pedersen, 1994; Sheets et al., 1990). Proximal bypassed pAs found in 3' UTRs, have generally been shown to have weaker PAS, as evidenced by a lower percentage of such sites bearing a canonical AAUAAA PAS (Jan et al., 2011; Tian et al., 2005). To confirm predicted trends of *Drosophila* PAS strength from previous studies of small subsets of EST derived ends (Retelska et al., 2006), we looked for a difference in PAS identity upstream of our atlas of pAs found in terminal 3' UTRs. When we only consider pAs that define the cleavage site of single end genes, we see strong over-representation of the canonical PAS, and other strong variants AUUAAA and AAUAUA (Fig. 2.9A). A similar enrichment of strong PAS is seen upstream of pAs that define the most 3' end of 3' UTRs of genes with multiple ends, although to a lesser extent. When we consider PAS identity upstream of proximal pAs, we observe a lower proportion of strong PAS, with increase in other weaker variants (Fig. 2.9A), confirming our previous finding of reduced levels of canonical AAUAAA upstream of proximal pAs (Smibert et al., 2012) and the general understanding that proximal pAs found in 3' UTRs are associated with weaker PAS (Elkon et al., 2013). To test if there is a correlation between PAS identity and 3' UTR isoform expression, we calculated a relative strength score at each pA, which represents the fraction of isoforms ending at a specific site compared to downstream sites. This was calculated from the fraction of reads at each pA out of all reads at that site or downstream (see chapter 7). When we look at the cumulative relative strength distribution of pAs according to

PAS, we observe that known stronger PAS variants associate more often with pAs with high relative strength (Fig. 2.9B), in an order of strength analogous to what observed through analysis of EST data. These results confirm that in *Drosophila*, as in other eukaryotes, PAS identity plays an important role in the generation of 3' end isoforms and PAS identity contributes to the final relative steady state levels observed for a certain isoform.

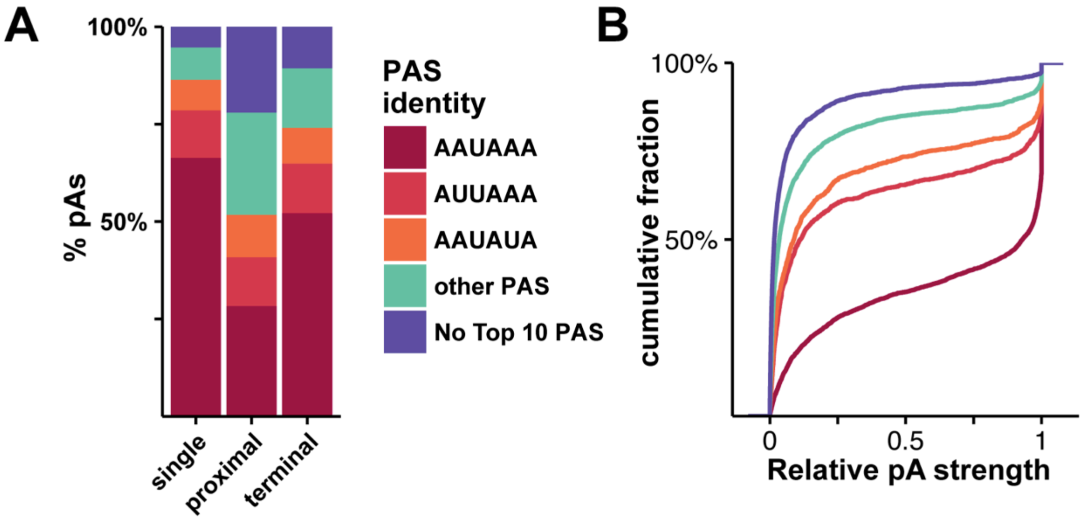


Figure 2.9 Analysis of PAS identity at terminal 3' UTR pAs
(A) Percentage of pAs that present in the 50 nt window upstream of the pA site a PAS as shown in the legend. pAs were divided based on 3' UTR class and position. Single - pAs that define the 3' end of genes with only evidence of single terminal 3' UTR isoform. Proximal - pAs found in the 3' UTR of genes that undergo APA, all pAs upstream of the most distal one. Terminal - most distal pAs that define the 3' most termini of terminal 3' UTRs that undergo APA. (B) Empirical cumulative distribution functions of the relative strength of all pAs in terminal 3' UTRs subdivided by the identity of the PAS (as shown in the legend on the left) in the 50 nt upstream of the cleavage site. The relative strength score represents the fraction of isoforms ending at a specific site compared to downstream sites. This was calculated from the fraction of reads at each pA out of all reads at that site or downstream ones.

Given that the strength of the PAS can influence the relative levels of 3' UTR isoform accumulation (Prescott and Falck-Pedersen, 1994; Sheets et al., 1990; Takagaki et al., 1996) and that differences in the identity of cis-elements surrounding regulated 3' ends have been noted to correlate with tissue specific accumulation of 3' UTR isoforms, such as during spermatogenesis (Liu et al., 2007), we asked if there was a correlation between tissue specific 3' UTR isoforms expression and putative pA strength, as this could suggest that PAS identity might play a role in the expression of tissue specific 3' UTR isoforms in *Drosophila*. We defined as tissue dominant pAs, the ones that have at least 30% higher relative strength in one tissue compared to the other two tissues, based on the strength score derived for each tissue. We identify a much greater number of testis dominant pAs (1053) compared to head (7) or ovary (108), suggesting that testis has different rules for PAS recognition. When we look at the PAS upstream of these putative tissue specific pAs, we notice that the dominant ends in testis have much less canonical AAUAAA PAS upstream of the cleavage site (Fig 2.10A). This observation is in line with what has been observed for testis pAs in human and mouse, where it has been found that a lower proportion of testis specific pAs presented the canonical AAUAAA PAS (Liu et al., 2007; MacDonald and Redondo, 2002). Despite a lower fraction of pAs preceded by a canonical AAUAAA, we nonetheless detect stronger recognition of sites that bare an AAUAAA hexamer (Fig. 2.10B), as shown by analysis of relative strength of pAs defined by different PAS; however the difference in recognition is not as pronounced compared to when analyzing all signals overall in 3' UTRs (Fig.

2.9A). Given the short nature of testis 3' UTRs (Fig. 2.6A,C) we hypothesized that testis might be able to strongly recognize the first canonical PAS available, and the decrease in canonical PAS recognition might be due to the presence of weaker upstream signals that can be recognized relatively strongly in testis (Fig. 2.10B). To test this, we took all the 1st recognized pAs that are dominant in testis with upstream hexamers other than AAUAAA and asked how often a canonical AAUAAA could be bypassed to recognize the weaker PAS. We were only able to identify 1% of 1st testis pAs which skip an AAUAAA to recognize a weaker PAS. To further investigate this, we looked for the presence of AAUAAA more than 50 bp upstream of the 1st recognized pA of all genes that are expressed at more than 1 RPM in testis. Only 10% of AAUAAA appear to not be 50 bp upstream of a pA site, however 46% of AUUAAA, a site of lesser strength, appears to be skipped before the first recognized PAS, confirming that AAUAAA is a stronger PAS in testis as well. Overall, these results confirm the PAS strength predictions derived through analysis of EST data (Liu et al., 2007; Retelska et al., 2006) and suggest that testis recognizes the first PAS available.

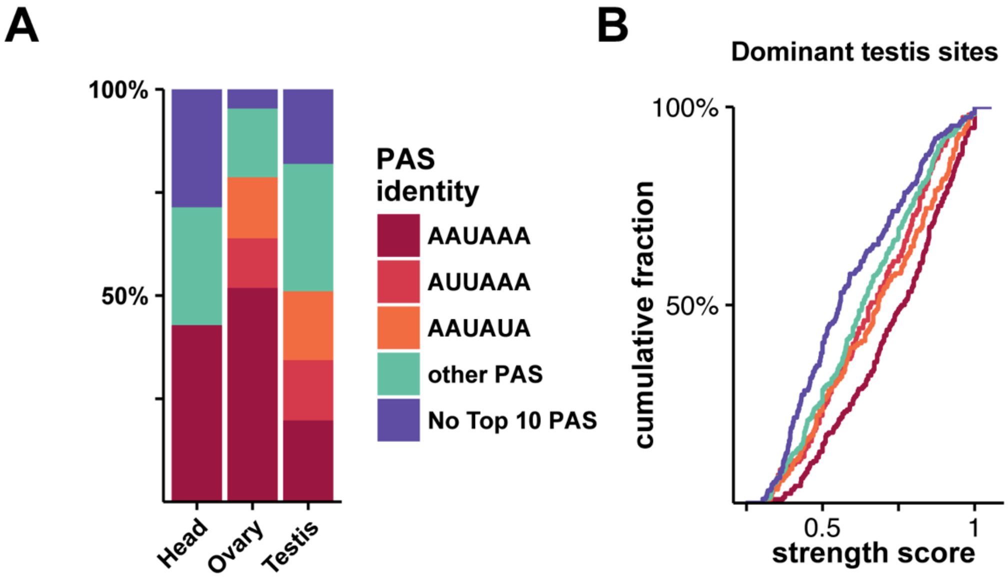


Figure 2.10 Analysis of PAS identity at tissue dominant pAs.
(A) Percentage of pAs that present in the 50 nt window upstream of the pA site a PAS as shown in the legend. pAs were divided based on tissue dominance in either head (n = 7), ovary (n = 108) or testis (n = 1053). (B) Empirical cumulative distribution functions of the relative strength of all pAs in terminal 3' UTRs of testis dominant pAs subdivided by the identity of the PAS (as shown in the legend on the left) in the 50 nt upstream of the cleavage site. The relative strength score represents the fraction of isoforms ending at a specific site compared to downstream sites. This was calculated from the fraction of reads at each pA out of all reads at that site or downstream ones.

Conservation of broad patterns of tissue specific 3' end isoform expression

To assess the breadth of conservation and novelty of tissue specific 3' UTR expression we decided to take advantage of the availability of good quality genomes of other Drosophilids (Clark et al., 2007). To this aim, we generated 3'-seq and RNA-seq libraries of head, ovary and testis for two additional *Drosophila* species (Table 2.2). We chose the closely related *D. yakuba*, and the more distantly related *D. virilis*, having respectively diverged from *D. melanogaster* ~5 and ~40 mya (Fig. 2.11A). Additionally, to supplement the 3' end annotations of these species, which are substantially less well annotated than *D. melanogaster*, we sequenced 3'-seq libraries of time-points throughout embryogenesis (Table 2.2). This effort allows us to significantly improve the 3' end annotation of these species, increasing the number of genes that undergo APA from one quarter to ~60% in *D. yakuba* (Fig. 2.11B) and ~75% in *D. virilis* (Fig. 2.11C).

D. yakuba

3'-seq library	Mapped reads
Head Female	14073832
Ovary	12505696
Testis	9529121
Embryo 0-12h	2653783
Embryo 12-24h	15519859

D. virilis

3'-seq library	Mapped reads
Head Female	16098260
Ovary	17914099
Testis	14877510
Embryo 0-12h	17723360
Embryo 12-24h	9275559
Embryo 24-36h	10064957
Embryo 36-48h	3687419

Table 2.2 Mapping statistics for 3'-seq samples from *D. yakuba* and *D. virilis*.
Total 3'-seq reads mapping onto the *D. yakuba* genome (droYak3) and the *D. virilis* genome (droVir3) after mapping the raw FASTQ files as described in Chapter 5.

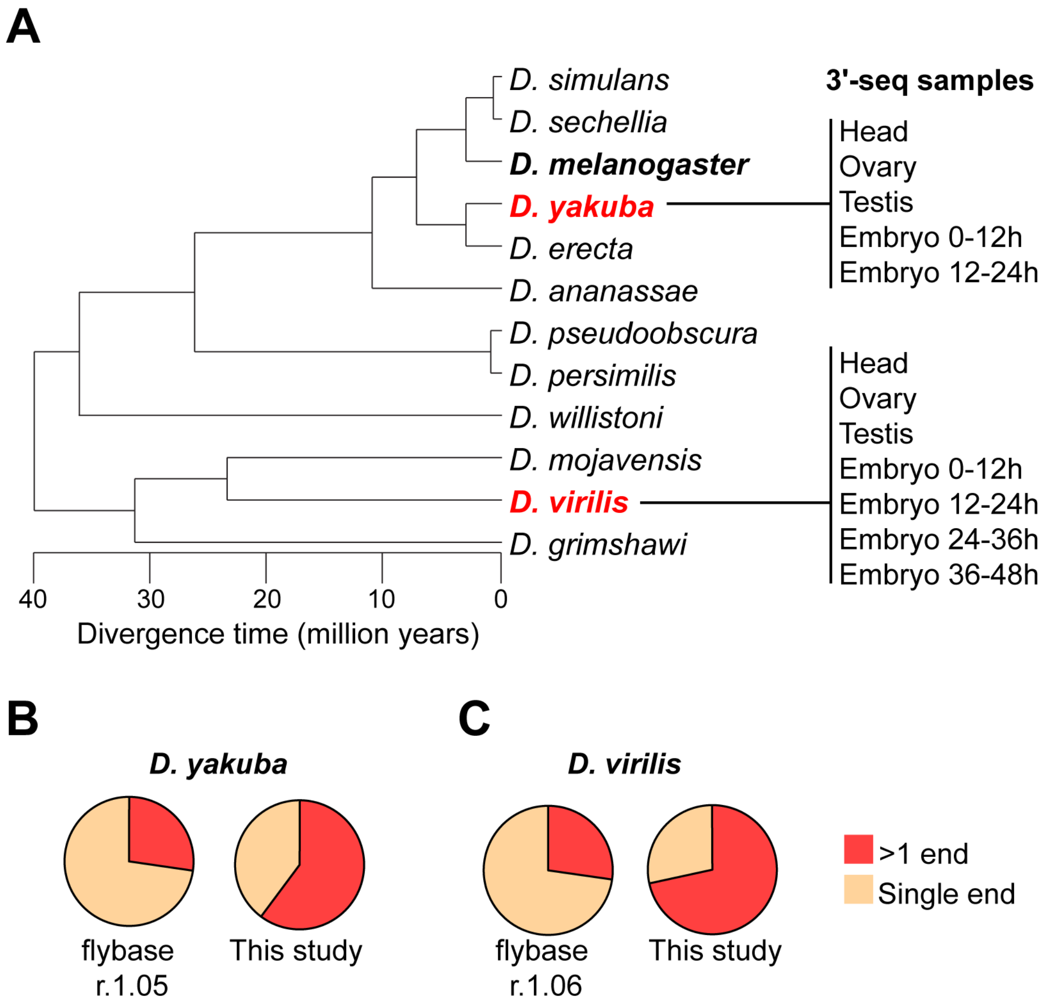


Figure 2.11 The majority of genes in *D. yakuba* and *D. virilis* undergo APA. (A) *Drosophila* phylogenetic tree showing the two species chosen for comparison with *D. melanogaster* in red. 3'-seq samples generated for each species are shown to the right. (B,C) Percentage of genes with either one annotated end or more than one for the current *D. yakuba* annotation (r1.05) or *D. virilis* annotation (r1.06) and our revised 3'-seq based atlas.

When manually browsing some of the genes that exhibit the most striking patterns of differential 3' UTR expression, we observed striking conservation of tissue specific 3' UTR isoform expression, as seen for *mei-P26*, the gene with the longest 3' UTR in *D. melanogaster* and with strong tissue specific patterns of 3' UTR expression (Fig. 2.12A). Of note, the orthologs of *mei-P26* in the other two species miss annotation of this incredibly long 3' UTR, reaching 22 kb in *D. virilis* (Fig 2.12A). However, we also noted a certain level of novel regulation of 3' UTR isoforms expression in a tissue specific manner, often in testis, as shown for *sick* which expresses the same pattern of 3' UTR isoforms in head, ovary and testis in *D. melanogaster* and *D. virilis* but which expresses as a dominant short 3' UTR isoform in testis (Fig.2.12B).

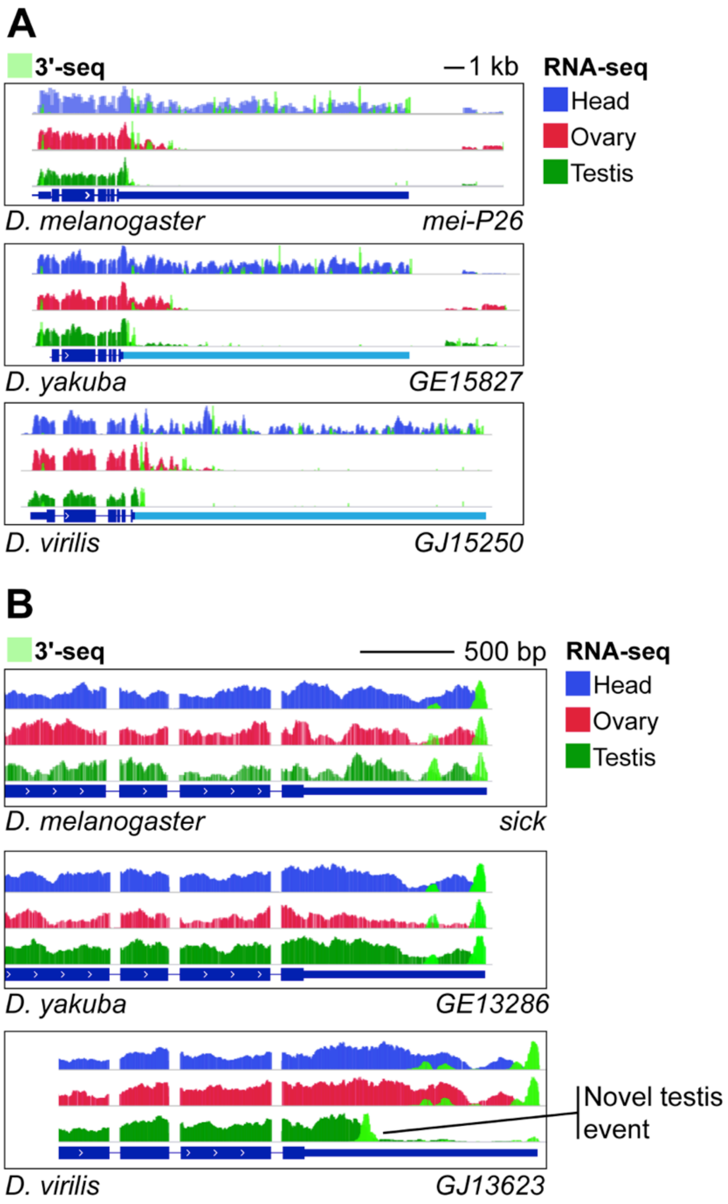


Figure 2.12 Conservation and divergence of tissue specific patterns of 3' UTR expression.

(A-B) Example of expression of genes with orthologs in *D. melanogaster*, *D. yakuba* and *D. virilis*. RNA-seq and 3'-seq of head, ovary and testis are shown as illustrated in the legend. 3'-seq is overlaid in light green onto the RNA-seq track. (A) Example of a gene that shows striking conservation of tissue specific patterns of 3' UTR expression. Extension of the annotation of *D. yakuba* and *D. virilis* is shown in light blue. (B) Example of a gene that shows de- novo expression of a dominant short 3' UTR isoform in the testis of *D. virilis*. All three orthologs shown are syntenic as suggested by the conservation of the same orthologous downstream gene (not shown). Scale for the genomic range is shown above and all species are shown at the same scale. Images were taken using the IGV genome browser.

To determine the genome-wide pattern of conservation of 3' UTR isoforms expression, we compared weighted length between tissues in the two species as computed for *D. melanogaster*. We observe a similar trend to express longer 3' UTR isoforms in head while expressing shorter 3' UTRs in testis in *D. yakuba* and *D. virilis* (Fig. 2.13A). Next, we wanted to test if the expression of tissue specific 3' UTR isoforms is conserved. When we computed the overlap of the orthologous genes that express more long 3' UTR isoforms in head compared to testis, we observe conservation of a majority of genes expressing longer 3' UTRs in Head (Fig. 2.13B), as seen for *mei-P26* (Fig. 2.12A). Similarly, a majority of genes that express more long isoforms in head compared to ovary appear to be conserved (Fig. 2.13C,D). A lower fraction of genes showed conserved expression of longer 3' UTR isoforms in ovary compared to testis (Fig. 2.14A,B) and we did not observe almost any conserved gene that showed expression of more longer 3' UTR isoforms in the testis compared to ovary (Fig. 2.14A,C). These observations point to expression of longer 3' UTR isoforms in the head compared to ovary and testis as a conserved aspect of gene expression in *Drosophila*. We also observe many species specific events, especially in the more distally diverged *D. virilis*, as shown for *sick* (Fig.2.12B). This appears to be especially true when comparing expression dynamics between testis and the two other tissues (Fig. 2.11C,D), suggesting that shortening of 3' UTRs in testis might be more subject to evolutionary change.

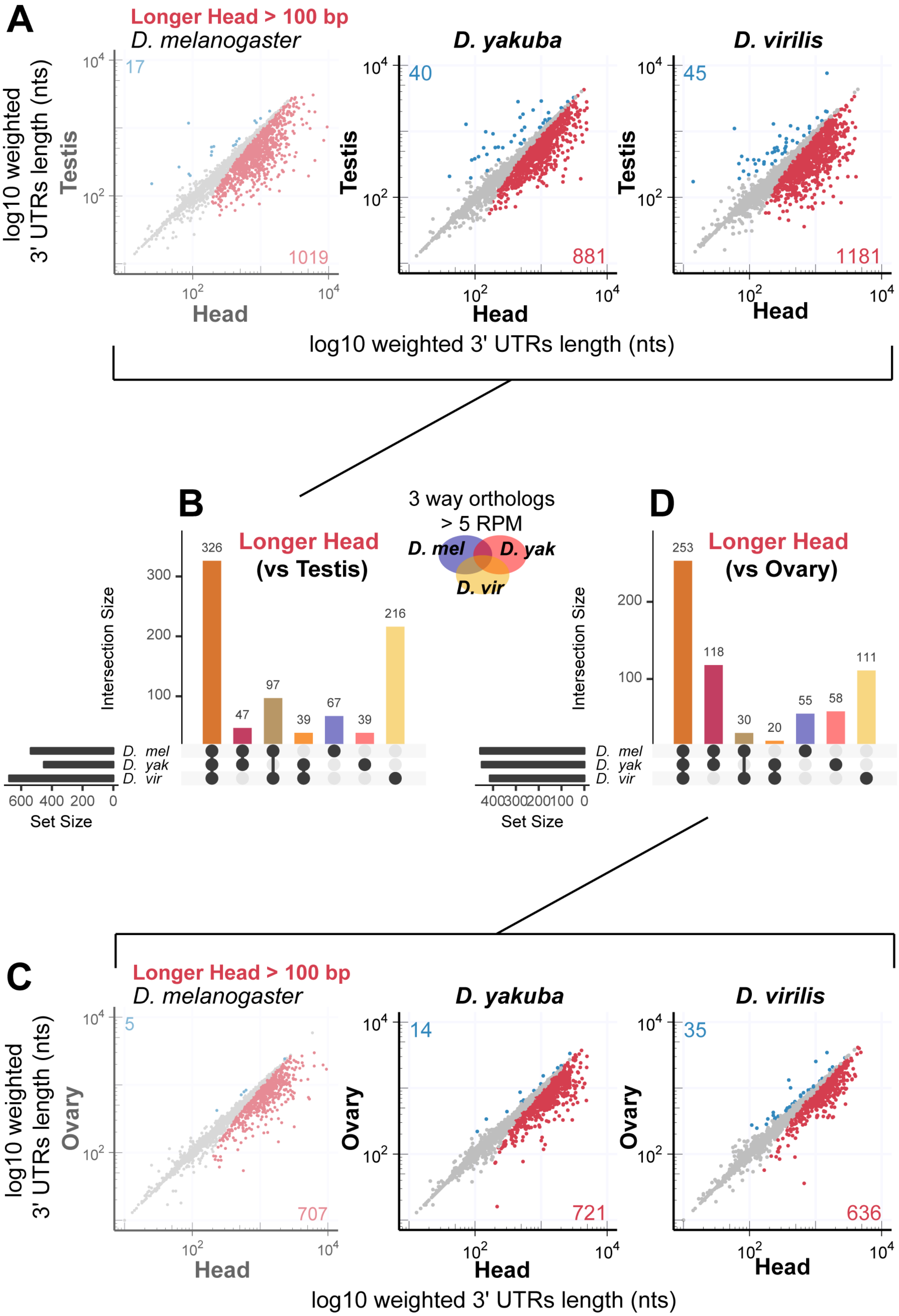

A
Longer Head > 100 bp
D. melanogaster
D. yakuba
D. virilis
log10 weighted
3' UTRs length (nts)
Testis
10^4
10^2
17
1019
40
881
45
1181
Head
log10 weighted 3' UTRs length (nts)
B
Longer Head
(vs Testis)
Intersection Size
300
200
100
326
47
97
39
67
39
216
D. mel
D. yak
D. vir
600 400 200 0
Set Size
3 way orthologs
> 5 RPM
D. mel
D. yak
D. vir
D
Longer Head
(vs Ovary)
Intersection Size
200
100
253
118
30
20
55
58
111
D. mel
D. yak
D. vir
400300200100 0
Set Size
C
Longer Head > 100 bp
D. melanogaster
D. yakuba
D. virilis
log10 weighted
3' UTRs length (nts)
Ovary
10^4
10^2
5
707
14
721
35
636
Head
log10 weighted 3' UTRs length (nts)

Figure 2.13 Expression of longer 3' UTR isoforms in head is a conserved aspect of gene expression.

(A,C) Weighted 3' UTR length comparison between tissues. Weighted 3' UTR length is obtained taking the average of all 3' UTR isoforms length per gene weighted by the contribution of each isoform expression. Genes are expressed at a minimum of 5 RPM in all samples. The genes which weighted 3' UTR length differs by 100 bp or more between samples are shown in color, red - longer weighted length in the sample on the x axis, blue - longer weighted length in the sample on the y axis. Comparisons are shown for *D. melanogaster*, *D. yakuba* and *D. virilis*. (A) Head vs. testis. (C) Head vs. ovary. (B,D) The overlap of the genes that express a weighted 3' UTR length of more than 100 bp in head compared to testis (B) or ovary (D) that have orthologs in all three species and are expressed at least at 5 RPM was taken and represented as a bar graph. Schematic Venn diagram and dots below the bar graph show the nature of each intersection. Horizontal bar graph shows the number of genes considered for each species.

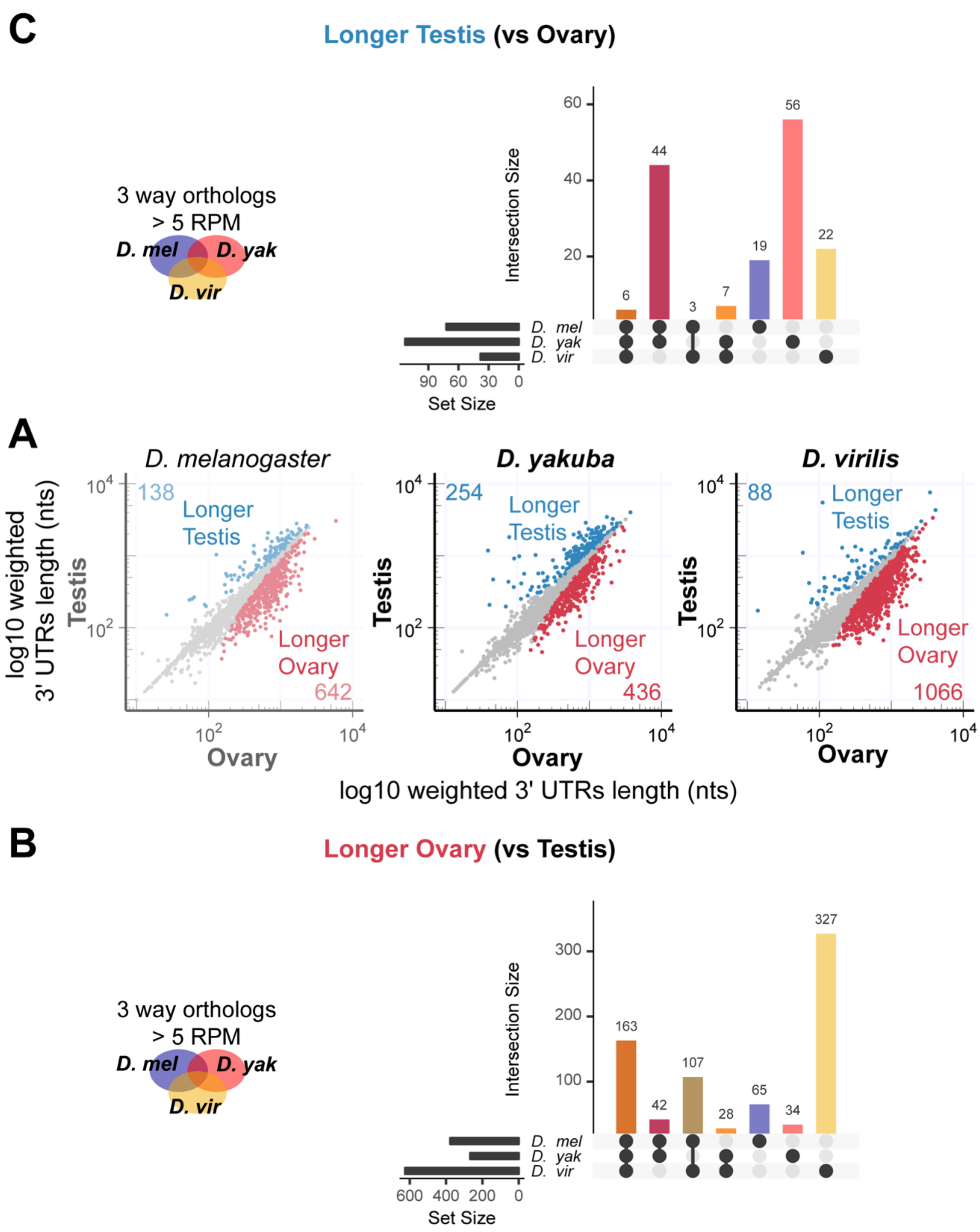

C
Longer Testis (vs Ovary)
3 way orthologs
> 5 RPM
D. mel
D. yak
D. vir
Intersection Size
60
40
20
6
44
3
7
19
56
22
D. mel
D. yak
D. vir
90
60
30
0
Set Size
A
D. melanogaster
D. yakuba
D. virilis
log10 weighted 3' UTRs length (nts)
Testis
10^4
10^2
138
Longer Testis
Longer Ovary
642
254
436
88
1066
Ovary
log10 weighted 3' UTRs length (nts)
B
Longer Ovary (vs Testis)
3 way orthologs
> 5 RPM
D. mel
D. yak
D. vir
Intersection Size
300
200
100
163
42
107
28
65
34
327
D. mel
D. yak
D. vir
600
400
200
0
Set Size

Figure 2.14 Testis specific expression dynamics of 3' UTR isoforms expression are not well conserved.

(A) Weighted 3' UTR length comparison between tissues. Weighted 3' UTR length is obtained taking the average of all 3' UTR isoforms length per gene weighted by the contribution of each isoform expression. Genes are expressed at a minimum of 5 RPM in all samples. The genes which weighted 3' UTR length differs by 100 bp or more between samples are shown in color, red - longer weighted length in the ovary compared to testis, blue - longer weighted length in the testis compared to ovary. Comparisons are shown for *D. melanogaster*, *D. yakuba* and *D. virilis*. (A) Ovary vs. testis. (B,C) The overlap of the genes that express a weighted 3' UTR length of more than 100 bp in ovary compared to testis (B) or longer in testis compared to ovary (C) that have orthologs in all three species and are expressed at least at 5 RPM was taken and represented as a bar graph. Schematic Venn diagram and dots below the bar graph show the nature of each intersection. Horizontal bar graph shows the number of genes considered for each species.

Conservation of pA usage

Given the observed patterns of conservation of 3' UTR isoforms expression (Fig. 2.13, 2.14), we dwelled deeper to investigate pA conservation at syntenic positions by leveraging pair-wise sequence alignments. We took advantage of the liftOver tool developed by the UCSC Genome Browser group (Kent et al., 2002) to identify syntenic positions of *D.* melanogaster 3' ends of terminal 3' UTR isoforms in either the *D. yakuba* or *D. virilis* genomes. We defined as conserved putative syntenic positions of *D. melanogaster* pAs that were within 25 bp of a 3'-seq annotated event in the other species (Fig. 2.15A, see Chapter 7 for details of the analysis). Of 38395 sites falling in terminal 3' UTRs, 17615 were conserved in *D. yakuba* (48%) while 8124 were conserved in *D. virilis* (22%). This is substantially a greater degree of pA site conservation then observed on other genomic features, where the pAs are conserved less than 10% in *D. yakuba* and less than 5% in *D. virilis* (Fig 2.15B). This is similar to what has been observed in a comparison of orthologous sites between mouse and human, where approximately 30% of pAs of single end genes was found to be conserved (Lee et al., 2008).

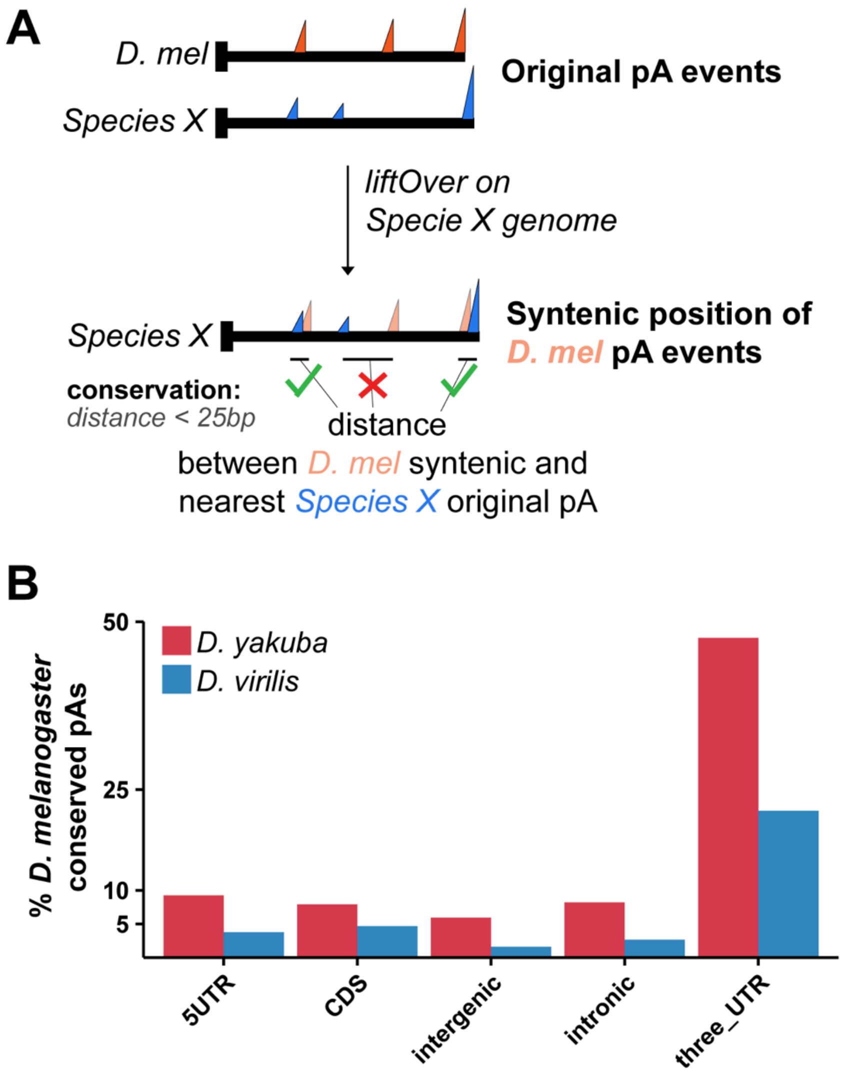


Figure 2.15 Conservation of pAs in *Drosophila.*

(A) Schematics of the pipeline for conservation analysis. *D. melanogaster* pAs were taken and syntenic locations on the genome of the species in the comparison were obtained using liftOver. *D. melanogaster* pAs were considered conserved if the syntenic position was within 25 bp of a 3'-seq annotated 3' end in the compared species. (B) Percentage of *D. melanogaster* sites annotated on different genomic features conserved in either *D. yakuba* or *D. virilis*. Conserved sites are syntenic positions of *D. melanogaster* pAs that are 25 bp or less away from the pA in the species under comparison.

Next, we asked what is the degree of PAS conservation at syntenic pAs. For the sites that have a conserved ortholog in the other species, defined as a pA 25 nt or closer to the syntenic position of the *D. melanogaster* pA, we observe a similar distribution of PAS identity at the syntenic positions in both the *D. yakuba* and *D. virilis* genomes (Fig. 2.16A,B, left panels), confirming conservation of 3' end formation. The presence of PAS upstream of syntenic positions is greatly reduced for the *D. melanogaster* pAs which are not conserved (Fig. 2.16A,B, right panels), suggesting that loss of PAS is one mechanism that leads to loss of 3' end formation. We further probed for the conservation profile of each PAS of syntenic sites that are either conserved or not. When we consider the conserved pAs, we see a great degree of specific PAS conservation at the syntenic positions in *D. yakuba* (Fig. 2.16C, left panel), while only the canonical AAUAAA appears to be highly conserved in *D. virilis* with other PAS changing at a greater degree than in *D. yakuba* (Fig. 2.16D, left panel). It is interesting to notice that in *D. virilis* there are more instances of weaker signals, such as AUUAAA and AAUAUA converting to canonical AAUAAA, suggesting that a set of signals could be under selective pressure to be more efficient and hence evolve optimal PAS through evolution. Altogether, this analysis shows that a significant portion of sites are under selective pressure to be maintained, both positionally and at the specific PAS level.

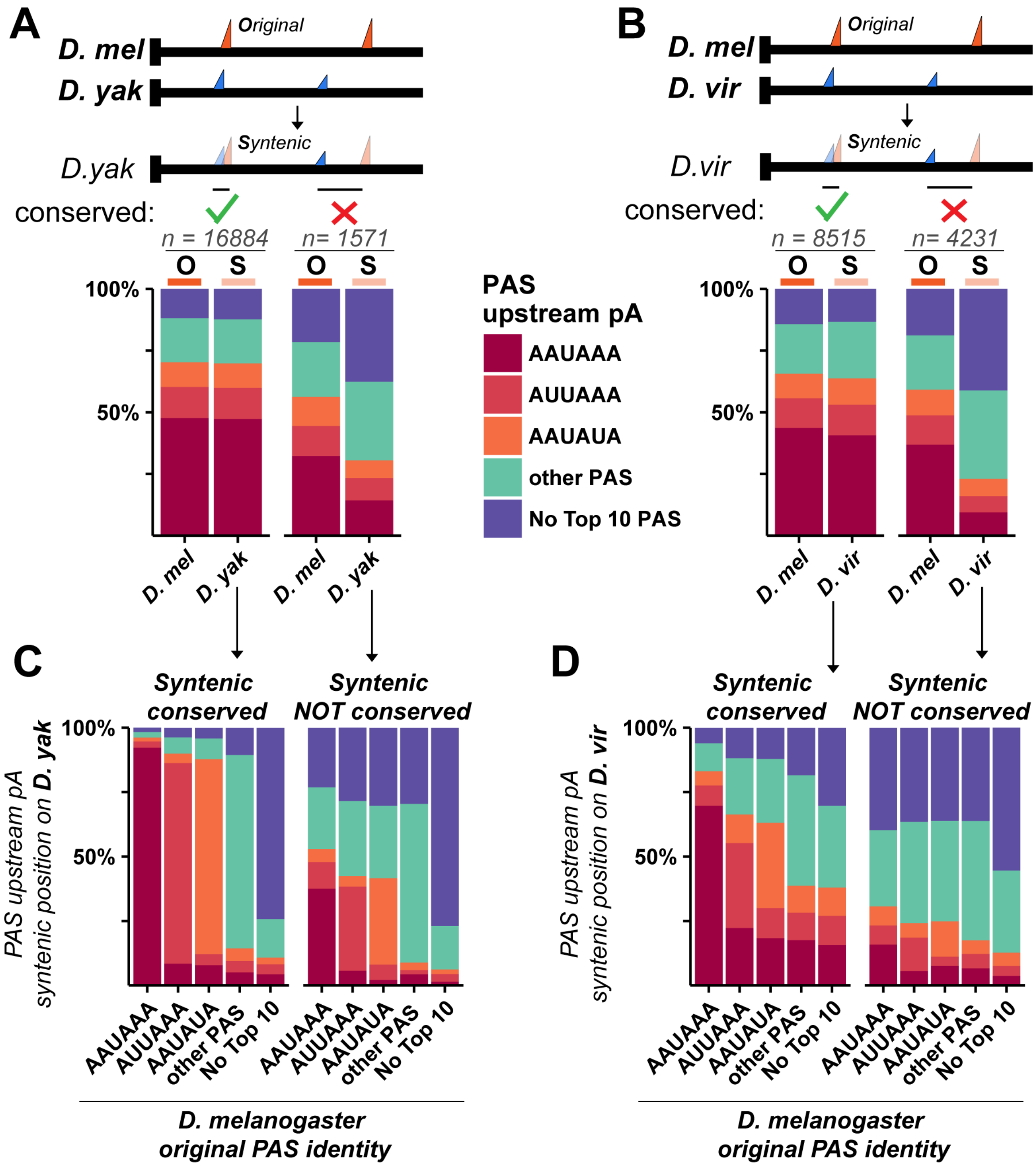


Figure 2.16 Conservation of PAS identity.
(A,B) Comparison of the identity of the PAS upstream of pA sites. Sites were considered conserved if the syntenic pA site is within 25 bp of an event in the species under comparison (green tick) or not conserved (red cross). PAS identity of the original pAs (**O**) was compared to the PAS identity of the syntenic position (**S**) in the other species. PAS identity is shown according to the legend. (C,D) Conservation of specific PAS upstream of original *D. melanogaster* pAs (x axis) in either conserved or not conserved syntenic positions.

Given that a fraction of syntenic pAs preserved the original PAS but loss evidence of 3' end formation in the other species, we tested the sequence composition around of pAs that showed syntenic conservation of AAUAAA and either evidence (Fig. 2.17A,B) or loss (Fig 2.17C,D) of a pA event in the other species, to se if other elements could have undergone change. Our analysis shows that the sites that retained the AAUAAA PAS but lost evidence of a pA event in the other species appear to have lost a U-rich DSE downstream to the cleavage site, suggesting that loss of DSE might be a mechanism adopted through evolution to inactivate pA sites.

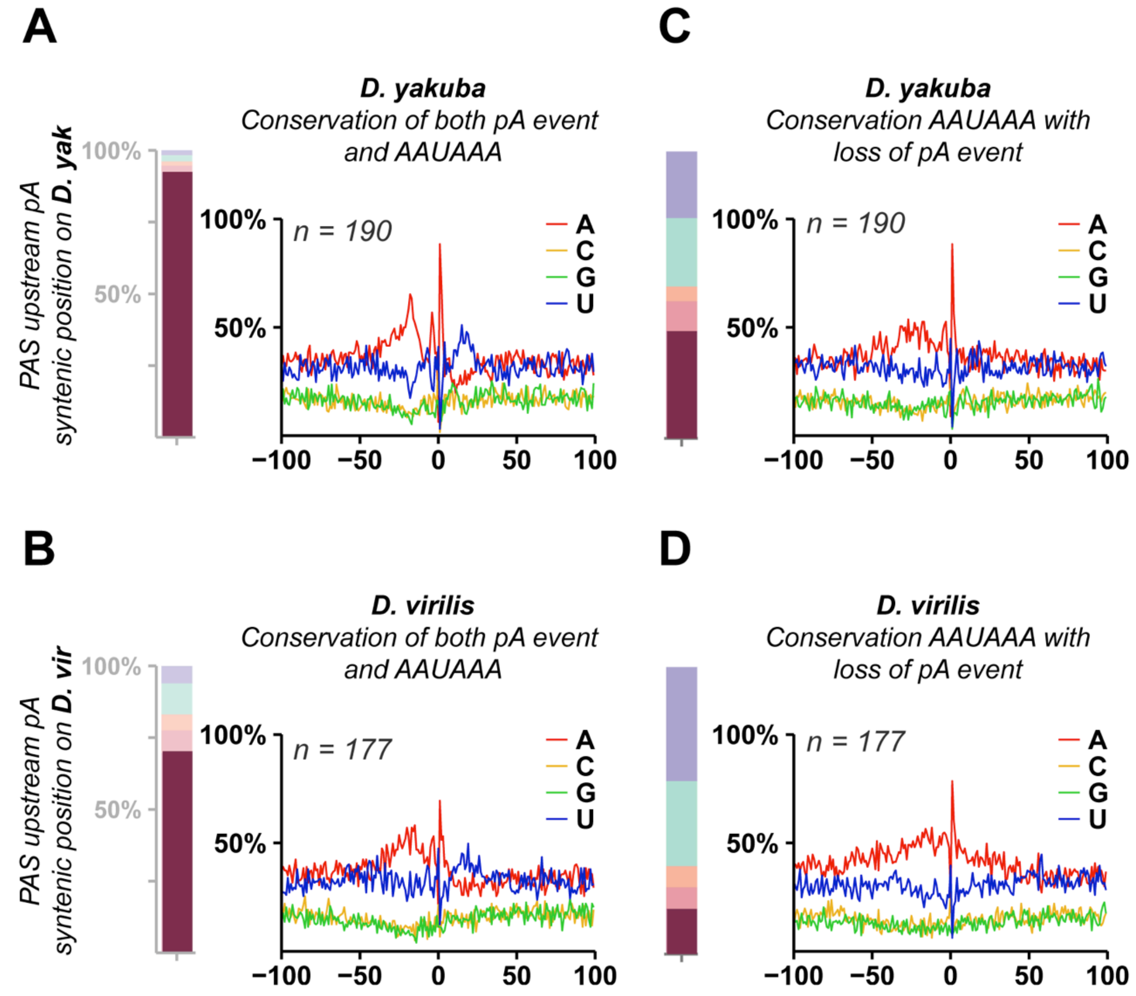


Figure 2.17 Loss of a U-rich DSE is associated with loss of cleavage and polyadenylation at syntenic sites with conserved AAUAAA PAS.
(A-D) Nucleotide distribution around syntenic position of *D. melanogaster* pAs in either *D. yakuba* (A,C) or *D. virilis* (B,D)(centered at 0 in a window of 200 nt). (A,B) Syntenic positions which exhibit both conservation of AAUAAA PAS as well as a conserved pA event in the other species. A random subset of these sites of size equal to the events in C and D were chosen. (C,D) Syntenic positions which exhibit conservation of AAUAAA PAS but loss of evidence for a pA event in the other species. Deep purple color in the bar plot shows the fraction of AAUAAA preserved at conserved or not conserved sites (taken from Fig. 2.16C,D).

To address if change in PAS observed through evolution could lead to changes in signal recognition, we compared the relative strength of the conserved orthologous pAs according to change in PAS identity in the two species to the relative strength of the original pAs in *D. melanogaster*. For the purpose of this analysis we analyzed the relative strength of pAs in head samples of conserved pAs between *D. melanogaster* and *D. virilis*. The relative strength of conserved pAs appears to change according to the change in PAS identity, as when PAS convert to better variants in *D. virilis* they associate with pA events that have a stronger relative strength, compared to the profile of the original PAS in *D. melanogaster* and vice versa (Fig. 2.18). These results were similar when comparing the relative score of conserved sites in ovary or testis (data not shown). These results support the importance of PAS strength in mediating recognition of pA sites (Sheets et al., 1990; Wilusz et al., 1989) and suggest that changes in PAS might be a strategy to modulate the expression of different 3' end isoforms through evolution in *Drosophila*.

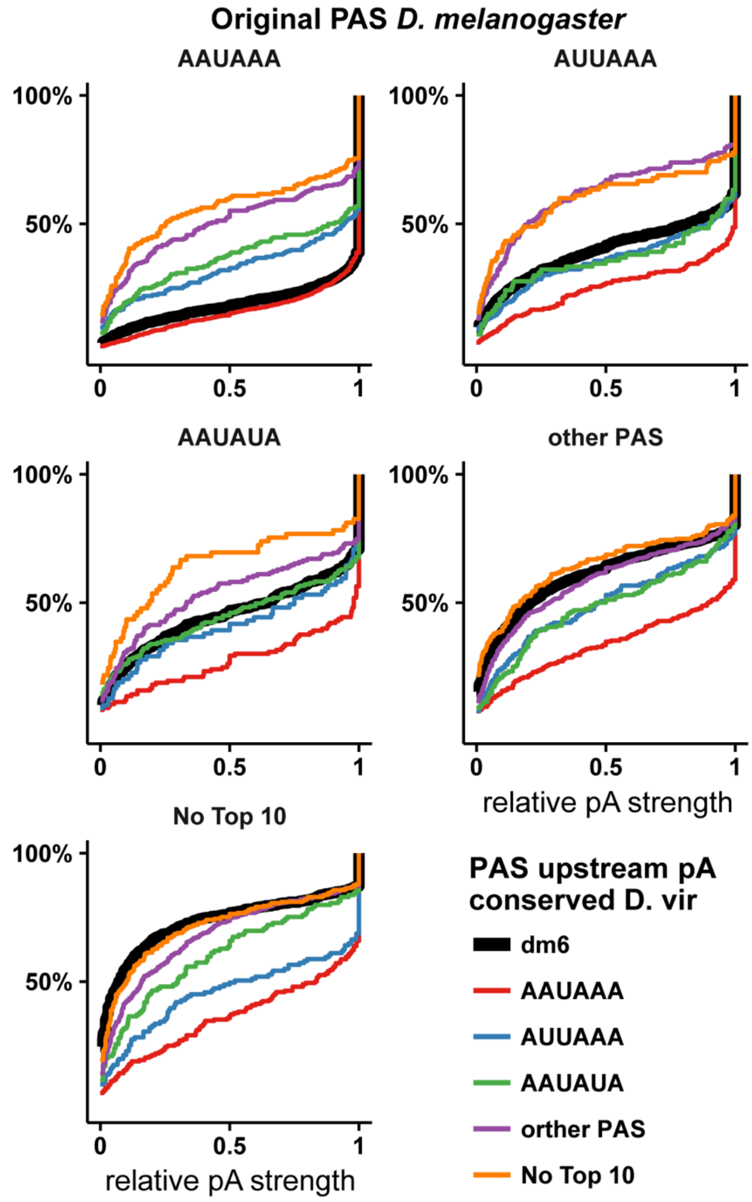


Figure 2.18 Relative strength of pAs change according to change in PAS. Empirical cumulative distribution functions of the relative strength of conserved pAs between *D. melanogaster* and *D. virilis*. Different graphs are subsets of conserved pAs according to the identity of the original *D. melanogaster* PAS. Relative strength of the conserved pAs in *D. virilis* are plotted according to PAS identity of conserved pAs in that species (color). The relative strength distribution of the original pAs in *D. melanogaster* is shown in thick black line. PAS identity of the original pAs in *D. melanogaster* is shown at the top of the graph. The relative strength score represents the fraction of isoforms ending at a specific site compared to downstream sites. This was calculated from the fraction of reads at each pA out of all reads at that site or downstream ones.

Discussion

A complex landscape of 3' UTR isoforms expression

Despite almost two decades of genome-wide efforts to annotate the *Drosophila* transcriptome, the 3' ends of genes have eluded careful characterization (Matthews et al., 2015). This in spite of recent reports of tissue specific expression of specific 3' end isoforms in *Drosophila* and RNA-seq based estimates of 3' end isoform complexity that suggested that at least 50% of all genes express multiple 3' end isoforms (Hilgers et al., 2011b; Smibert et al., 2012). Here, we deploy a 3'-seq protocol that we developed to generate a genome-wide atlas of sites of polyadenylation. We annotate additional 3' UTR isoforms for thousands of genes and provide a dataset that can be mined to characterize intronic polyadenylation as well as the polyadenylation status of ncRNAs. We expand the number of genes that undergo APA to more than 65% of all genes, bringing *Drosophila* in line with what has been observed in other high eukaryotes (Shi, 2012). We provide an atlas of 3' end isoforms that covers 97% of all expressed genes, which we hope will be incorporated in a future annotation of *D. melanogaster*. Additionally, we generated 3' end annotation of head, ovary and testis, as well as time-points that cover all of embryonic development, for two additional species of *Drosophila*. We believe that these combined datasets will be of great interest to both the *Drosophila* community as well as to scientists interested in investigating aspects of the evolutionary conservation of 3' UTR isoform expression and formation.

Despite the lack of a next-generation sequencing based annotation of *Drosophila* APA isoforms, the character of the most utilized PAS was already investigated through analysis of EST data (Graber et al., 1999; Retelska et al., 2006). Strikingly, these early analyses were already able to define the PAS strength hierarchy that we identify in our study. Not surprisingly, the AAUAAA hexamer acts as the canonical PAS in *Drosophila* as in all other high eukaryotes investigated to date, and as predicted at the time of its discovery in 1976 (Proudfoot and Brownlee, 1976). We confirm the previous observation of a significant use of the AAUAUA hexamer in *Drosophila* and identify this variant upstream of 10% of all pAs. Interestingly, this PAS does not appear to be significantly present upstream of cleavage sites in neither *C. elegans* (Jan et al., 2011) or humans (Derti et al., 2012). This PAS appears to be functional in *Drosophila*, as shown in experiments that mutated the proximal AAUAUA of *Dl* to AAGAGA, causing loss of proximal 3' UTR isoform expression and sole recognition of the distal PAS (Shepherd et al., 2010). The higher use of the AAUAUA PAS to define 3' ends in *Drosophila* and perhaps in insects more generally, brings question on the nature of the increase in recognition of this specific variant. Future studies will be needed to understand the molecular basis of this increase in recognition, which could occur through a difference in the specificity of the cleavage and polyadenylation machinery in *Drosophila* or the binding of RBPs that cooperate with recognition of this PAS, among several possible mechanisms.

One of the most intriguing aspects of the expression of APA isoforms in eukaryotes is the specific accumulation of isoforms in a tissue and length specific manner. This was observed early on through EST data analysis (Gautheret et al., 1998; Tian et al., 2005), which was subsequently confirmed through sequencing of 3' end isoforms (Derti et al., 2012; Jan et al., 2011) and RNA-seq analysis (Smibert et al., 2012). We greatly expand the number of genes that express 3' UTR isoforms in a tissue specific manner in *Drosophila*. Confirming our previous observations, head samples express longer 3' UTR isoforms when compared to all tissues we analyzed, and indeed we identify hundreds of additional genes exhibiting this pattern of expression. However, not all transcripts that express longer isoforms in head samples are head specific, as many of these appear to express in an analogous manner in somatic tissues found in carcass and cell lines. We identify ovary specific shortening events, which brings into question if female germline specific dynamics of 3' UTR isoforms expression might be a conserved aspect of gene expression, as this was also observed in zebrafish (Ulitsky et al., 2012). Since the lengthening of these germline 3' UTRs in *Drosophila* appears to occur with the advent of zygotic transcription (Tadros and Lipshitz, 2009), it will be interesting to investigate the role of APA in the deposition of maternal RNA in the oocyte and the transition from maternal to zygotic transcription. Germline expression of specific 3' UTR isoforms appears to be very complex and subject to change in evolution. For example, expression of ovary specific short isoforms occurs in zebrafish, where these isoforms are found to be the shortest amongst all tissues analyzed, including testis, while in

Drosophila we observe the shortest 3' UTR expression pattern in testis. On the other hand, the expression of shorter 3' UTRs in testis appears to be more conserved amongst metazoan organisms (Bao et al., 2016; Li et al., 2016; MacDonald and McMahon, 2010), however possibly more labile to evolution of the specific termini (Fig 2.14). The role that these tissue specific dynamics play in biology is just beginning to emerge and will be a very active field of investigation in the near future.

Conservation and evolution of APA in Drosophila

Conservation of sites of cleavage and polyadenylation has been investigated in mammalian systems before (Derti et al., 2012), showing conservation of tissue specific patterns. Here, we provide 3' end evidence for head, ovary, and testis of three different species of *Drosophila*, at a depth much greater than previously achieved in studies of evolutionary conservation of APA (Derti et al., 2012), allowing deeper analysis of site-specific conservation. Our analysis begins to mine this rich dataset, demonstrating great conservation of patterns of longer 3' UTRs expression in heads and shortest in testis. Furthermore we identify many syntenic positions that are conserved for 3' end formation between species and observe a strong role of the PAS and DSEs to mediate conservation or change in the degree of pA recognition. This is only the beginning of an analysis that promises to uncover important cis-regulatory elements that mediate tissue specific accumulation of 3' end isoforms. As part of this effort, we have identified thousands of deeply conserved predicted microRNA binding sites in our extended 3' UTR annotation (data not shown), largely extending our understanding of possible biologically relevant sites of post-transcriptional regulation.

A recent study of quantitative trait loci that affect the expression of alternative 3' UTR isoforms uncovers many sites under genetic control, identifying variants that cluster around pAs as well as putative RBP binding motifs that appear to have a role in specific 3' end isoforms expression (Cannavò et al., 2017). This and our study confirm the importance of cis-regulatory elements in modulating the expression of specific isoforms, some of which play significant biological function *in vivo* (Cannavò et al., 2017). Our comprehensive annotation of 3' UTRs in *Drosophila* sets the base for using the fly as a model for understanding the regulation of tissue specific APA and the functional ramifications of this complex phenomenon in multiple aspects of biology.

CHAPTER 3: Regulation of 3' UTR expression by the Elav family of RNA binding proteins in *Drosophila*

Summary

The expression of long 3' UTR isoforms in the nervous system has recently emerged as a conserved aspect of gene expression. However, the mechanism that controls the expression of these isoforms is not well understood. Here, we address the role of Elav, an RBP proposed to regulate 3' UTR extensions in *Drosophila*, in mediating this phenomenon. Analysis of gain and loss of function of Elav provides evidence that this RBP controls 3' UTR expression for hundreds of targets, however it does not appear to be necessary to express CNS specific long 3' UTR isoforms. We further extend this function to known Elav paralogs Rbp9 and Fne and propose a redundant role for these factors in mediating 3' UTR extensions in the nervous system. Finally, computational analysis of putative Elav regulated pAs provides further evidence that extension of 3' UTRs by Elav might be in part mediated through the inhibition of pA recognition.

Introduction

The expression of mRNA isoforms with differential 3' UTR length through APA has emerged as a dominant property of eukaryotic transcriptomes where up to 75% of all genes express this class of isoforms (Shi, 2012). The expression of 3' UTR isoforms is strongly biased depending on differentiation state, tissue and disease state (Tian and Manley, 2017). The nervous system in particular shows a striking pattern of 3' UTR isoform expression, where hundreds of genes express much longer 3' UTRs when compared to other tissues (Miura et al., 2014; Tian and Manley, 2017; Zhang et al., 2005). The expression of longer 3' UTR isoforms in the nervous system appears to be a property of neural transcriptomes deeply conserved through evolution, with neural extension of 3' UTRs being observed from fly to human (Hilgers et al., 2011b; Miura et al., 2013; Smibert et al., 2012). Expression of these differential isoforms has an impact on the regulatory potential of the 3' UTRs of these genes, changing the number and identify of regulatory elements such as miRNA binding sites or RBP motifs (Dai et al., 2015; Smibert et al., 2012). Despite the potential biological importance of the expression of different 3' UTR isoforms in neurons, for example as evidenced by studies implicating differential 3' UTR expression to transcript localization and local translation (An et al., 2008b; Yudin et al., 2008), little is known about the mechanisms that lead to the differential expression of these isoforms in the nervous system.

Several mechanisms involved in regulating the steady state accumulation of 3' UTR isoforms have been described in recent years. These mechanisms include changes in the levels of core components of the cleavage and polyadenylation machinery (Takagaki et al., 1996), promoter mediated modulation of signal recognition through differential recruitment of trans factors (Calvo and Manley, 2001; Dantonel et al., 1997) or modulation of RNA polymerase II elongation rate (Pinto et al., 2011) to modulation of PAS recognition by the differential binding of RBPs proximal to regulated PAS (Batra et al., 2014; Masuda et al., 2015). The regulation of APA by RBPs has received a lot of attention, as the list of RBPs that are able to impact the identity of 3' UTR isoforms expressed keeps rising (Zheng and Tian, 2014).

One RBP class that has been implicated in modulating 3' end processing early on, prior to the realization of widespread regulated APA, is the Elav/Hu family of RBPs. Elav, the founding member of this family of RBPs. was first discovered in *Drosophila* in 1985 as a mutation leading to embryonic lethality and impairment of the architecture of the visual system (Campos et al., 1985). Elav, as well as the other members of the family, contain 3 conserved RRM motifs (Samson and Chalvet, 2003). These domains mediate a plethora of RNA processing and metabolism events, from regulation of alternative splicing and stability to localization (Colombrita et al., 2013). In *Drosophila*, Elav as well as two other paralogs, Fne and Rbp9 are predominantly found in the nervous system where they were shown to mediate neural specific splicing while Rbp9 is also expressed in gonads and has been shown to mediate stabilization of transcripts in the ovary (Colombrita et al., 2013; Kim-Ha et al., 1999). The binding of *Drosophila* Elav in proximity of intronic PAS was first shown to lead to inhibition of 3' end processing, resulting in splicing into downstream exons of *ewg* (Lisbin et al., 2001; Soller, 2003). A similar role of Hu family RBPs in humans was shown early on in HeLa cells where Hu proteins can block cleavage and polyadenylation in the presence of U-rich sequences adjacent to the PAS (Zhu et al., 2007). Given the predominant neuronal expression of most of the Elav/Hu family of RBPs, these proteins are good candidate regulators of CNS specific 3' UTRs extension. Indeed, recent studies have shown that Elav can mediate the extension of 3' UTRs of a subset of genes. Hilgers and colleagues have shown that loss of Elav during embryogenesis correlates with a loss of 3' UTR extensions of 7 genes and ectopic overexpression of Elav outside the nervous system in embryonic tissue leads to the expression of long 3' UTR isoforms (Hilgers et al., 2011a). The regulation of 3' UTR length by Elav appears to be in part mediated by differential recruitment of Elav at the promoter of genes that express long 3' UTR isoforms (Oktaba et al., 2015). However, identification of putative regulatory Elav binding motifs on extended 3' UTR targets by computational means has failed to identify potential regulatory sequence elements (Oktaba et al., 2015) and our understanding of the breadth of genes that are responsive to Elav mediated regulation of 3' UTR length remains unclear.

To gain a more comprehensive understanding of the scope of Elav mediated 3' UTR isoform regulation and attempt to elucidate the mechanism that mediates the steady state increase of long 3' UTR isoforms in the nervous system we turned to genome-wide approaches. Here, we created an S2 cell based system that can recapitulate the lengthening of 3' UTRs observed in the nervous system by overexpression of Elav. However, loss of Elav *in vivo* in larval CNS fails to show complete loss of expression of long 3' UTR isoforms, suggesting that the Elav paralogs Rbp9 and Fne might play a redundant role in regulating these isoforms. Indeed, we uncover that Elav family proteins Rbp9 and Fne mediate Elav like 3' UTR extensions. Finally, we identify candidate pAs that could be under Elav regulation. Computational analysis of sequences around candidate Elav regulated sites reveals an enrichment of U-rich elements downstream of the PAS, similar to consensus sequences characterized to bind this family of RBPs. Our data favors a model in which Elav family proteins can suppress recognition of proximal PAS followed by U-rich DSEs leading to the extension of long 3' UTRs in the nervous system.

Results

Elav can regulate expression of longer 3' UTR isoforms in S2 cells

We and others have shown that hundreds of genes in *Drosophila* express 3' UTR isoforms that are longer in the nervous system when compared to other tissues (Chapter 2) (Hilgers et al., 2011b; Smibert et al., 2012). Recently, the RBP Elav has been implicated in mediating neural specific extension of 3' UTRs (Hilgers et al., 2012; Oktaba et al., 2015). However, the scope of Elav regulation of 3' UTR length of genes expressed in the CNS and the underlying mechanism remain unclear. To begin to address both of these questions we decided to develop a cell-based system to investigate the role of Elav in a gain of function setting, to test if we could recapitulate the pattern of 3' UTR isoforms expression seen in the nervous system. To this aim, we chose S2 cells, a commonly used *Drosophila* cell line which is embryo derived and exhibits a hemocyte-like gene expression profile (Cherbas et al., 2011), allowing us to test the consequence of Elav expression in a non-neuronal setting. As a control, we engineered previously known point substitutions in all 3 RNA binding RRM domains of Elav, which have been shown to impair both the binding of Elav to RNA and the ability of these mutants to rescue loss of function Elav mutations *in vivo* (Lisbin et al., 2000). To test the ability of Elav to mediate expression of mRNA isoforms with longer 3' UTRs, we transiently transfected WT Elav or the 3X RRM mutant (3X-MT) under the control of a ubiquitous *actin* promoter (Fig. 3.1A). Western blot analysis shows stable expression of the two Elav constructs at the expected ~55 kDa size (Fig. 3.1B). Northern blot analysis of genes known to express long 3' UTRs in the CNS shows steady state accumulation of longer 3' UTR isoforms, similarly to what is seen in the nervous system, as shown for *goα* and *AcCoAs* (Fig. 3.1C). These results confirm the ability of Elav to induce the expression of longer neural like 3' UTR isoforms in ectopic settings, similarly to what seen with mis-expression of Elav in embryonic ectoderm (Hilgers et al., 2012).

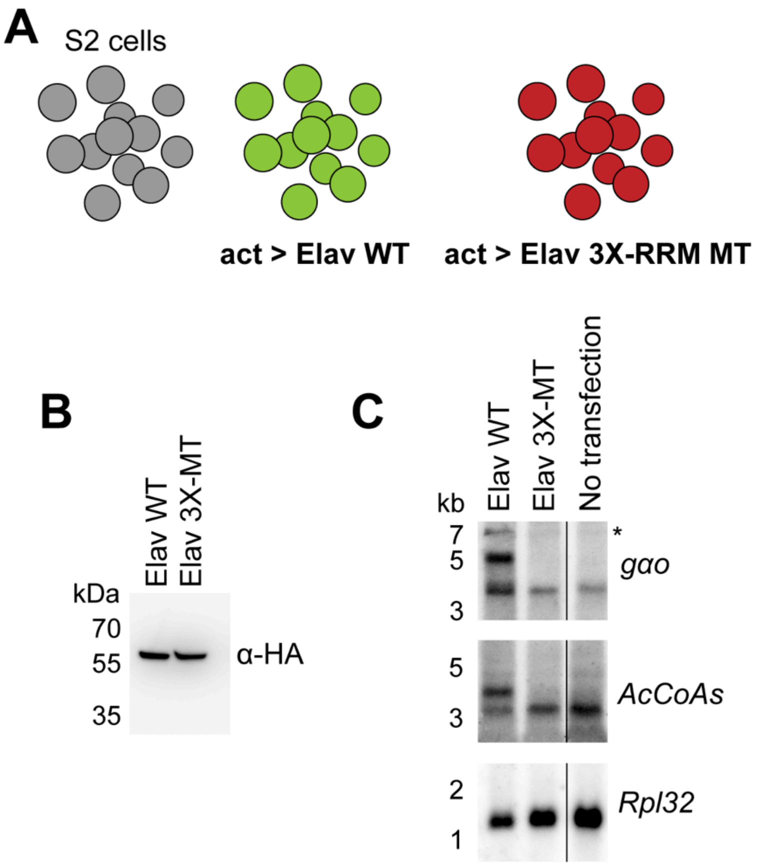


Figure 3.1 Elav over-expression in S2 cells induces expression of 3' UTR isoforms with longer 3' UTRs.

(A) Schematic of S2 cell based system where we either over-express WT Elav (green) or an Elav mutant version which has mutations in all 3 RRM domains (3X-MT) which abrogate binding to RNA (red) (Lisbin et al., 2000). (B) Western blot showing expression of the Elav constructs with HA antibody. (C) Northern blotting of cells transfected with either Elav WT or 3X-MT using probes against goα, AcCoAs or Rpl32 as a control. Asterisks denote non specific bands. (All experiments performed by Sonali Majumdar)

To capture the genome-wide effect of Elav on its targets, we performed 3'-seq of S2 cells over-expressing either WT Elav or 3X-MT Elav. Analysis of weighted 3' UTR length shows hundreds of genes changing pattern of 3' UTR isoform expression, with a majority of genes expressing longer 3' UTRs in the presence of WT Elav (Fig. 3.2A). In agreement with our Northern blot analysis, the 3X-MT Elav fails to change to a great degree 3' UTR length when expressed in S2 cells (Fig. 3.2B) and a similar number of genes show a change in 3' UTR isoforms expression when we compare WT and 3X-MT Elav (Fig 3.2C), confirming that the 3X-MT Elav mutant does not support accumulation of long 3' UTR isoforms. These results expand the number of genes susceptible to Elav regulation when ectopically expressed outside of the nervous system. To understand the degree to which ectopic expression of Elav can recapitulate the pattern of longer 3' UTR expression seen in the head we looked at genes that we previously found to express longer 3' UTR isoforms in head compared to S2 cells (Fig 2.8A). When we overlay this subset of gene onto our analysis, we detect only half of these genes expressing longer 3' UTRs upon over-expression of Elav in S2 cells (Fig. 3.2D). For example, the gene *ps* expresses the longest 3' UTR isoform seen in head samples upon over expression of Elav (Fig. 3.2E) while *vsg* fails to express the longer 3' UTR isoform found in head in an Elav dependent manner (Fig. 3.2F), suggesting that Elav alone in S2 cells is not sufficient to lead to the expression of longer 3' UTR isoforms of a subset of head extended genes. These observations point to the possibility that factors other than Elav might be necessary to express longer isoforms of a subset of neural extended genes.

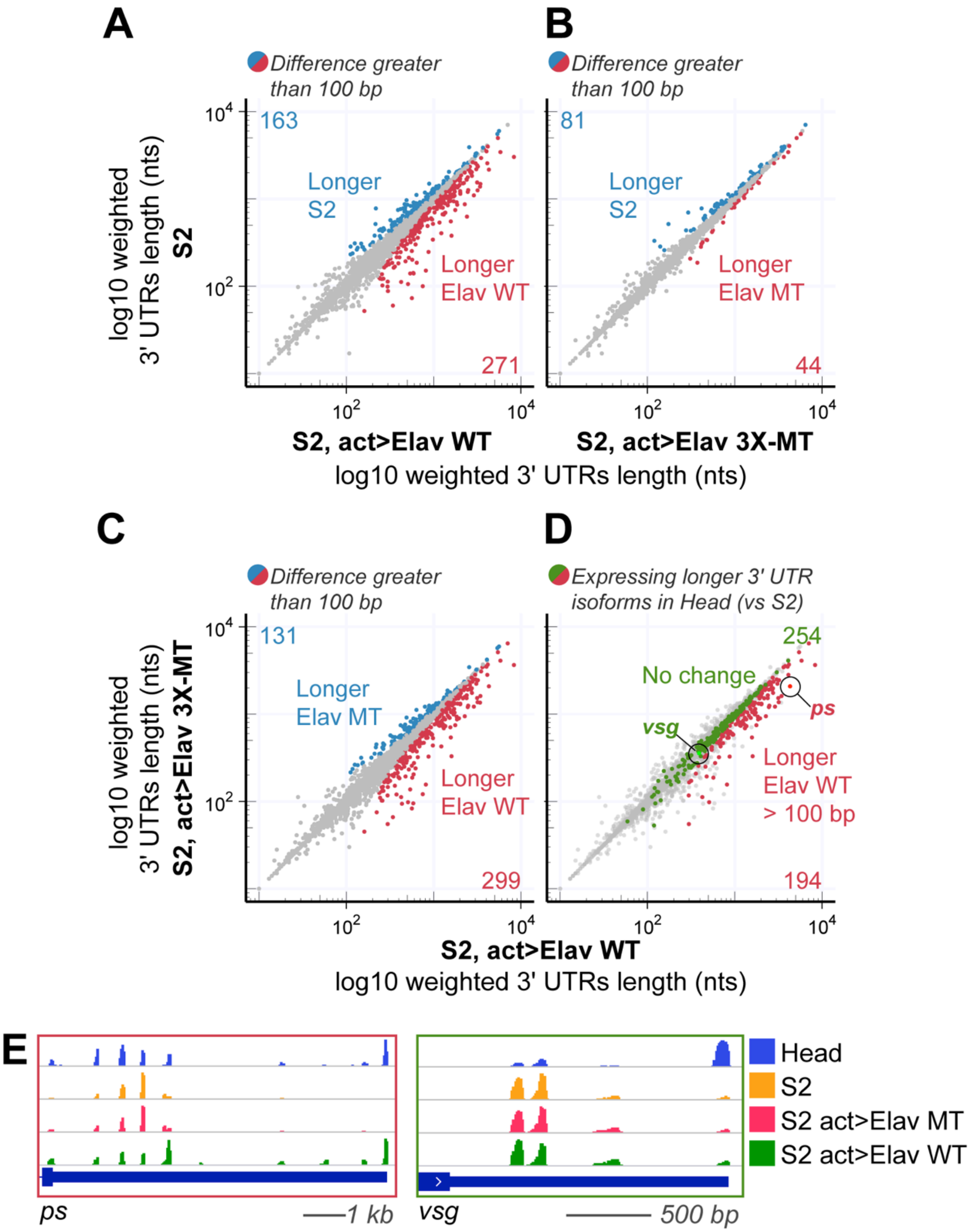

A
Difference greater than 100 bp
163
Longer S2
Longer Elav WT
271
10^4
10^2
log10 weighted 3' UTRs length (nts)
S2
S2, act>Elav WT
B
Difference greater than 100 bp
81
Longer S2
Longer Elav MT
44
S2, act>Elav 3X-MT
log10 weighted 3' UTRs length (nts)
C
Difference greater than 100 bp
131
Longer Elav MT
Longer Elav WT
299
log10 weighted 3' UTRs length (nts)
S2, act>Elav 3X-MT
D
Expressing longer 3' UTR isoforms in Head (vs S2)
254
No change
vsg
ps
Longer Elav WT > 100 bp
194
S2, act>Elav WT
log10 weighted 3' UTRs length (nts)
E
Head
S2
S2 act>Elav MT
S2 act>Elav WT
ps
1 kb
vsg
500 bp

Figure 3.2 Elav mediates extension of 3' UTRs of only a subset of neural extended genes.

(A-D) Weighted 3' UTR length comparison between samples. Weighted 3' UTR length is obtained taking the average of all 3' UTR isoforms length per gene weighted by the contribution of each isoform expression. Genes are expressed at a minimum of 5 RPM in all samples. (A-C) The genes which weighted 3' UTR length differs by 100 bp or more between samples are shown in red when longer weighted length is in the sample on the x axis, blue when longer weighted length is in the sample on the y axis. (D) Genes that were longer in head when compared to S2 cells are overlaid on top of the comparison of over-expression of Elav WT vs. 3X-MT in S2 cells. In green are genes that were longer in head compared to S2 cells but do not change pattern of 3' UTR isoforms when over-expressing WT Elav. In red are genes that express longer isoforms upon over-expression of Elav WT, as seen in head samples. (A,B) Comparison of S2 cells vs. over-expression of either Elav WT (A) or Elav 3X-MT (B). (C,D) Comparison of over-expression of Elav WT vs. 3X-MT in S2 cells. (E) Examples of 3'-seq of genes that express longer 3' UTR isoforms upon over-expression of Elav WT (red frame) or which do not respond to over-expression of WT Elav (green frame. Samples are identified as shown in the legend. Scale for the genomic range is shown in grey. Images were taken using the IGV genome browser. (3'-seq libraries generated by Sonali Majumdar)

Elav is not essential for the expression of a majority of long 3' UTRs in vivo

In the process of analyzing the previously published *elav*^{5} LOF allele, which is a deletion of the *elav* locus (Yao et al., 1993), we discovered that contrary to the reported embryonic lethal phenotype, the embryos can develop up to the 1st instar larval stage (L1), when they die presumably because they cannot escape the eggshell. We took advantage of this observation to address the role of Elav in shaping the pattern of 3' UTR expression seen in the nervous system *in vivo* by analyzing the transcriptome of dissected L1 CNS. As expected, we cannot detect Elav expression by IHC in isolated L1 CNS from homozygote animals (Fig. 3.3A,B). Furthermore, BP102 staining, which visualizes neuronal morphology and axon projections (Hummel et al., 2000), shows aberrant crossing of axons through the CNS midline upon loss of Elav, a phenotype previously recorded in Elav MT embryo CNS (Simionato et al., 2007). We first generated RNA-seq libraries of L1 CNS from WT, *elav*^{5} (LOF) and *elav*^{5} rescued with a genomic insertion containing the whole *elav* locus (*elav*^{5} rescue). As expected, we do not detect any signal arising from the *elav* locus in the RNA-seq from the mutant L1 CNS, but recover expression in rescued animals (Fig 3.3C).

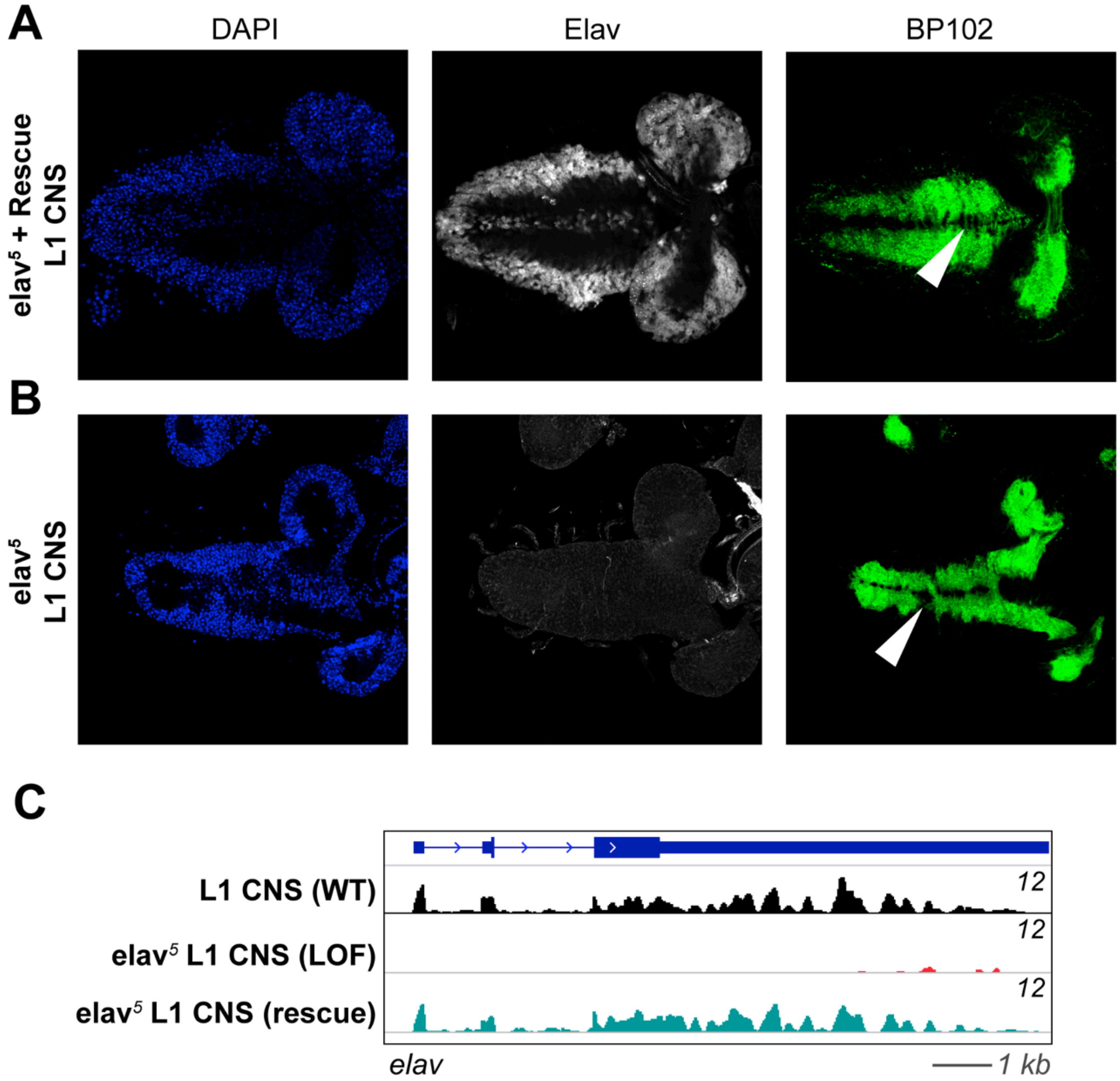


Figure 3.3 Characterization of Elav^5 LOF allele.

(A,B) Immunohistochemistry of L1 CNS dissected 24h after the start of embryogenesis, at the onset of the 1st instar larval stage (L1). L1 CNS is stain with DAPI (blue) to visualize cell nuclei, Elav (grey) which is a marker of neurons and BP102 (green) antibodies, which visualizes axonal projections. Arrows point to commissural axons projections which are aberrant in the elav LOF sample and properly cross the midline upon rescue of Elav expression. (C) RNA-seq of the *elav* locus, normalized to the same RPM value. Scale for the genomic range is shown in grey. Images were taken using the IGV genome browser. (L1 CNS dissected by Alexandra Panzarino)

Following mutant validation, we turned to 3'-seq to query the effect of Elav on expression of long 3' UTR isoforms in the CNS by performing 3'-seq on *elav*^{5} MT and *elav*^{5} rescue L1 CNS. Analysis of the weighted length of 3' UTRs between the two samples shows fewer genes showing expression of longer 3' UTR isoforms (Fig. 3.4A) than observed when comparing gonads, cell lines and somatic samples to head (Chapter 2). We observe a decrease in the expression of extended 3' UTR isoforms for genes such as *Imp* and *mtd* (Fig. 3.4B,C), however we do not observe complete loss of extensions, as previously reported for Elav regulated genes (Hilgers et al., 2012). This is true for *ps* as well, a gene that shows absence of expression of the longest isoform in S2 cells but which is responsive to Elav over-expression (Fig. 3.2E). We also detect gain of 3' UTR extensions upon loss of Elav for a handful of genes, such as *AP2-α* (Fig. 3.4E). These results show a more complex role of Elav in shaping the expression pattern of 3' UTR isoforms expression and suggest that there are other factors that can mediate the expression of long 3' UTR isoforms in L1 CNS. It is important to notice that loss of Elav does not seem to cause a significant change in the expression levels of genes that undergo a change in the expression of 3' UTR isoforms. As expected, many of the isoforms of genes that undergo a change in 3' UTR isoform expression show significant change in gene expression (Fig. 3.4F), however only a handful appear to significantly change in expression level at the gene level (Fig. 3.4G). These results support a model in which the observed change in 3' UTR isoform expression is caused by differential recognition of pAs rather than differential stabilization of 3' UTR isoforms.

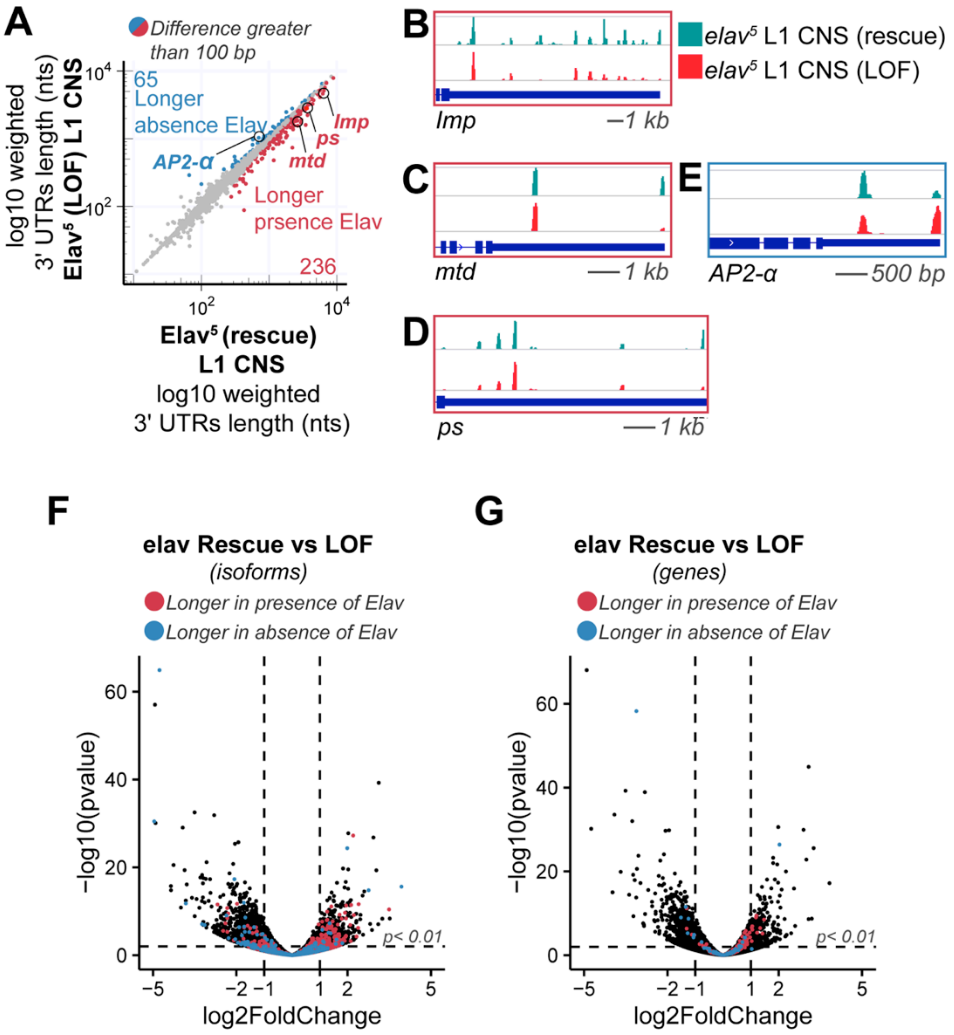


Figure 3.4 Loss of Elav in L1 CNS does not lead to loss of expression of long 3' UTR isoforms.

(A) Weighted 3' UTR length comparison between samples. Weighted 3' UTR length is obtained taking the average of all 3' UTR isoforms length per gene weighted by the contribution of each isoform expression. Genes are expressed at a minimum of 5 RPM in all samples. Genes which express longer weighted 3' UTRs by at least 100 bp are shown in red when they are longer in the presence of Elav and blue when they express longer isoforms in the absence of Elav. (B-E) Example of 3'-seq of genes that express less long 3' UTR isoforms upon loss of Elav (B-D, red frame) or show expression of more longer isoforms (E, blue frame) in the absence of Elav. Scale for the genomic range is shown in grey. Images were taken using the IGV genome browser. (F,G) Volcano plot of differential gene expression analysis of either 3' end isoforms (F) or genes (G). Genes that show expression of longer 3' UTR isoforms in the presence of Elav are shown in red while genes that show expression of longer 3' UTR isoforms in the absence of Elav are shown in blue. (F) Each point represents one 3' end isoform. (G) Each point represents a gene.

To address the discrepancy between the previously published role of Elav as a master regulator of neural 3' UTR extensions and our *in vivo* analysis of loss of Elav in the L1 CNS, which showed persistence of the expression of long 3' UTR isoforms, we wondered if known Elav paralogs, Rbp9 and Fne, could play a redundant role with Elav to extend 3' UTRs in the L1 CNS. All of the three Elav family RBPs are mostly restricted to the nervous system. Previous studies have shown that the 3 members of the Elav family are expressed in a temporal fashion, where Elav is expressed most highly early in embryogenesis, Fne appears mostly during late embryogenesis and during larval and pupal development while Rbp9 is the dominant RBP in pupal and adult stages (Zaharieva et al., 2015). To test the ability of Fne and Rbp9 to mediate 3' UTR extensions, we expressed either WT or 3X-MT of the two proteins in S2 cells. We verified protein expression (Fig. 3.4A) and observed extension of the 3' UTR of known Elav responsive targets, similarly to the pattern observed upon expression of Elav (Fig. 3.4B). Indeed, 3'-seq analysis of S2 cells over-expressing these two paralogs shows similar extension of the 3' UTRs of hundreds of genes, as seen for Elav (Fig. 3.4C-H). These results suggest that the other members of the Elav family, Fne and Rbp9 could play a redundant role with Elav to mediate 3' UTR extension and could explain the mild effect of loss of Elav in the L1 CNS, as both are expressed at this stage.

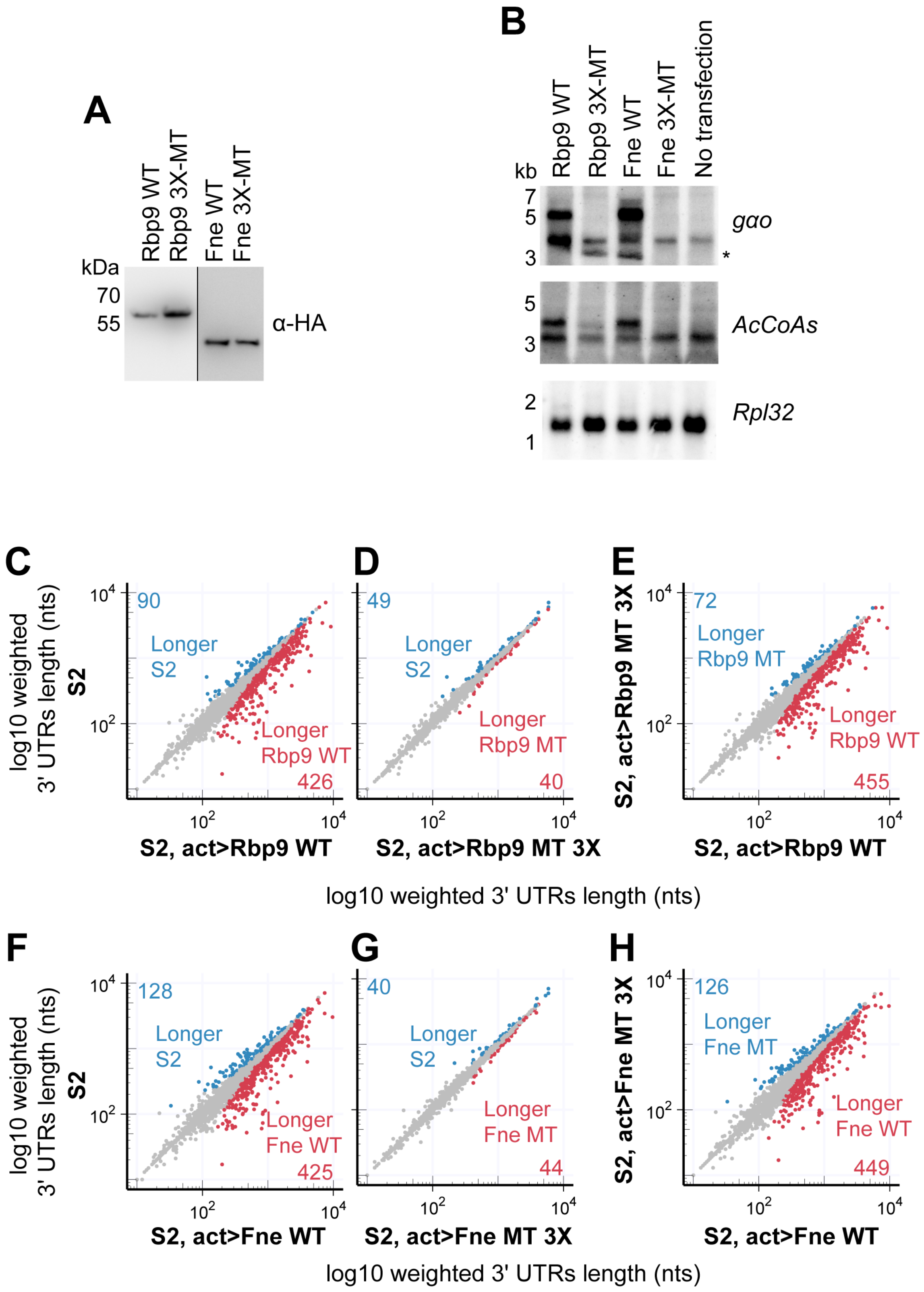

A
Rbp9 WT
Rbp9 3X-MT
Fne WT
Fne 3X-MT
kDa
70
55
α-HA
B
kb
Rbp9 WT
Rbp9 3X-MT
Fne WT
Fne 3X-MT
No transfection
7
5
3
gαo
*
5
3
AcCoAs
2
1
Rpl32
C
D
E
F
G
H
log10 weighted 3' UTRs length (nts)
S2
10^4
10^2
90
Longer S2
Longer Rbp9 WT
426
S2, act>Rbp9 WT
49
Longer S2
Longer Rbp9 MT
40
S2, act>Rbp9 MT 3X
S2, act>Rbp9 MT 3X
72
Longer Rbp9 MT
Longer Rbp9 WT
455
S2, act>Rbp9 WT
log10 weighted 3' UTRs length (nts)
128
Longer S2
Longer Fne WT
425
S2, act>Fne WT
40
Longer S2
Longer Fne MT
44
S2, act>Fne MT 3X
S2, act>Fne MT 3X
126
Longer Fne MT
Longer Fne WT
449
S2, act>Fne WT
log10 weighted 3' UTRs length (nts)

Figure 3.5 The Elav paralogs Rbp9 and Fne can induce expression of longer 3' UTR isoforms in S2 cells.

(A) Western blot showing expression of the Rbp9 and Fne construct with HA antibody. (C) Northern blotting of cells transfected with either Rbp9 or FNE, WT or 3X-MT using probes against goα, AcCoAs or Rpl32 as a control. Asterisks denote non specific bands. (C-H) Weighted 3' UTR length comparison between samples. Weighted 3' UTR length is obtained taking the average of all 3' UTR isoforms length per gene weighted by the contribution of each isoform expression. Genes are expressed at a minimum of 5 RPM in all samples. (C-H) The genes which weighted 3' UTR length differs by 100 bp or more between samples are shown in red when longer weighted length is in the sample on the x axis, blue when longer weighted length is in the sample on the y axis. (C,D) Comparison of S2 cells vs. over-expression of either Rbp9 WT (C) or Rbp9 3X-MT (D). (E) Comparison of over-expression of Rbp9 WT vs. 3X-MT in S2 cells. (F,G) Comparison of S2 cells vs. over-expression of either Fne WT (F) or Fne 3X-MT (G). (H) Comparison of over-expression of Fne WT vs. 3X-MT in S2 cells. (All experiments, excluding computational analysis, were generated by Sonali Majumdar)

Given that Elav/Hu family proteins are known to stabilize AU-rich transcripts (Fan and Steitz, 1998; Peng et al., 1998), we wondered if transcript stabilization might play an important role in mediating extension of alternative expression of 3' UTRs upon expression of Elav in S2 cells. To assay the effect of Elav paralog over-expression on gene expression in S2 cells, we pulled the WT RBP samples as pseudo-replicates compared to the 3X-MT samples to gain some statistical power to our analysis. When we look at differentially expressed genes, we observe a population of genes that change expression, with a predominance of genes showing an increase in expression upon expression of the Elav family RBPs (Fig. 3.6A). This suggests that indeed Elav family proteins might predominantly act as stability factors. Next, we asked if the genes that undergo a change in 3' UTR isoform expression, might also increase in expression, suggesting that Elav might lead to expression of long 3' UTR isoforms by differentially stabilizing transcripts. When we look at the genes that change in weighted length when over-expressing Elav WT compared to Elav 3X-MT, we only see a minority of genes undergoing a significant change in expression levels (Fig 3.6B). These results suggest that Elav family proteins are more likely regulators of APA rather than stability of alternatively expressed 3' UTR isoforms.

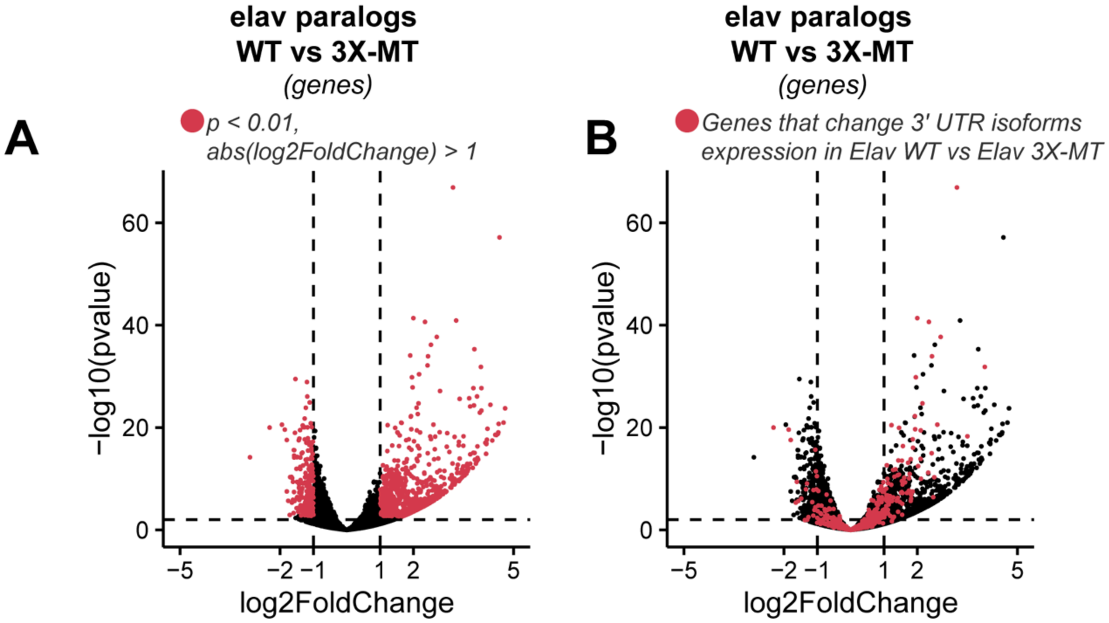


Figure 3.6 Most genes that express Elav dependent longer 3' UTRs in S2 cells do not change in gene expression levels.

(A,B) Volcano plot of differential gene expression analysis. (A,B) Gene expression analysis comparing over-expression in S2 cells of the three members of the Elav RBP family (treated as pseudo replicates) vs. the three 3X-MT variants. (A) Shown in red are genes that are significantly changed in expression ($p < 0.01$, fold change > 1). (B) Shown in red overlay are genes that show a greater than 100 bp change in weighted 3' UTR length, in either direction, when comparing over-expression of WT or 3X-MT Elav in S2 cells. Each point represents a gene. (3'-seq libraries were generated by Sonali Majumdar)

Putative Elav regulated pAs have U-rich DSEs

The gene expression analysis of APA targets suggests that Elav might more likely mediate 3' UTR lengthening by modulating PAS recognition. This would be analogous to the role of other RBPs in regulating APA (Batra et al., 2014; Licatalosi et al., 2008; Masuda et al., 2015). Furthermore, RIP analysis showed binding of Elav downstream of putative regulates PAS (Hilgers et al., 2012). To start to address this hypothesis, we calculated the relative strength of each pA, which represents the fraction of isoforms ending at a specific site compared to downstream sites. We hypothesized that PAS recognition might play a role, hence we might detect changes in associated pA properties in our data. When we looked for the PAS upstream of the cleavage site of ends that change in relative strength upon expression of Elav, we failed to detect any significant difference in the identity of the putative PAS upstream to the cleavage site (Fig 3.7A-B). However, when we looked for the presence of the 10 most represented hexamers downstream of the cleavage site, which contains U-rich sequences as shown previously (Conway and Wickens, 1985; Gil and Proudfoot, 1984; Gil and Proudfoot, 1987; Hart et al., 1985; Salisbury et al., 2006), we found a significant enrichment of U-rich hexamers downstream of the cleavage site of ends that are predicted to be recognized at lower level upon over-expression of Elav (Fig. 3.7C-D). U-rich stretches of RNA have been shown to be binding sites for Elav (Lisbin et al., 2001). Our analysis suggests that Elav might lead to expression of longer 3' UTR isoforms by binding downstream of proximal pAs with U-rich DSEs and inhibiting their recognition.

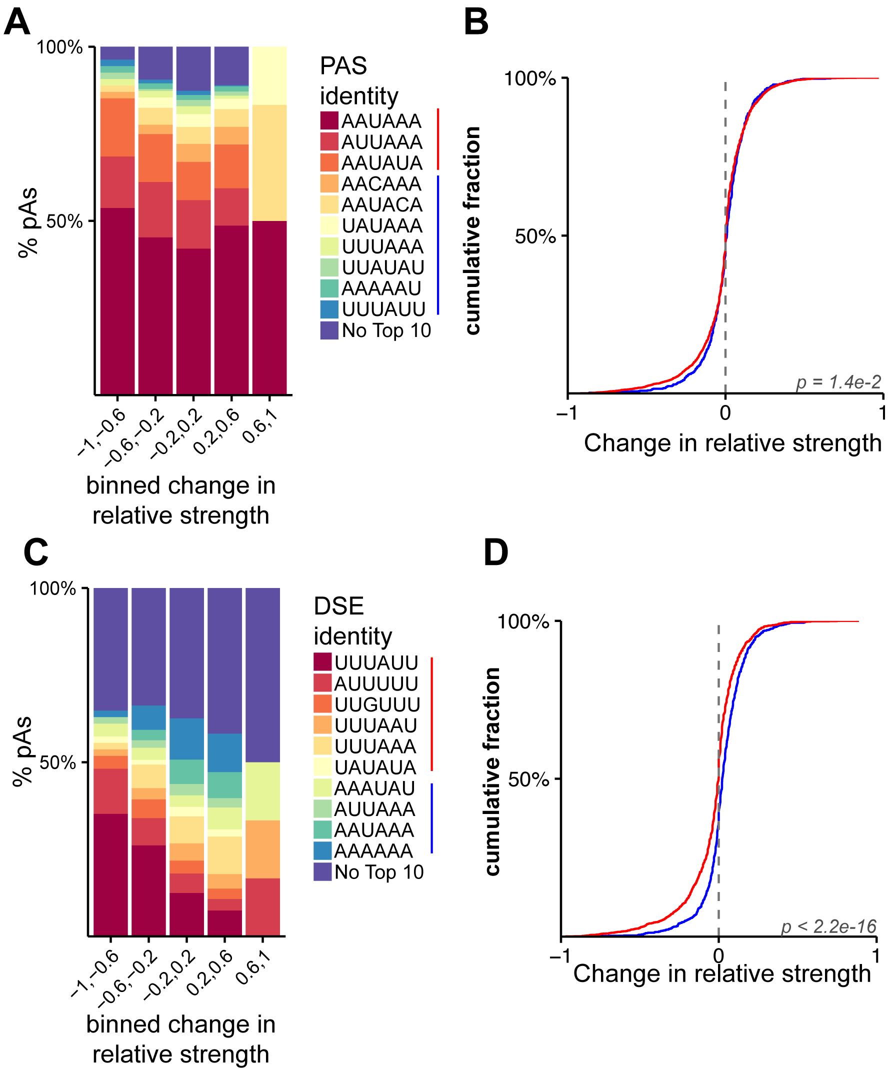


Figure 3.7 pAs that are predicted to be bypassed in the presence of Elav show U-rich DSEs.
(A,C) Percentage of pAs that present in the 50 nt window upstream of the pA site a PAS as shown in the legend (A) or a DSE (C). pAs were binned according to the difference in relative strength when comparing over-expression of WT to 3X-MT Elav. Negative values correspond to sites that decrease in relative strength upon over-expression of WT Elav and positive values show increase in relative strength. (B,D) Empirical cumulative distribution functions of the change in relative strength of pAs in terminal 3' UTRs subdivided by either PAS strength, as shown color coded next to the legend (A) or DSE motif U richness, red 3 or more Us, blue 2 or less (D). The relative strength score represents the fraction of isoforms ending at a specific site compared to downstream sites. This was calculated from the fraction of reads at each pA out of all reads at that site or downstream ones. Only pA sites of genes that showed a change in weighted 3' UTR length of greater than 100 bp in either direction were considered in the analysis. p value of Kolmogorov-Smirnov goodness of fit test (B,D).

Discussion

The expression of longer 3' UTR isoforms in the nervous system has been recently found to be a conserved aspect of gene expression in this tissue (Hilgers et al., 2011b; Miura et al., 2013; Miura et al., 2014; Smibert et al., 2012). Recent reports suggest that this phenomenon in *Drosophila* is regulated by the neural specific expression of Elav in the CNS (Hilgers et al., 2012; Oktaba et al., 2015). Despite these advancements, the mechanism and breadth of regulation of 3' UTR isoforms by Elav are not currently well defined. To better understand the role that Elav plays in the nervous system, we queried the gain and loss of function of Elav genome-wide. Our analysis shows that hundreds of gene are under the control of Elav to express longer 3' UTR isoforms, however Elav does not seem to be necessary to mediate expression of most long 3' UTR isoforms *in vivo*. This could be in part due to the presence of Elav paralogs expressed in the same tissue, which we show can also mediate extension of 3' UTRs in a gain of function setting. This could potentially explain the results seen by Hilgers *et al*., where they noticed complete loss of 3' UTR extensions *in vivo* in Elav null embryos (Hilgers et al., 2011b), which we do not recapitulate in Elav null L1 CNS. It is possible that the previous study could show complete loss of long 3' UTR isoforms in 8-12 hours embryos since Elav appears to be the dominant paralog to be expressed at that time point (Zaharieva et al., 2015). Studies of double and triple LOF mutants of the Elav paralogs will be needed to further test this redundancy hypothesis. If indeed these proteins are redundant in regulating expression of long 3' UTRs in the CNS, it will be interesting to understand what determines their specificities, as Elav is the only one which results in a lethal phenotype when lost (Zaharieva et al., 2015). Interestingly, these three paralogs appear to be differentially localized between the nucleus and cytoplasm (Zaharieva et al., 2015), and we even observed changes in Elav localization at different points in development (data not shown). Furthermore, we discovered ubiquitous expression of the previously believed to be neural restricted Elav and found this basal level of expression to be actively regulated by the microRNA pathway (Sanfilippo et al., 2016). A better understanding of these aspects of the expression of these RBPs will certainly shed light on the role that these evolutionary conserved neural RBPs play in the biology of the nervous system.

Elav is the funding member of the Hu/Elav family of RNA binding proteins (Campos et al., 1985). These RBPs have been implicated in the regulation of many aspects of RNA biology, amongst which stability (Peng et al., 1998), splicing (Toba et al., 2002) and APA (Hilgers et al., 2012). Early reports in fly had already implicated Elav in the regulation of pA recognition (Soller, 2003), finding that Elav binds at U-rich elements in introns in proximity of pAs leading to repression of recognition of the sites. However, an exact mechanism of how Elav mediates regulation of pAs or the relevance of regulation of stability in setting the landscape of 3' UTR isoform expression is still lacking. Recent reports propose that Elav is recruited at regulated loci through interaction with promoters which are characterized by stalled polymerase (Oktaba et al., 2015), however this does not address the downstream effect of Elav eventually interacting with binding sites on the nascent RNA. Here, we provide further evidence that Elav might setup expression of long 3' UTRs through binding downstream of pA sites. Regulation of stability does not appear at the moment to play a significant role, as the targets that undergo changes in 3' UTR isoform accumulation do not significantly change in overall gene expression in both a gain and loss of function situation. Further studies will be necessary to more carefully rule out differential stabilization of 3' UTR isoforms as a mechanism to extend 3' UTRs. Nascent RNA-seq datasets as well as a more careful understanding of the sites of Elav binding, through methods such as CLIP, will provide more evidence to the role that stability plays in this process. Despite evidence that Elav might modulate pA recognition, the exact mechanism is not readily apparent. Mammalian Hu proteins have been shown to mediate repression of pA recognition in part by blocking interaction of CstF-64 with RNA (Dai et al., 2012; Zhu et al., 2007), however in *Drosophila* so far Elav has not been found to compete with CstF-64 (Soller, 2003). Conversely Sxl, another Elav family RBP has been found to compete with CstF-64 causing female specific 3' UTR elongation of the gene *e(r)* (Gawande et al., 2006). We hope that our S2 cells based system might provide a good system to biochemically characterized the role of Elav family proteins in modulating 3' UTR length in the nervous system.

CHAPTER 4: Neural specificity of the RNA binding protein Elav is achieved by post-transcriptional repression in non-neural tissues

Summary

Drosophila Elav is the founding member of the conserved family of Hu RNA binding proteins (RBPs), which collectively play critical and diverse roles in post-transcriptional regulation. In particular, Elav has for >20 years served as the canonical neuronal marker, owing to the availability of specific monoclonal antibodies. Surprisingly, although Elav has a well-characterized neural cis-regulatory module, we find endogenous Elav is also ubiquitously transcribed and post-transcriptionally repressed in non-neural settings. In particular, mutant clones of multiple miRNA pathway components derepress ubiquitous Elav protein. Our re-annotation of the elav transcription unit shows that its universal 3' UTR isoform is much longer than previously believed. This longer universal 3' UTR region includes multiple conserved, high-affinity sites for the miR-279/996 family. Notably, out of several miRNA mutants tested, we find that endogenous Elav and a transgenic elav 3' UTR sensor are derepressed in mutant clones of mir-279/996. We also observe cross-repression of Elav by another RBP derepressed in non-neural miRNA pathway clones, namely Mei-P26. Finally, we demonstrate that ubiquitous Elav has regulatory capacity, since derepressed Elav can stabilize an Elav-responsive sensor. Altogether, we define unexpected post-transcriptional mechanisms that direct appropriate cell-type specific expression of a conserved neural RBP.

Introduction

microRNAs (miRNAs) are ~22 nucleotide (nt) RNAs that regulate broad target networks and play diverse biological roles (Bartel, 2009; Sun and Lai, 2013). While it is difficult to identify processes that are not regulated by miRNAs, the general activity of the miRNA pathway, and by extension bulk miRNAs, has often been considered to be important for differentiation. This concept is based on (1) the broad diversity of miRNAs expressed in specific organs or terminally differentiated cells (Lagos-Quintana et al., 2002) (2) general downregulation of miRNAs in tumors compared to normal tissues (Lu et al., 2005), (3) that certain miRNA mutants including the founding locus lin-4 reiterate early cell lineages (Chalfie et al., 1981; Lee et al., 1993) and (4) the fact that certain stem cell types including ES cells (Wang et al., 2007) and neural stem cells (Andersson et al., 2010; Kawase-Koga et al., 2010) tolerate deletion of core miRNA biogenesis factors but are unable to differentiate. Still, it is clear that miRNAs affect the behavior of stem cells and other undifferentiated cells, and are otherwise embedded in a dizzying array of biological settings (Flynt and Lai, 2008; Shenoy and Blelloch, 2014; Sun and Lai, 2013).

In this study, we report surprising observations on the role of post-transcriptional regulation in determining the spatial accumulation of *Drosophila* Elav. This was one of the first loci whose transcript and protein products were recognized to be restricted to neurons (Campos et al., 1987; Robinow et al., 1988). Antibodies against this nuclear RBP were the first reagent to label post-mitotic *Drosophila* neurons (Robinow and White, 1991), and its status as the standard neuronal marker was solidified by the development of high-quality mouse and rat monoclonal Elav antibodies more than twenty years ago (O'Neill et al., 1994). Despite reports that Elav is transiently detected in embryonic neuroblasts and glial cells (Berger et al., 2007; Lai et al., 2012), its robust and specific accumulation in post-mitotic neurons makes Elav the "go-to" marker for this terminally differentiated fate.

We unexpectedly find that endogenous Elav protein is ectopically expressed in non-neuronal mutant clones of miRNA pathway components, due to loss of post-transcriptional repression via the elav 3' UTR. Thus, this classic cell-specific differentiation marker is under spatially broad repression by the miRNA pathway. Moreover, we demonstrate the seemingly background staining detected by Elav antibodies actually reflects native accumulation in wildtype non-neural cells. Out of many miRNAs bearing conserved target sites in the elav 3' UTR, we identify a substantial role for mir-279/mir-996 in restricting Elav expression. We also provide evidence for an auxiliary repression mechanism for elav mediated by the RNA binding protein (RBP) Mei-P26. Although basal levels of ubiquitous Elav are modest, its derepression in miRNA pathway clones has functional impact on a transgenic Elav sensor. Moreover, directed misexpression of Elav outside of the nervous system, but not within the nervous system, is profoundly deleterious. Altogether, we demonstrate unexpected post-transcriptional circuitry that restricts the expression and activity of the canonical post-mitotic neural RBP Elav.

Results

Loss of the miRNA pathway derepresses Elav in non-neuronal territories

In the course of examining clonal phenotypes of miRNA pathway mutants (Smibert et al., 2011), we were surprised to observe that arbitrary imaginal disc clones expressed Elav, the canonical nuclear neuronal marker in *Drosophila*. This was notable as beyond the photoreceptors of the eye disc, few differentiated neurons are found in other larval imaginal discs. For example, while there are only a few Elav positive neurons in the wing imaginal disc, homozygous mutant clones of the core miRNA pathway factors such as drosha, pasha, and dicer-1 reliably exhibited cell-autonomous accumulation of Elav protein (Fig. 4.1A,B).

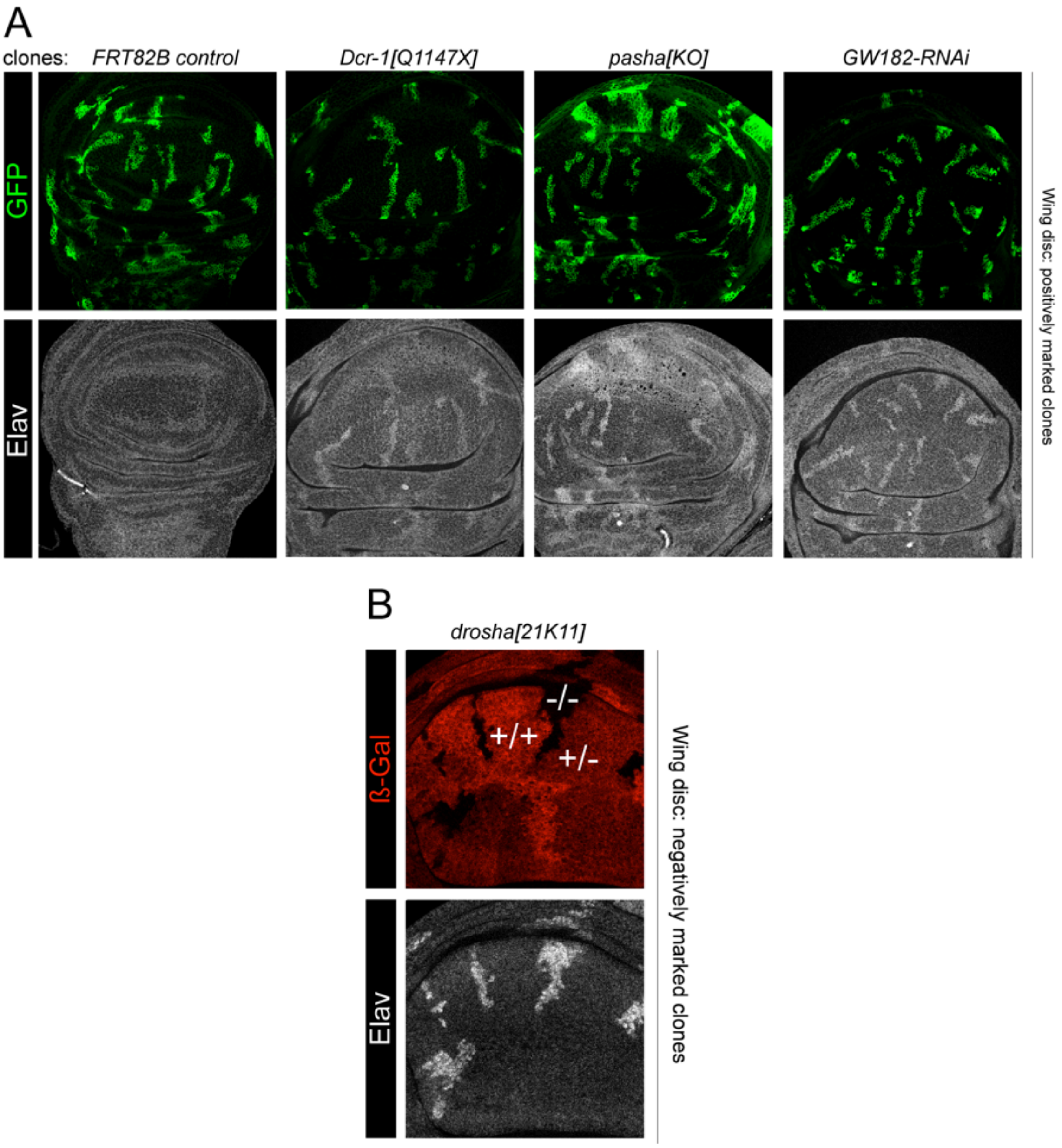


Figure 4.1 Loss of the microRNA pathway derepresses Elav in non-neural tissues in conventional mitotic clones.
Shown are regions of the wing imaginal disc pouch, costained for a clonal marker (GFP–green or ß-Gal–red) and Elav (shown in gray scale). Positively marked MARCM clones (A) or negatively marked clones (B) generated with conventional technique of LOF alleles of the microRNA pathway. Conventional clones of miRNA pathway components show de-repression of Elav. Reprinted with permission with modifications. B panel generated by Peter Smibert. (Sanfilippo et al., 2016).

Although the clonal derepression we observed in multiple mutants was unequivocal, ectopic Elav accumulated to a lower level than in differentiated neurons, and was not restricted to the nucleus as in neurons. As miRNA pathway clones are growth disadvantaged (Herranz et al., 2010), we sought to improve their recovery using the Minute technique (Blair, 2003). Curiously, not only did we obtain larger clones, these derepressed Elav protein more robustly than conventional clones (Fig. 4.2B-D). We previously showed substantial perdurance of miRNAs when depleting upstream miRNA biogenesis components (Smibert et al., 2013). Consequently, cells that are chromosomally null for miRNA biogenesis factors may retain variable amounts of miRNA functionality. We infer the extent of Elav derepression is sensitive to the loss rate of cognate mRNA/protein products and/or existing miRNAs, which may be influenced by dilution upon cell division and/or potentially distinct turnover rates of individual miRNAs.

Post-transcriptional repression of elav might be due to miRNA-mediated silencing, or alternatively reflect direct mRNA cleavage by miRNA pathway nucleases (Han et al., 2009; Karginov et al., 2010; Smibert et al., 2011). We therefore examined Flp-out clones expressing a knockdown transgene against the AGO1 co-factor GW182 (Smibert et al., 2013), which specifically impairs miRNA regulatory activity. Clones expressing GW182-RNAi similarly accumulated Elav protein (Fig. 4.2E), indicating that miRNA activity per se restricts Elav in non-neuronal territories.

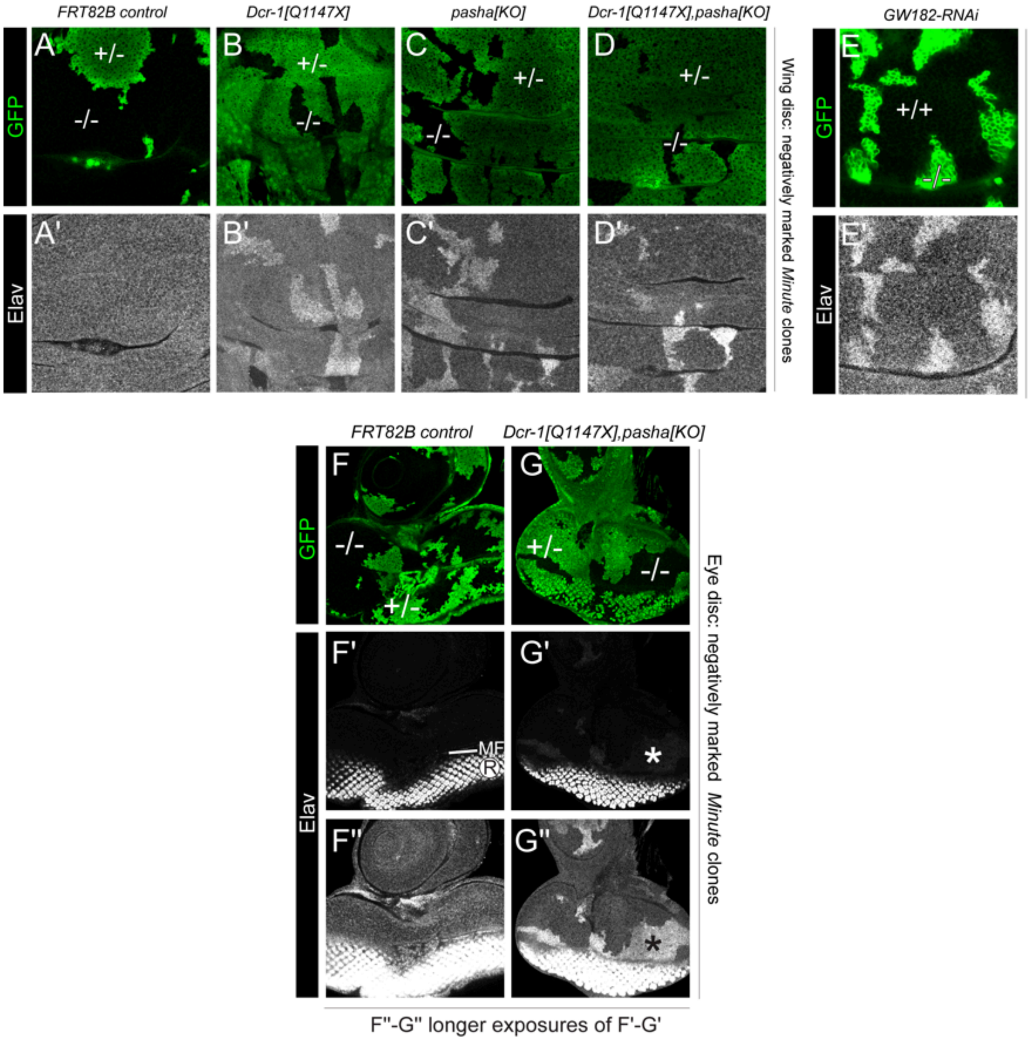


Figure 4.2 Loss of the miRNA pathway derepresses the canonical neural marker Elav in non-neural territories in Minute clones and clones of effector components of the miRNA pathway.

Shown are regions of the wing imaginal disc pouch (A-E) and are eye imaginal discs (F-G), costained for a clonal marker (GFP–green) and Elav (shown in gray scale). Negatively marked clones were generated using Minute technique in A-D,F-G. Positively marked clones using MARCM technique are shown in E. Example wildtype territories are indicated by +/+ and example clones are noted by -/-; heterozygous (+/-) tissue. Elav protein is derepressed in mutant clones lacking diverse core miRNA pathway factors (Dcr-1, pasha or drosha) (B-D, F-G), or that are depleted for the miRNA effector GW182 (E). (F-G) "Conventional" imaging technique for Elav shows typical photoreceptor (R) expression posterior to the morphogenetic furrow (MF), which is not substantially affected by miRNA pathway loss (A'-B'). Longer exposure (A''-B'') emphasizes ectopic Elav anterior to MF and in antennal region. Reprinted with permission with modifications (Sanfilippo et al., 2016).

Evidence for endogenous, ubiquitous accumulation of Elav

Since Elav is popularly considered to accumulate in post-mitotic neurons, this phenomenon might be interpreted at face value as ectopic Elav. However, as miRNAs operate post-transcriptionally, we considered whether Elav might be deployed more broadly than appreciated. In the larval eye-imaginal disc, Elav is well-known to accumulate in photoreceptor neurons located posterior to the morphogenetic furrow (MF). Eye discs bearing control clones showed the normal pattern of strong signal in differentiating photoreceptors posterior to the morphogenetic furrow (Fig. 4.2F-F'), with longer exposure showing ubiquitous signals that are typically interpreted as background staining (Fig. 4.2F''). Essentially all of the thousands of publications that utilize this standard eye marker employ imaging settings that minimize non-photoreceptor signals. Eye discs bearing Dcr-1, pasha clones showed substantial derepression of Elav anterior to the MF and in the antennal disc. This was evident even in "typical" exposure (Fig. 4.2G') but became quite obvious by increasing the gain (Fig. 4.2G'').

We used an elav-RNAi transgene to assess whether non-neuronal Elav signals were genuine. Indeed, positively-marked Flp-out Gal4 clones expressing elav-RNAi exhibited cell-autonomous depletion of Elav, in both photoreceptors and non-neuronal territories of the eye-antennal disc (Fig. 4.3A). Similarly, we observed that wing disc clones expressing elav-RNAi eliminated endogenous Elav staining (Fig. 4.3B). These data suggest the miRNA pathway clones do not reveal spatially ectopic Elav, but rather cause derepression of an unappreciated basal level of Elav present in most cells.

We confirmed these results by staining mitotic clones of the characterized null allele elav[4]. We used a negatively-marked clone strategy in females (as elav resides on the X chromosome), in which homozygous mutant cells, their wildtype twinspots, and the heterozygous unrecombined tissue can all be distinguished. Remarkably, we observed distinct levels of Elav protein in territories bearing two, one or no copies of the elav locus (Fig. 4.3C). These tests firmly establish ubiquitous, non-neuronal, accumulation of endogenous Elav protein in imaginal discs.

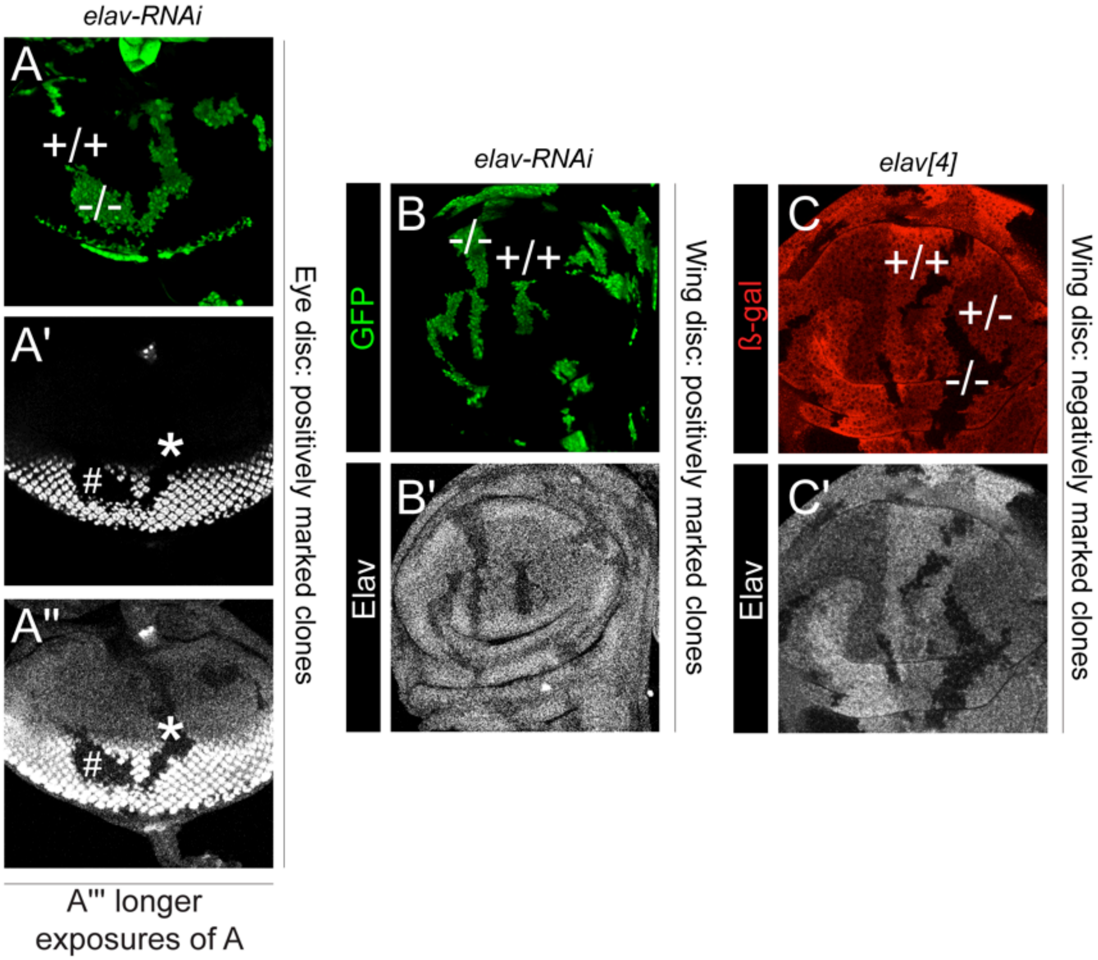

elav-RNAi
A
+/+
-/-
A'
*
#
A''
*
#
A''' longer exposures of A
Eye disc: positively marked clones
elav-RNAi
GFP
B
-/-
+/+
Elav
B'
Wing disc: positively marked clones
elav[4]
ß-gal
C
+/+
+/-
-/-
Elav
C'
Wing disc: negatively marked clones

Figure 4.3 Elav is expressed outside of the nervous system.
Shown are eye imaginal discs (C) and pouch regions of wing imaginal discs (B-C) stained for clonal markers (GFP–green or ß-Gal–red), Elav (in gray scale). (A) Clonal expression of *elav-RNAi* eliminates Elav in R cells (C'), and longer exposure also shows loss of basal Elav in non-neural disc regions (C''). (B) Clonal expression of *elav-RNAi* eliminates basal Elav in the wing disc. (C) Mitotic clonal analysis of null allele *elav[4]* shows graded expression of basal, non-neural Elav in heterozygous and mutant regions. Reprinted with permission with modifications (Sanfilippo et al., 2016).

With this revised perspective in mind, we searched for evidence of non-neural expression of elav using modENCODE mRNA-seq (Graveley et al., 2011) and total RNA-seq data(Brown et al., 2014). Surprisingly, we observe that elav is maternally deposited (in 0-2 hour embryos), clearly elevated at the onset of zygotic expression (in 2-4 hour embryos, Fig. 4.4A) and continues to be upregulated prior to neurogenesis (Fig. 4.4A) which occurs at 9-10 hours (Hartenstein, 1993). This provides definitive evidence for non-neuronal, even non-neuroblast (prior to 4 hours), transcription of elav in the early embryo.

We also detected elav transcripts in all *Drosophila* cell lines profiled (Cherbas et al., 2011), most of which lack any documented neural character (Fig. 4.4A). Furthermore, we detect Elav protein by Western blotting in the commonly used S2 cell line (Fig. 4.4B), which has hemocyte character. To address whether the miRNA pathway restricts Elav in this setting, we treated S2 cells with dsRNA against GFP, Dcr-2 or Dcr-1. We observed specific upregulation of Elav protein in cells depleted of Dcr-1 (Fig. 4.4B).

In summary, even though elav bears a neural cis-regulatory enhancer (Yao and White, 1994) and has had two decades usage as the canonical neural antigen in *Drosophila*, elav is also transcribed in non-neuronal cells of diverse developmental stages and cell types. Moreover, Elav is detectably translated in non-neuronal cells and tissues, but its basal protein accumulation is restricted by the miRNA pathway.

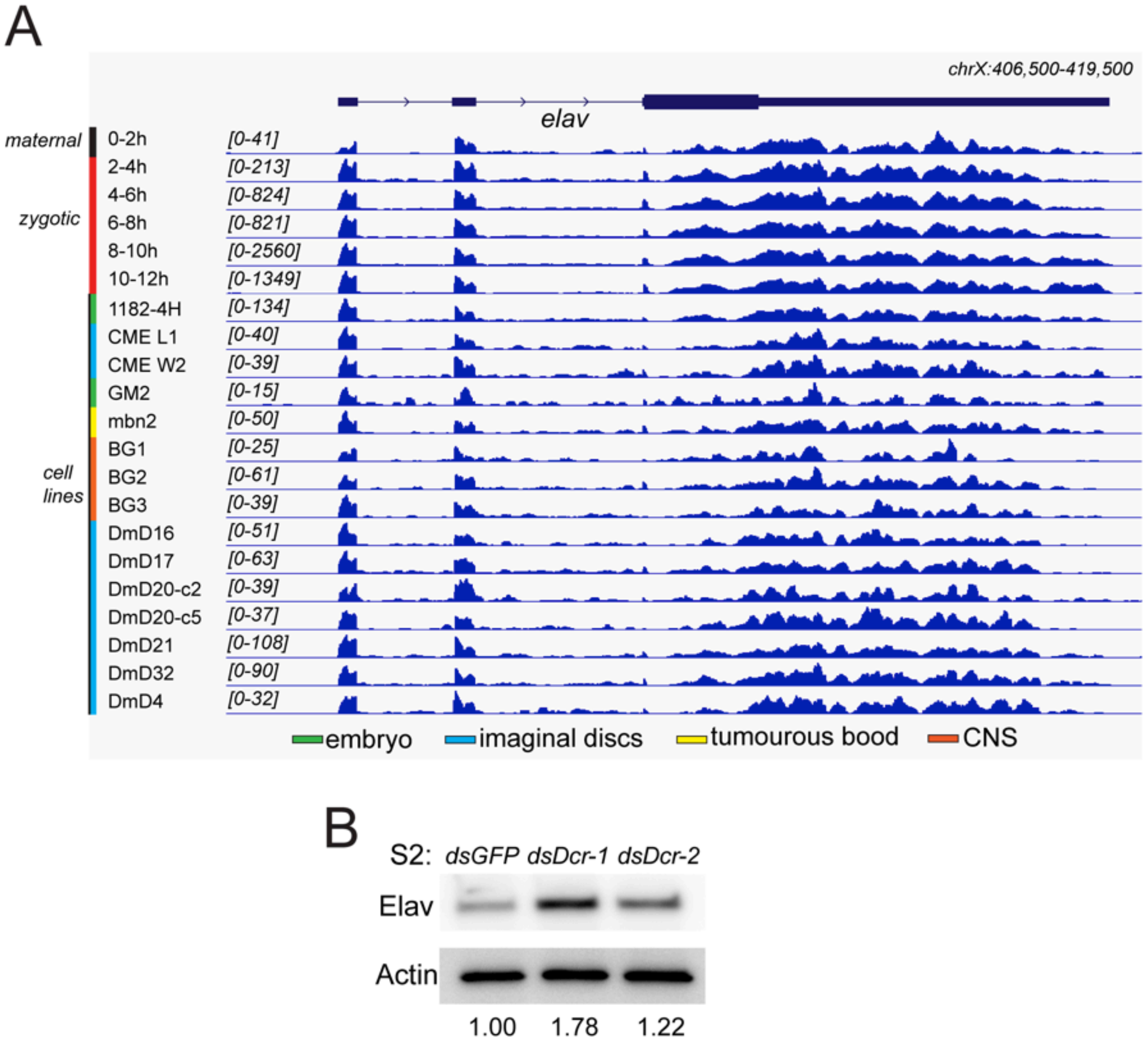


Figure 4.4 ***elav*** **is ubiquitously transcribed and is translated and regulated in S2 cells.**

(A) RNA-seq of the elav locus in a panel of embryonic stages and cell lines from the modENCODE dataset. Although Elav is considered as the canonical marker of post-mitotic neurons, elav transcripts exhibit broad temporal and spatial accumulation throughout all embryonic stages and in all cell lines. *elav* transcripts are maternally deposited (0-2h embryos) and the locus initiates zygotic transcription at the onset of maternal-zygotic transition (as indicated by increased transcript levels in 2-4h embryos). Levels of *elav* transcripts continue to increase long before the first appearance of differentiated Elav+ neurons (9-10h). (B) Western blot of S2 cells shows Elav is specifically derepressed upon *Dcr-1* knockdown. Reprinted with permission with modifications (Sanfilippo et al., 2016).

miRNA pathway loss does not result in neural transformation or transcriptional activation of elav

Although the miRNA pathway mutant data support the scenario that one or more miRNAs repress elav, one could still hypothesize indirect mechanisms leading to elav derepression. For example, loss of the miRNA pathway might reveal a default neural program, which was reported as the ground differentiation state of the vertebrate ectoderm (Ozair et al., 2013). Curiously, Mei-P26 is another factor with specific (although not exclusive) neural expression that is derepressed in non-neuronal miRNA pathway mutant clones (Herranz et al., 2010).

We addressed this by staining miRNA pathway clones for a panel of neural markers. For example, the transcription factors Scratch and Deadpan are long-established neural markers in CNS and PNS (Emery and Bier, 1995). We detected their endogenous expression in wing disc (Fig. 4.5A) and eye disc (Fig. 4.5B), but did not observe elevated Scratch and Deadpan proteins in miRNA pathway mutant clones, as observed with parallel Elav stainings. We assayed other neural antigens (Futsch, Shep, FasII, FasIII, and Orb), and none of these exhibited ectopic staining in Dcr-1 and/or pasha clones (Figure 4.5B). The specific effects on Elav argue against a general neural transformation in the loss of miRNA activity.

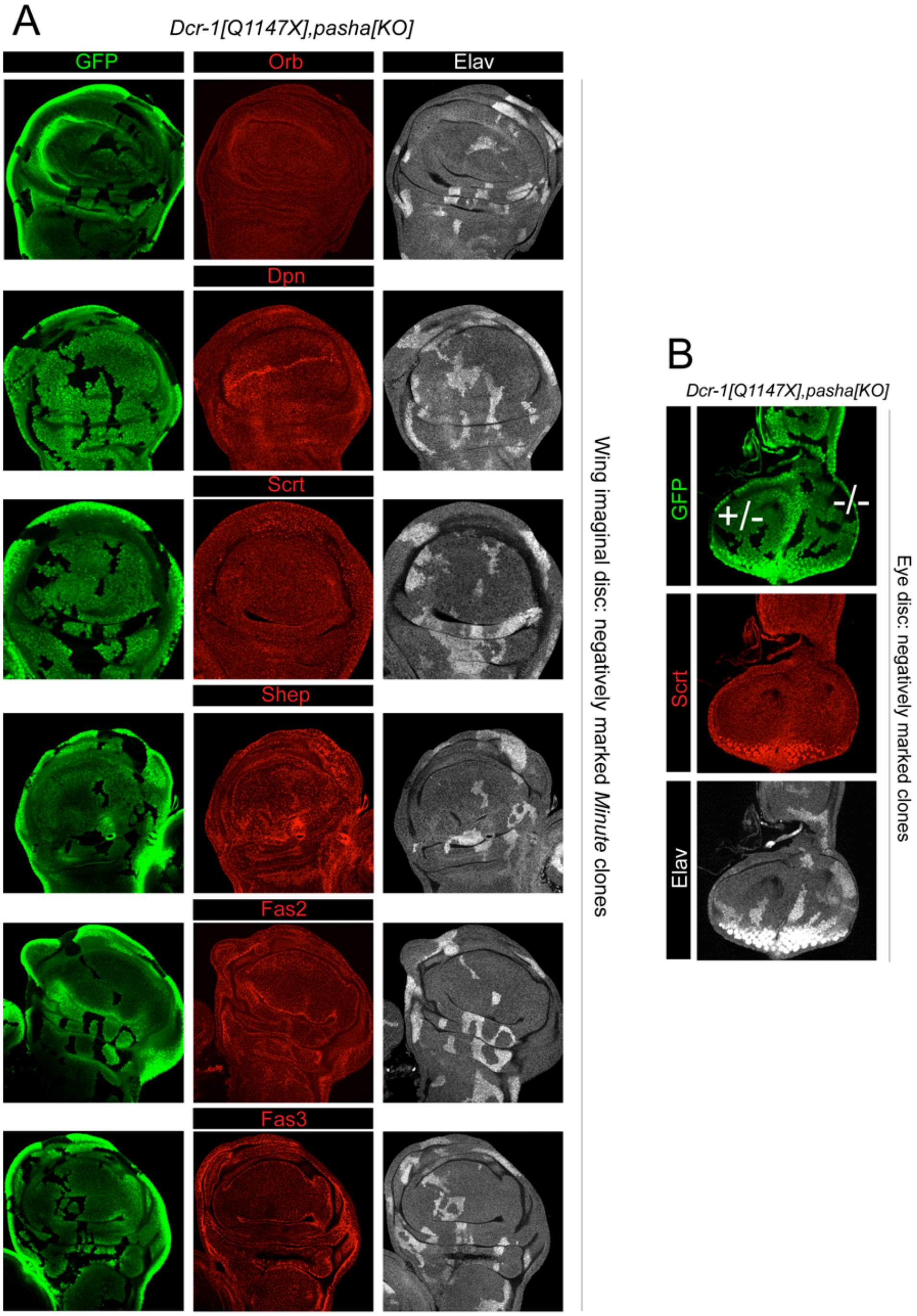


Figure 4.5 Loss of the miRNA pathway does not cause neuronal fate conversion.

Negatively marked Minute clones of double Dcr-1, pasha clones in the pouch region of the wing imaginal disc (A) or eye disc (B) show robust Elav (gray scale) derepression while no effect is detectable for a panel of neural expressed antigens (red). Reprinted with permission with modifications (Sanfilippo et al., 2016).

miRNA-mediated repression via the elav 3' UTR can generate its spatial pattern

The above negative data implied that the miRNA pathway directly represses elav. To test this, we generated a transgenic tub-GFP-elav 3' UTR (GFP-elav) sensor (Fig. 4.6A). We note that extensive transcriptome data (Brown et al., 2014) does not support the annotated elav 3' UTR utilized in TargetScan predictions (www.targetscan.org) as a bona fide expressed isoform. Instead, we identified a longer proximal isoform as well as genuinely extended isoforms (Fig. 4.6A), consistent with previous Northern blotting experiments (Smibert et al., 2012). Our GFP-elav sensor includes the full untranslated region. Interestingly, while control tub-GFP-tub 3' UTR sensor lacked patterned expression (Fig. 4.6B), GFP-elav precisely recapitulated the endogenous Elav pattern (Fig. 4.6C-E). That is, GFP was low throughout imaginal discs, but accumulated in Elav+ photoreceptors, and neurons in the brain and ventral nerve cord. This was particularly striking in the case of specific cells that express high levels of GFP-elav in the vicinity of early arising sensory organs of leg discs. Double labeling with Elav confirmed these were indeed neurons (Fig. 4.6D). Thus, regulation via the elav 3' UTR is sufficient to generate the Elav spatial expression pattern.

We then introduced GFP-elav sensor into backgrounds suitable for generating negatively-marked miRNA pathway clones. Cells lacking pasha (Fig. 4.6F-G) or drosha (not shown) exhibit concomitant, cell-autonomous, derepression of both endogenous Elav and GFP-elav sensor. Thus, a major aspect of elav spatial control is mediated post-transcriptionally.

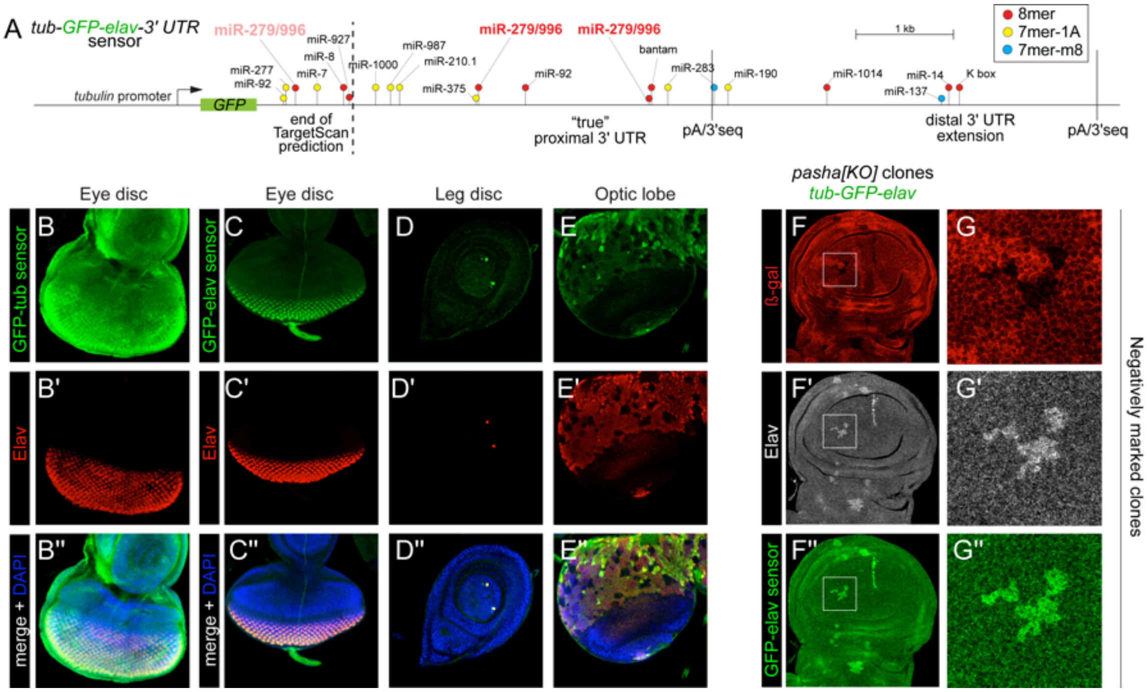


Figure 4.6 Direct repression of the *elav* 3' UTR via the miRNA pathway.
(A) Schematic of *tub-GFP-elav 3' UTR* sensor transgene. We find that the 3' terminus used in public miRNA annotations (e.g. TargetScan) is not detected *in vivo*. Rather, the genuine proximal 3' UTR isoform is nearly 3 kb longer, and we identify an extended 3' UTR isoform, both of which are supported by 3'-seq tags (as annotated). Images depict eye imaginal discs (B-C), leg imaginal disc (D), brain optic lobe (E) and pouch regions of wing imaginal discs (F-G) stained for reporter GFP (B-E, F-G), clonal markers (GFP–green or ß-Gal–red) and Elav (in gray scale). (B-E) Comparison of a *tub-GFP-tubulin-3' UTR* and *tub-GFP-elav-3' UTR* sensor transgenes. (B) The *tub* 3' UTR sensor is broadly expressed in the eye disc, as well as other tissues (not shown). (C-E) The *elav* 3' UTR restricts ubiquitously expressed GFP into the pattern of endogenous Elav protein, as seen by their co-localization in several tissues. (F-G) Mitotic *pasha[KO]* clones (marked by absence of ß-Gal staining) show coincident derepression of Elav and the *elav* 3' UTR sensor. Reprinted with permission with modifications (Sanfilippo et al., 2016).

The mir-279/mir-996 cluster is an endogenous repressor of elav

As with many genes, the elav 3' UTR contains conserved binding sites for multiple miRNAs (Fig. 4.6A). Amongst these, miR-7 and miR-8 are known to be active in imaginal discs. However, null clones of mir-7 and mir-8 did not derepress Elav (Fig. 4.7A-C). In general, few miRNA targets have been shown to exhibit cell-autonomous derepression in miRNA mutant clones, so this is a very stringent test. These negative results do not distinguish if potential regulation is non-existent or very mild, or requires coordinated action of multiple miRNAs.

Cognizant of the longer elav 3' UTR, we performed de novo assessment of potential conserved miRNA sites in elav 3' extended regions. Notably, the "true" longer proximal 3' UTR contains two deeply conserved 8mer sites for the miR-279/996/286 seed family (Fig. 4.6A). The 8mer site constitutes the highest affinity type of canonical site (Grimson et al., 2007), and both sites reside in locally conserved domains in the newly recognized common 3' UTR. We also identified a less-conserved 7mer-1A site for the miR-279/996/286 family. Expression of miR-286 is restricted to the early embryo, whereas miR-279/miR-996 are co-expressed from a genomic cluster and are detected throughout development (Mohammed et al., 2014; Sun et al., 2015). We recently characterized a double deletion allele [15C] that specifically removes mir-279/mir-996 (Sun et al., 2015), and showed both miRNAs contribute to a variety of neural phenotypes previously ascribed to sole loss of miR-279 (Cayirlioglu et al., 2008; Luo and Sehgal, 2012).

We examined mir-279/996[15C] null clones and observed clearly elevated Elav in mutant cells (Fig. 4.7D), in contrast to the other miRNAs tested (Fig. 4.7 B-C). Therefore, this individual miRNA locus is a substantial mediator of non-neural repression of Elav. With this knowledge in hand, we assessed for direct regulation of the tub-GFP-elav 3' UTR sensor in mir-279/996[15C] clones. Indeed, accumulation of this sensor was elevated in cells deleted for miR-279/996 (Fig. 4.3E), demonstrating them as critical effectors of the miRNA pathway for post-transcriptional suppression of ubiquitous elav.

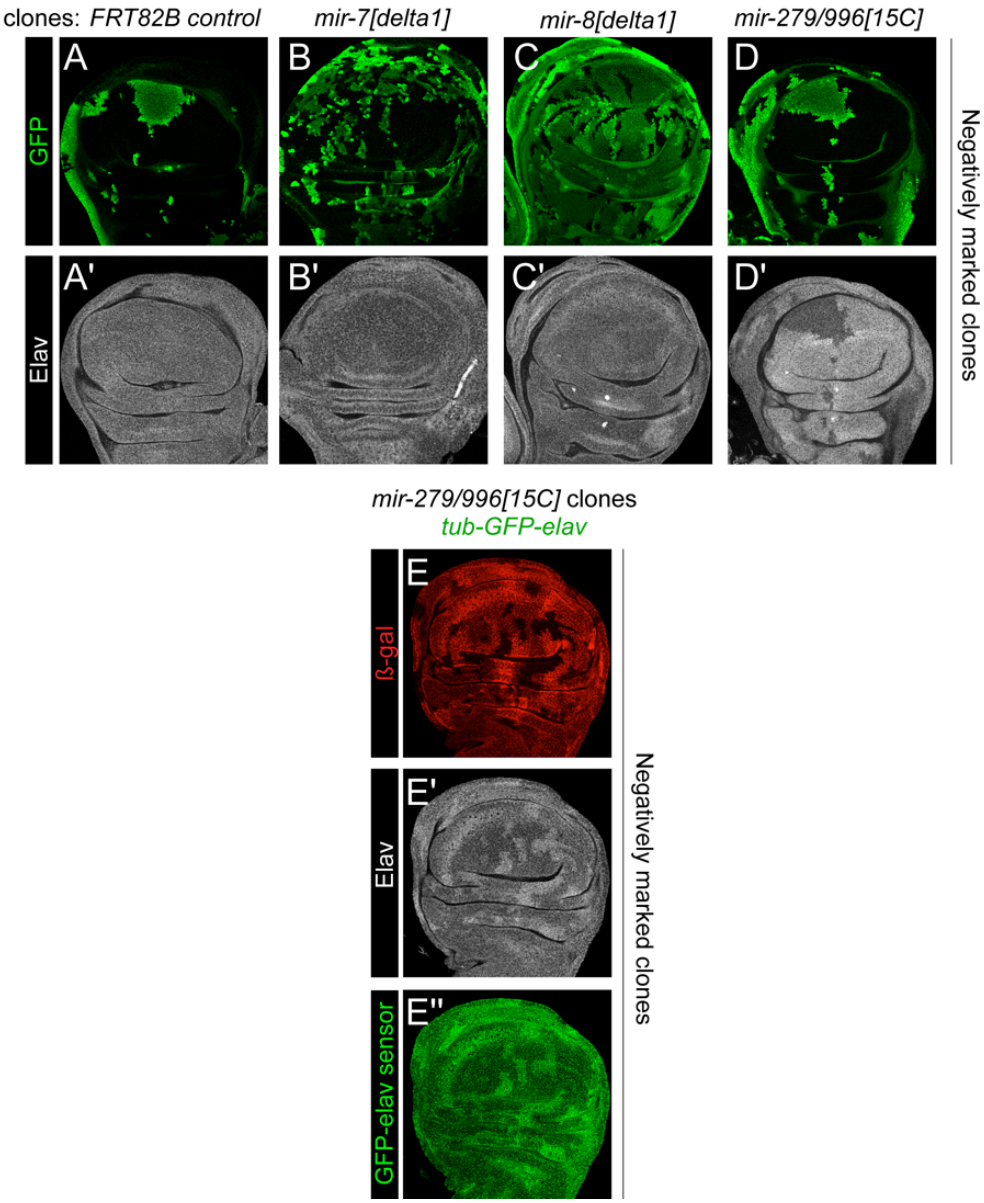


Figure 4.7 Direct repression of the *elav* 3' UTR is mediated by miR-279/996. Images depict pouch regions of wing imaginal discs (A-E) stained for reporter GFP (E), clonal markers (GFP–green or ß-Gal–red) and Elav (in gray scale). (A-D) Tests of individual miRNA loci on post-transcriptional repression of Elav. (A) Control clones and (B) *mir-7* or *mir-8* (C) null clones do not affect Elav, while clones of *mir-279/996* derepress Elav protein (D). (E) Clonal deletion of *mir-279/996* causes concomitant elevation of Elav and the *elav* 3' UTR sensor. Reprinted with permission with modifications (Sanfilippo et al., 2016).

Cross-regulatory interactions of the RBPs Mei-P26 and Elav

Mei-P26 is also broadly depressed in miRNA pathway disc clones (Herranz et al., 2010), and Mei-P26 was proposed to inhibit the miRNA pathway (Neumüller et al., 2008). According to these prior data, ectopic Mei-P26 might be hypothesized to mimic miRNA pathway mutant clones, leading to derepression of Elav. Unexpectedly, we observed the opposite effect, in that Flp-out Gal4 clones expressing Mei-P26 showed strong loss of basal Elav. This was true in both an "Elav-hi" domain such as the photoreceptor field (Fig. 4.8A), and "Elav-lo" domains such as the wing pouch (Fig. 4.8B). Therefore, Mei-P26 appears to repress Elav, and this is not apparently due to an effect on the miRNA pathway in general.

Given that Mei-P26 was shown to bind RNA via its NHL domain (Loedige et al., 2014), we tested if Mei-P26 might directly regulate Elav via its 3' UTR. Indeed, clonal misexpression of Mei-P26 resulted in cell-autonomous reduction of the GFP-elav 3' UTR sensor (Fig. 4.8C). This effect was milder than observed with endogenous Elav protein, but close examination of well-positioned clones clearly showed that the GFP-elav sensor was lower in clones that misexpress Mei-P26 (Fig. 4.8D). Thus, Mei-P26 may contribute to post-transcriptional repression of Elav outside of the nervous system. Consistent with this, clonal knockdown of Mei-P26 results in mild upregulation of endogenous Elav (Fig. 4.8E).

Since ectopic Mei-P26 acts oppositely to the miRNA pathway with respect to basal Elav, and both RBPs are derepressed in miRNA pathway clones, we assessed the consequences of simultaneously removing Mei-P26 and miRNAs. To do so, we compared control pasha-KO MARCM clones with ones that expressed UAS-mei-P26-RNAi. The former exhibited cell-autonomous derepression of Mei-P26 protein (Fig. 4.8F) while the latter did not (Fig. 4.8G), confirming efficacy of the mei-P26-RNAi transgene. The accumulation of Elav protein was sometimes higher in pasha-KO+UAS-mei-P26-RNAi compared to pasha-KO clones, but they were not reliably different.

These data suggest that activity of the miRNA pathway predominates over Mei-P26 with respect to repression of basal Elav. Nevertheless, the existence of multiple strategies for post-transcriptional repression of Elav supports the notion that it is a biologically significant imperative to restrict Elav outside of neurons.

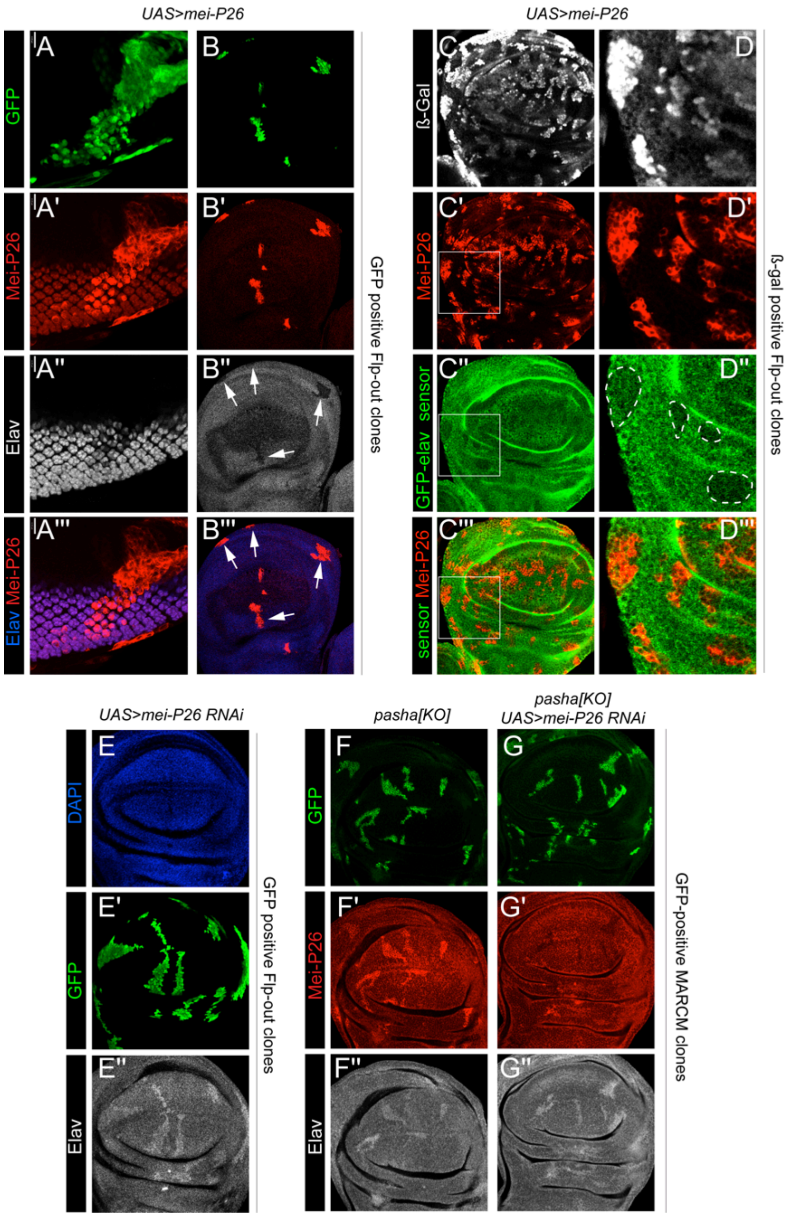


Figure 4.8 Mei-P26 represses *elav* via its 3' UTR.
Shown are eye imaginal disc (A) and pouch regions of wing imaginal discs (B-G). (A-B) Clonal activation of Mei-P26 leads to cell-autonomous decrease in Elav protein in both a high expression domain (photoreceptors, A) and a low expression domain (wing pouch, B, arrows). (C) Ectopic Mei-P26 represses the *tub-GFP-elav-3' UTR* sensor. Although the effect is quantitatively mild, it is more clearly observed in higher magnification clones (D, dotted circle regions). (E) Clonal knockdown of *mei-P26* causes cell-autonomous increase in basal Elav. (F-G) MARCM analysis of *pasha[KO]* clones. (F) Control *pasha* clones derepress both Mei-P26 and Elav proteins. (G) Knockdown of *mei-P26* reverses the accumulation of Mei-P26 protein in *pasha* clones, but does not reliably super-activate Elav. Reprinted with permission (Sanfilippo et al., 2016).

Regulatory and phenotypic impact of non-neural Elav

Elav is mostly considered to influence neuronal gene expression. Do the lower levels of ubiquitous Elav have detectable regulatory impact? To address this, we took advantage of a transgenic Elav activity sensor (ub-GFP-hsp70Ab 3' UTR or "UgGH", Fig. 4.9A) to assess in vivo function of Elav (Toba et al., 2002). We confirmed that Flp-out clones expressing Elav can upregulate the UgGH reporter (Fig. 4.9B). We then introduced UgGH into a background bearing either Dcr-1 or Pasha null mutant clones. Both types of mutant clones elevated both Elav and Elav sensor (Fig. 4.9C-D), demonstrating palpable regulatory activity of derepressed, basal Elav outside of the nervous system.

In the course of Elav activity sensor tests (Fig. 4.9B), we noticed that clones of cells that overexpress Elav were much smaller than control clones, or even miRNA pathway clones. This suggested that elevation of Elav in non-neural settings might not be tolerated. To further investigate this, we activated Elav with a panel of Gal4 drivers. We found that activation of Elav using da-gal4 (ubiquitous), ap-gal4 (dorsal compartment of wing disc), and rn-gal4 (wing pouch) were all fully lethal (Fig. 4.9E). Therefore, elevation of Elav in non-neural settings is highly deleterious. By contrast, misexpression of Elav in neurons using elav-Gal4 was compatible with viability, consistent with its normally high levels in this cell type.

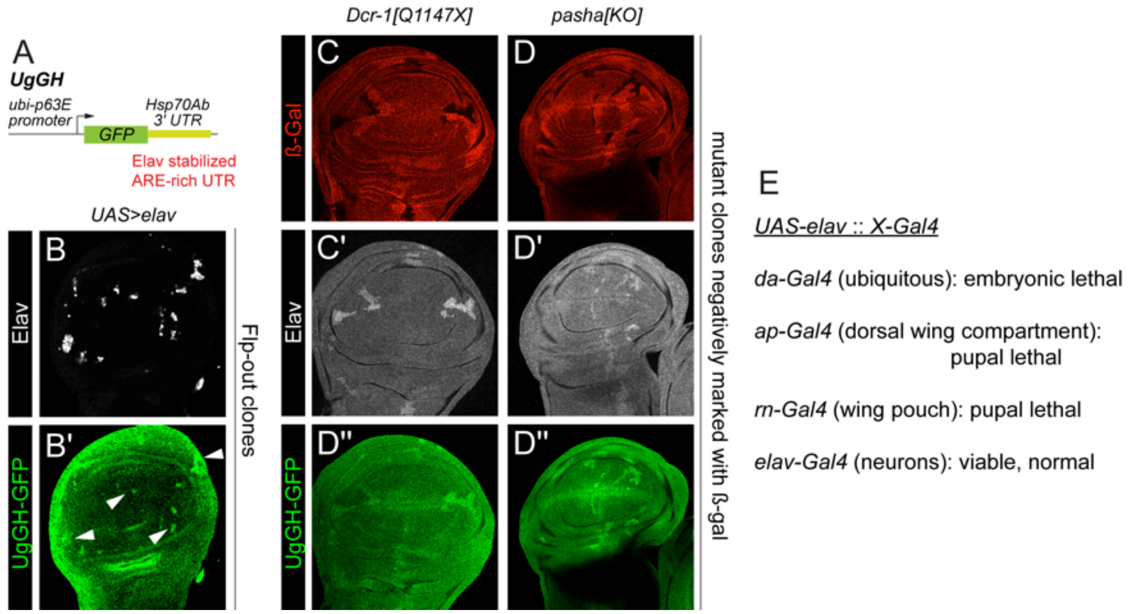

A
UgGH
ubi-p63E promoter
GFP
Hsp70Ab 3' UTR
Elav stabilized ARE-rich UTR
UAS>elav
B
Elav
B'
UgGH-GFP
Flp-out clones
Dcr-1[Q1147X]
pasha[KO]
C
D
ß-Gal
C'
D'
Elav
D''
D''
UgGH-GFP
mutant clones negatively marked with ß-gal
E
UAS-elav :: X-Gal4
da-Gal4 (ubiquitous): embryonic lethal
ap-Gal4 (dorsal wing compartment): pupal lethal
rn-Gal4 (wing pouch): pupal lethal
elav-Gal4 (neurons): viable, normal

Figure 4.9 Functional consequences of elevating Elav in wing imaginal discs.
(A) The UGgH Elav sensor consists of ubiquitously-expressed GFP followed by the ARE-rich 3' UTR of *Hsp70Ab*, a sensor previously shown to be stabilized by Elav. (B) Clonal expression of *UAS-elav* increases UGgH expression (arrowheads). (C-D) *Dcr-1* (C) or *pasha* (D) mutant clones that derepress Elav also stabilize the UGgH Elav activity sensor. (E) Summary of *elav* misexpression tests shows it induces lethality when activated with multiple non-neuronal drivers, but not when activated neuronally. Reprinted with permission with modifications (Sanfilippo et al., 2016).

Immunostaining of Flp-out expression clones provided cellular insight into region-specific effects of Elav. Control GFP-expressing clones were easily induced throughout the eye disc (Fig. 4.10A). However, inspection of eye discs bearing Elav Flp-out clones showed that labeled cells persisted robustly only within the normal "Elav-hi" domain, namely in the photoreceptor field (Fig. 4.10B). Elav-expressing clones were poorly recovered elsewhere in undifferentiated portions of the retina or antennal domain. This is consistent with the viability of elav-Gal4>UAS-elav animals, and the overall notion that Elav exerts its normal function in neurons, but is poorly tolerated at high levels elsewhere.

In the wing disc, control GFP clones were again recovered throughout (Fig. 4.10C). In contrast, Elav-expressing clones exhibited distinct behaviors in different wing disc regions. They were poorly recovered in the prospective notum epithelium (Fig. 4.9D, N), although the large adepithelial cells that cover the notum could express ectopic Elav (Fig. 4.9D, AE). We also observed clones in the wing pouch disc proper (Fig. 4.9D, WP), but these exhibited poor morphology suggestive of apoptosis. Indeed, pouch clones that overexpressed Elav reacted strongly with the apoptotic marker cleaved caspase-3 (c-casp3), while adepithelial cells with high Elav did not react similarly (Fig. 4.9D"). Close examination of wing pouch clones showed fragmented, pyknotic nuclei that were in the process of being removed from the disc epithelium (Fig. 4.10E). This was more evident in cross sections through the wing pouch, which showed that medially located Elav+ cells accumulated high levels of c-casp3 and delaminated

(Fig. 4.10F, asterisk), whereas lateral Elav+ cells in the wing hinge area did not robustly activate c-casp3 and largely remained in the epithelium (Fig. 4.10F, pound signs). Overall, these results highlight a biological imperative to restrict the accumulation of Elav low outside of the nervous system, particularly within distinct disc compartments.

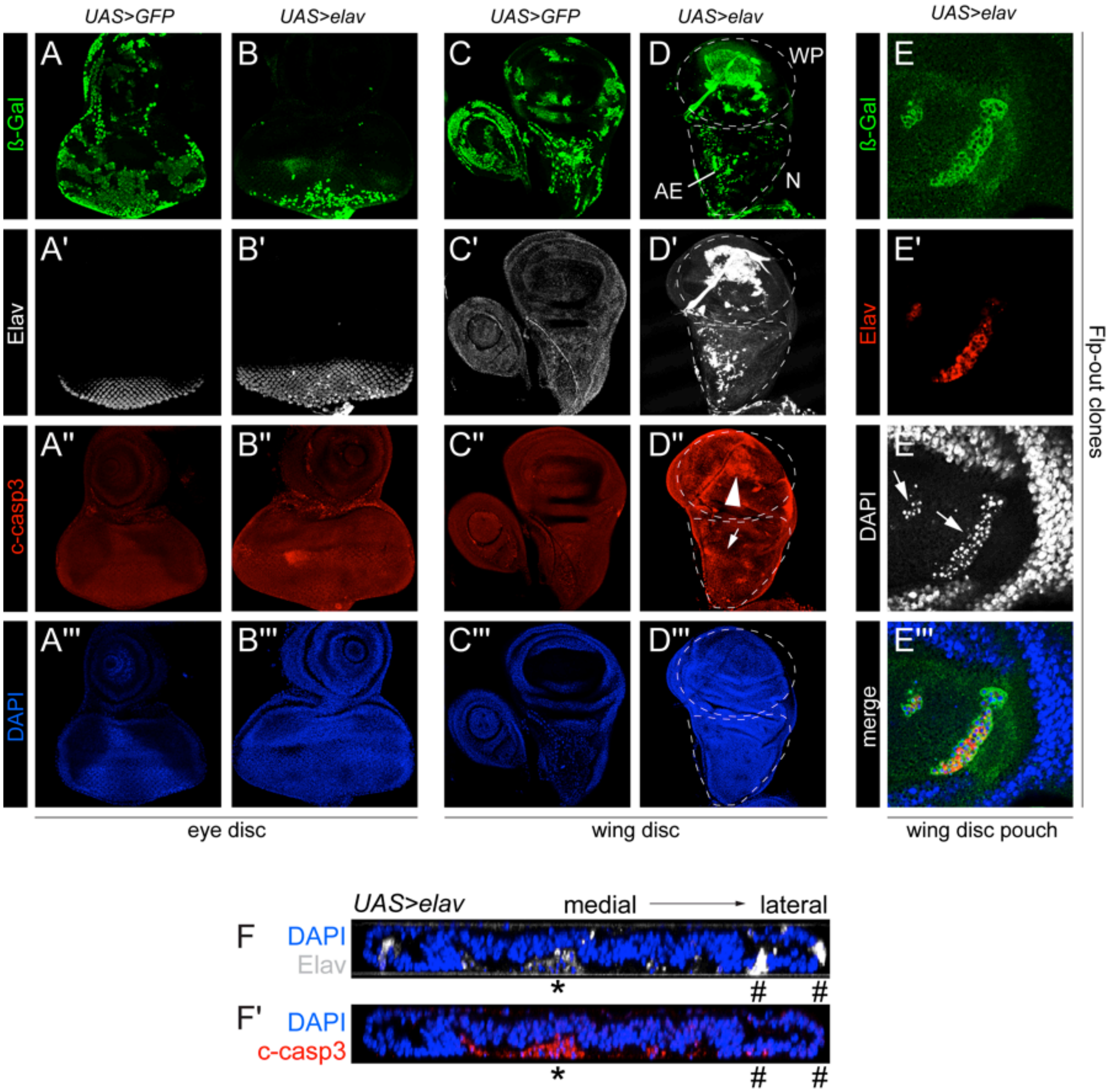

UAS>GFP
UAS>elav
UAS>GFP
UAS>elav
UAS>elav
ß-Gal
Elav
c-casp3
DAPI
A
B
C
D
WP
AE
N
E
A'
B'
C'
D'
E'
A''
B''
C''
D''
E''
A'''
B'''
C'''
D'''
E'''
ß-Gal
Elav
DAPI
merge
Flp-out clones
eye disc
wing disc
wing disc pouch
UAS>elav
medial
lateral
F
DAPI
Elav
F'
DAPI
c-casp3
*
#
#
*
#
#

Figure 4.10 Detrimental consequences of elevating Elav in wing imaginal discs.

(A,B) Flp-out expression clones in eye discs, marked by activation of *UAS-lacZ* (ß-Gal, green). (A) Control GFP clones are recovered throughout the eye disc. (B) Elav expressing clones are preferentially recovered in photoreceptors, and do not induce cell death as marked by cleaved caspase-3 (B"). (C-F) Flp-out expression clones in wing discs, marked by activation of *UAS-lacZ* (ß-Gal, green). (C) Control GFP clones are recovered throughout the wing disc and do not induce c-casp3 reactivity. (D) Elav-expressing clones are recovered in the wing pouch (WP) and prospective notum (N), but cells in the latter region are not in the disc epithelium but rather reside in the adepithelial layer (AE). Elav-expressing cells in the wing pouch accumulate high levels of c-casp3 (I", arrowhead), while Elav-expressing adepithelial cells do not (D", arrow). (E) Close-up of wing pouch region shows that Elav-expressing clones are fragmented. Lack of continuous DAPI signal is due to visualization of a narrow z-section, and the pyknotic nuclei delaminate from the epithelium. (F) Cross-section through the wing pouch illustrates how dying, Elav+/c-casp-3+ clones in the center of the wing pouch are removed from the epithelium, while laterally located clones remain integrated and express little or no c-casp3+ (pound signs). Reprinted with permission (Sanfilippo et al., 2016).

Discussion

Unexpected expression of cell-specific or compartment-specific markers

Qualitative techniques for assessing mRNA and protein accumulation in tissues can sometimes provide unsettlingly distinct impressions from quantitative techniques. A classic example is that whole mount in situ hybridizations of *Drosophila* egg chambers and oocytes provide striking visual evidence for highly localized transcripts of critical anterior-posterior patterning determinants. Nevertheless, quantitative analysis reveals such signals reflect a small minority of total cellular transcripts. For example, the sharp posterior localization of nanos and oskar in situ signals represent only 4% and 18% of total oocyte transcripts, respectively (Bergsten and Gavis, 1999). Thus, while it originally appeared that mRNA localization determines protein localization, later observations demonstrated that translational control is critical for appropriate spatial restriction of cognate proteins.

As another example, antibodies to the Notch transcription factor Su(H) have long served to mark socket cells of peripheral sense organs (Gho et al., 1996), and its characteristic expression there is driven by an autoregulatory socket enhancer (Barolo et al., 2000). Nevertheless, as most cells can execute Notch signaling, as evidenced by profound cell-autonomous effects of activated Notch, it is implicit that Su(H) must be ubiquitously expressed. In these as in all staining experiments, the investigator chooses when to stop a colorimetric reaction or how much to expose a fluorescent image. However, when utilizing cell-specific or subcellular-specific markers, one typically tries to minimize apparent background signals.

The case we present for Elav is particularly surprising, given its broad usage as a post-mitotic, neural-specific *Drosophila* antigen. In fact, Elav was reported to accumulate transiently in embryonic glia (Berger et al., 2007), and that it can be detected in a small fraction (~10%) of larval neuroblasts (Lai et al., 2012). Nevertheless, these findings have not detracted from its broad use and reliable utility to mark mature neurons. Here, we show that Elav protein is modestly but ubiquitously expressed, and substantially derepressed in miRNA pathway mutant clones. We acknowledge the endogenous regulatory impact of basal Elav remains to be demonstrated. For example, tissue-specific knockdown of Elav in the wing pouch did not overtly affect wing development (data not shown). Nevertheless, Elav is a powerful and multifaceted post-transcriptional regulator that orchestrates alternative splicing, 3' end formation and alternative polyadenylation (Hilgers et al., 2011b; Koushika et al., 2000; Lisbin et al., 2001; Soller, 2003), and it would not be surprising for basal Elav to have molecularly demonstrable effects. Indeed, we visualized that ectopic Elav generated in miRNA pathway mutant clones upregulates a transgenic Elav sensor.

Elav was previously shown to be transcriptionally repressed in neuroblasts by the intrinsic factor Worniu, and that this is required to prevent their premature neural differentiation (Lai et al., 2012). Our data indicate another unanticipated tier of

Elav repression acting via miRNAs, and further suggest that the RBP Mei-P26 also contributes to post-transcriptional repression of the elav 3' UTR. It is conceivable that other genetic situations might activate Elav in unexpected ways in non-neuronal settings. Thus, the reality of broadly transcribed and translated Elav should be taken into consideration in *Drosophila* studies.

miR-279/996 represses Elav within sensory organs and outside of the nervous system

Amongst the many miRNAs that have captured elav within their target cohorts, the mir-279/996 locus is particularly notable. The collected knowledge on this miRNA operon points to important roles in sensory organ development. This locus was one of the first to be characterized by primary transcript in situ hybridization, revealing expression in embryonic CNS and PNS (Aboobaker et al., 2005), but not in differentiated neurons (Stark et al., 2005). This was coupled to bioinformatic evidence that the miR-279/996 seed is enriched for conserved targets that are neurally expressed (Stark et al., 2005). Finally, deletion mutants of miR-279/996 reveal defects in olfactory sensory organs, causing inappropriate specification of ectopic CO2-sensing, Elav+ neurons within the maxillary palp (Cayirlioglu et al., 2008; Sun et al., 2015). The developmental basis of this phenotype remains unknown, but a plausible hypothesis based on their expression and computational patterns as "anti-neuronal" determinants might be that they stem from the transformation of a non-neuronal sensory lineage cell into an ectopic neuron.

These observations stem from the locations of overt expression of mir-279/996 and Elav within the nervous system, comprising "miRNA-hi" and "Elav-hi" territories, which we extend using in situ and transcriptional reporter studies (Sanfilippo et al., 2016). However, we unexpectedly show that their antagonistic relationships extend more broadly to "miRNA-lo" and "Elav-lo" territories, which comprise all imaginal cells and possibly include other settings. In particular, we use stringent knockout analyses to show derepression of Elav protein and GFP-elav sensor in mir-279/996 mutant clones. We do not rule out potential contribution of other miRNAs to elav control, but miR-279/996 exert substantial regulation and constitute a notable example of a potent single miRNA locus-target interaction. Strikingly, post-transcriptional regulation is sufficient to generate the appropriate Elav spatial pattern. Indeed, a ubiquitous reporter linked to the elav 3' UTR actually mimics Elav expression more closely than the elav transcriptional reporter, since tub-GFP-elav 3' UTR is neural-restricted but is responsive to the miRNA pathway and to miR-279/996 in non-neural territories.

Non-neural expression of neural Elav family members is observed in a subset of human malignancies associated with co-occurring paraneoplastic syndromes. Remarkably, characterization of abundant immunoglobulins in these patients led to the discovery of the human Elav ortholog HuD. Subsequent studies revealed that HuD, like Elav in flies, is expressed specifically in mature neurons but is aberrantly expressed in cancers such as small cell lung cancer. The ectopic expression outside of the immune-privileged nervous system mounts a strong immune response that crosses the brain-blood barrier. Here, HuD+ neurons are destroyed leading to neurological paraneoplastic syndromes that often cause patient death (Albert and Darnell, 2004; Darnell, 1996). The mechanism leading to ectopic HuD outside of the nervous system has long been elusive, but HuD notably bears one of the most conserved mammalian 3' UTRs (Siepel et al., 2005). Thus, it is plausible that a post-transcriptional mechanism similar to the one we identified in flies may help restrict HuD/Elav to the nervous system.

CHAPTER 5: Genome-wide profiling of the 3' ends of polyadenylated RNAs

Summary

Alternative polyadenylation (APA) diversifies the 3' termini of a majority of mRNAs in most eukaryotes, and is consequently inferred to have substantial consequences for the utilization of post-transcriptional regulatory mechanisms. Since conventional RNA-sequencing methods do not accurately define mRNA termini, a number of protocols have been developed that permit sequencing of the 3' ends of polyadenylated transcripts (3'-seq). We present here our experimental protocol to generate 3'-seq libraries using a dT-priming approach, including extensive details on considerations that will enable successful library cloning. We pair this with a set of computational tools that allow the user to process the raw sequence data into a filtered set of clusters that represent high-confidence functional pAs. The data are single-nucleotide resolution and quantitative, and can be used for downstream analyses of APA.

Introduction

The final step in the maturation of an mRNA is the recognition of a polyadenylation signal (PAS) at its 3' end leading to cleavage and polyadenylation of the nascent transcript. Although a few genes were known from decades ago to exhibit alternative definition of their 3' termini (Alt et al., 1980; Amara et al., 1982), a process referred to as alternative cleavage and polyadenylation (APA), genome-wide studies eventually revealed this process to be the rule and not the exception. Indeed, a majority of genes in diverse eukaryotic organisms probed to date appear to undergo APA (Derti et al., 2012; Jan et al., 2011; Lianoglou et al., 2013; Mangone et al., 2010; Ozsolak et al., 2010; Smibert et al., 2012). While some of the first examples of APA occur within internal gene regions and affect coding potential (Alt et al., 1980; Amara et al., 1982), most APA sites occur within 3' UTRs (Tian and Manley, 2017).

Since 3' UTRs act as hubs of post-transcriptional regulation, APA has substantial implications for determining alternative usage of diverse regulatory regimes. 3' UTRs can affect transcript stability, localization and/or translation, regulatory events that are often mediated via binding of trans-acting regulators such as RNA binding proteins (RBPs) (Gerstberger et al., 2014) and microRNAs (miRNAs) (Bartel, 2009). Furthermore, the accumulation of alternative 3' UTR isoforms is highly regulated, with isoforms differentially accumulating depending on tissue, developmental and disease states

(Derti et al., 2012; Ji et al., 2009; Masamha et al., 2014; Miura et al., 2013; Smibert et al., 2012). Beyond correlative genome-wide studies, the impact of APA on individual genes can be tangible and substantial. For example, the expression of shorter 3' UTRs on some oncogenes may be associated with transforming properties (Mayr and Bartel, 2009), expression of the long 3' UTR of *α-synuclein* has been shown to be linked to the accumulation and translation of *α-synuclein* transcripts in Parkinson's disease (Rhinn et al., 2012), while the expression of the long 3' UTR isoform of *BDNF* leads to transport and translation of *BDNF* transcript in dendrites (An et al., 2008b).

Despite great interest in this topic, the underlying mechanisms that lead to the expression of different 3' UTR isoforms still remain to be clarified
(Gruber et al., 2014a; Tian and Manley, 2017). Moreover, from the phenotypic viewpoint, much remains to be understood as to the extent that switching of 3' UTR isoforms affects gene regulation. For example, in at least some settings, it was proposed that global 3' UTR isoform modulation has subtle effects on protein outputs (Gruber et al., 2014b; Spies et al., 2013). Therefore, there are clearly great needs for ongoing investigations of APA mechanism and biology. Both of these efforts will often need to utilize strategies to profile 3' UTR isoforms in order to examine the regulation, perturbation, and function of 3' UTR isoforms.

Early genome-wide attempts to understand the diversity of transcripts generated by APA involved analyses of cDNA clones (Gautheret et al., 1998; Tian et al., 2005). These studies were followed by attempts to infer 3' UTR isoforms expression first by microarrays and later by RNA sequencing. However, these techniques can only infer the genomic location of the cleavage events by either looking at already known events in the case of microarrays (Sandberg et al., 2008) or by only recognizing large changes in 3' UTR length by leveraging distinctive changepoints in RNA-seq coverage

(Masamha et al., 2014; Shenker et al., 2015). Because of these limitations several groups have developed specialized protocol to specifically sequence just the 3' ends of mRNAs (Hoque et al., 2013; Jan et al., 2011; Pelechano et al., 2012; Shepard et al., 2011); see (Elkon et al., 2013) for review of these and other strategies. In the course of our efforts to annotate 3' UTR isoforms in *Drosophila* (in preparation), we also developed methods to sequence and analyze 3' termini of polyadenylated transcripts. We present detailed experimental and bioinformatic protocols that permit quantitative, single-nucleotide resolution measurements of sites of cleavage and polyadenylation, allowing APA to be assessed at the transcriptome wide level in a rapid and cost-effective manner.

Description of the method

Overview

3'-seq reports on alternative polyadenylated RNA isoforms by sequencing the junction of the end of the transcript (3' UTR in the case of mRNA) and the polyA tail. This allows for the genome-wide quantification of RNA isoforms that differ at the 3' end (Fig. 6.1). The method can report on any polyadenylated RNAs, including coding and non-coding species. Briefly, the protocol starts by synthesizing cDNA from fragmented total RNA using an RT primer with a biotin at the 5' end, a part of an Illumina adapter, and oligo-dT with a terminal anchor at the 3' end. The oligo-dT sequence recognizes the polyA tail and the anchor at the end of the primer ensures that the oligonucleotide binds at the junction of the polyA tail with the terminus of the transcript. cDNA is converted to dsDNA and bound on magnetic beads. This step allows for direction specific ligation of the Universal Illumina adapter and ease of washing between different steps. The library is PCR amplified and size selected to enrich for reads that contain the junction between the polyA tail and the end of the transcript. Sequencing of these reads followed by mapping to a reference genome enables determination of 3' ends at single nucleotide resolution, and can be used for differential expression analysis.

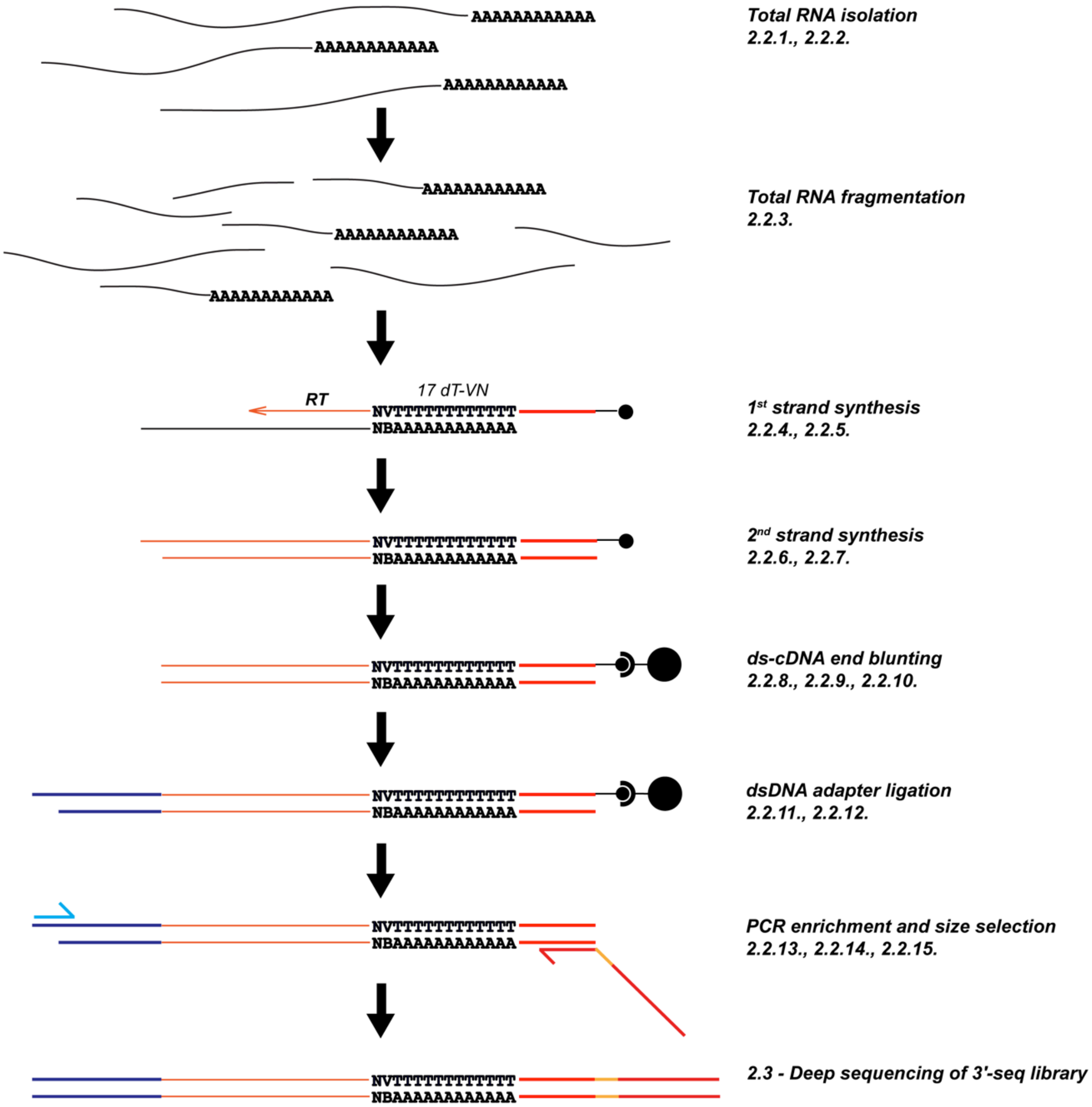


Figure 5.1 Overview of the 3'-seq protocol.
IUPAC codes, V=A/C/G; B=C/G/T; N=A/C/G/T.

Detailed protocol

1.1. Total RNA isolation

This protocol provides a quantitative, genome-wide readout of the 3' ends of all polyadenylated coding and non-coding transcripts. 3'-seq can be performed on any tissue or cell sample, and requires at least 500 ng of total RNA.

Isolate tissue or cell lines on ice. It is important to perform this step quickly, keeping isolated material on ice to avoid RNA degradation.

Prepare total RNA using Trizol® Reagent (Ambion) according to manufacturer's instructions. *Be sure to completely homogenize samples in Trizol®.*

DNase treat total RNA to remove DNA contamination according to standard procedures.

Resuspend total RNA to a concentration of at least 100 ng/µl in nuclease-free water. Store RNA samples at -80 °C.

1.2. Total RNA QC

Determine the quality of prepared total RNA on Bioanalyzer 2100 using the Agilent RNA 6000 Kit. A typical profile for *Drosophila* high quality total RNA is shown, where peaks indicate 18S and 28S rRNA (Fig. 6.2).

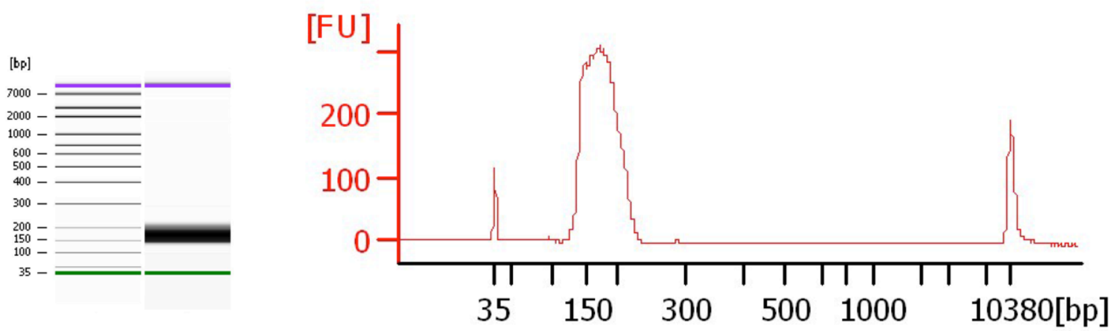


Figure 5.2 Example of high quality total RNA.
Total RNA quality is assessed on the basis of the quality of the predominant signal from rRNA. The trace is from high quality total RNA from *Drosophila melanogaster* run on an Agilent Bioanalyzer 2100. Notice that insect 28S rRNA dissociates into two subunits of equal size that co-migrate with the 18S rRNA. The migration of rRNA of other organisms will vary and should be taken into account when validating total RNA quality.

1.3. Total RNA fragmentation

Total RNA is fragmented using divalent cations under high temperature to obtain small fragments that allow us to pick up the junction between the 3' end of the transcript and the polyA tail upon sequencing. The RT primer is already present in the reaction, however RNA fragments and RT primer do not interact at temperatures lower than 55 °C to minimize internal priming and polyA tail only annealing. Steps 1.3. to 1.7. can be carried out either in 1.5 mL tubes when making up to 16 libraries in parallel or in PCR tubes if processing more than 16 samples (see supplies and equipment for details – sample number limitations are due to the size of the magnet used).

1. Prepare the fragmentation reaction with 500-2000ng of total RNA, 0.8 μM RT primer and 4 μL 5X FS buffer (SSIII kit, Invitrogen) diluted to a final volume of 10 μL with water.
2. Fragment total RNA by placing the reaction on a thermocycler for 10 min at 94 °C.
3. Decrease the temperature of the fragmented total RNA to 55 °C to prepare the RNA for 1st strand cDNA synthesis

Note: The fragmentation time used above has been optimized for Drosophila total RNA. The fragmentation time was chosen by performing a time course of the same reaction with species specific RNA. Here the optimal time that leads to small fragments of 150-200 nucleotides in length prior to complete RNA fragmentation was chosen (Fig. 6.3). Fragmentation might need to be optimized for different RNA preparations.

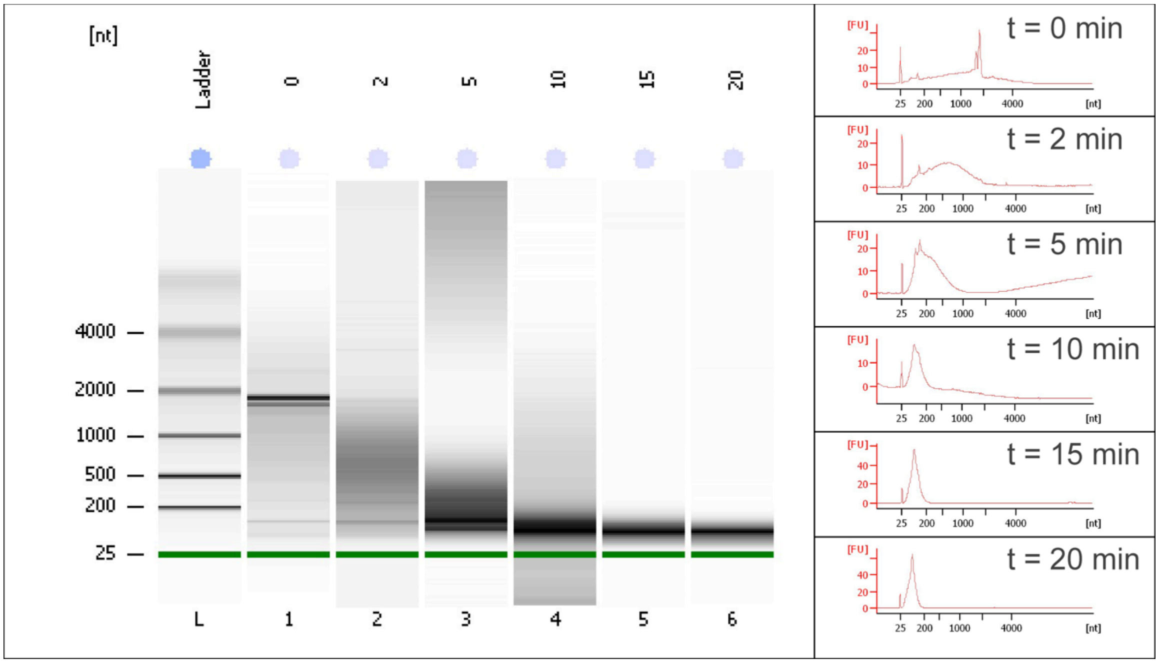


Figure 5.3 Optimization of fragmentation time.
Agilent Bioanalyzer 2100 traces of the reaction outlined in 1.3 stopped at different time intervals. The chemical fragmentation reaction should be stopped when the total RNA peak is around 150-200 nt (10 min) but before the RNA is completely fragmented as shown in the later time points. This step was optimized for fragmentation of *Drosophila* total RNA and should be optimized when using this protocol to determine 3' ends from total RNA of other organisms.

1.4. 1st strand synthesis

In this step cDNA of the 3' end of transcripts is synthesized. The anchor on the oligo-dT RT primer as well as the relatively high temperature of 55 °C ensures proper annealing onto the junction between the end of the 3' UTR and the polyA tail. The RT primer also includes a portion of one of the adapters required for Illumina sequencing to which one of the PCR amplification primers in the final step of the protocol will anneal.

4. Prepare 1st strand synthesis reaction mix by adding 1 mM dNTPs, 20 mM DTT, 20U of RNaseOUT (Invitrogen) and 200U of SuperScript III (Invitrogen).
5. Equilibrate 1st strand synthesis reaction mix to 55 °C on thermocycler.
6. Add mix to the fragmented total RNA and mix by pipetting up and down at least 4 times.
7. Incubate the reaction for 1 h at 55 °C.
8. Inactivate the reaction for 15 min at 70 °C.

Possible stop point. Store at -20 °C.

1.5. 3' terminal biotinylated cDNA cleanup

The cDNA is cleaned up using Ampure XP beads (AGENCOURT BECKMAN) to remove smallest fragments as well as enzymes and buffers.

9. Add 1.5 volumes (45 µL) of Ampure XP beads equilibrated at room temperature.
10. Allow binding of nucleic acids to beads for 5 min.
11. Place on magnetic stand for 5 min or until solution appears clear and remove the supernatant.
12. Wash the beads two times with 70% ethanol according to manufacturer's instructions. *Make fresh 70% ethanol on the same day*.
13. Air dry the beads for 1 min.
14. Elute biotinylated cDNA in 40 µL of 10 mM Tris-HCl pH 8.0 by re-suspending the beads in the elution buffer and place on magnetic stand.
15. Transfer eluted cDNA in a new tube.

Possible stop point. Store at -20 °C.

Note: It is important to use 1.5 volumes of beads to cDNA volume to select the right fragment size range. To ensure this, check that the volume has not changed significantly do to evaporation during the cDNA synthesis step.

1.6. Second strand synthesis

In this step, double stranded biotinylated cDNA is generated. RNase H is used to nick the cDNA/RNA duplex and *E. coli* DNA polymerase I is used to synthesize the second strand of DNA by nick translation.

16. Prepare 2nd strand synthesis reaction mix by adding 0.5 mM dNTPs, 1X NEB2 buffer, 2.5U RNase H (Thermo Scientific), and 20U of *E. coli* DNA polymerase I (NEB) to the above generated cDNA and diluting if necessary to 50 µL with water.
17. Incubate the reaction for 2.5 h at 16 °C.

Possible stop point. Store at -20 °C.

1.7. 3' terminal biotinylated cDNA cleanup

ds-cDNA is cleaned up using Ampure XP beads as in step 1.5. using 1.5 volumes of beads (75 µL) and eluting ds-cDNA in 50 µL of 10 mM Tris-HCl pH 8.0.

1.8. Bind ds-cDNA to magnetic streptavidin beads

Biotinylated ds-cDNA is bound to magnetic streptavidin beads (M-280). This protects one of the ends leading to end specific ligation of the Illumina Universal adapter in the next steps.

18. Wash 50 µL M-280 beads (Invitrogen) two times in 2X B&W buffer.
19. Re-suspend beads in 50 µL 2X B&W buffer.
20. Add the bead mixture to 50µL of the ds-cDNA solution and re-suspend to generate a homogeneous solution.

21. Incubate the solution at room temperature for 30 min with rotation to allow binding.
22. Wash two times with 1X B&W buffer.
23. Wash two times with 1X NEB2 buffer.
24. Transfer the solution to a new tube.

1.9. ds-cDNA end blunting

In this step Klenow fragment is used to generate blunt ends of the ds-cDNA to ensure blunt-end specific ligation of the ds-DNA adapter in step 1.11.

25. Prepare reaction mix by mixing 1 mM dNTPs, 1X NEB2 buffer and 5U of Klenow fragment (NEB) and dilute the reaction mix to a volume of 100 µL with water.
26. Remove the buffer from the ds-cDNA bound beads from 1.8..
27. Add the blunting reaction mix to the bead bound ds-cDNA.
28. Incubate the reaction for 15 min at 25 °C on a thermomixer using interval mixing of 1400 rpm for 15 s followed by 2 min pause.

1.10. Beads clean up

Once the library is bound to the M-280 beads buffers and enzymes are changed using a series of washes that also include a mixture of proteases to ensure enzyme inactivation.

29. Wash the beads one time with 1X buffer C.
30. Remove buffer C and add 100 µL of cleaning solution
31. Incubate the cleaning reaction for 15 min at 37 °C to inactivate enzymes on a thermomixer using interval mixing of 15 s 1400 rpm followed by 2 min pause.
32. Wash three times with 1X buffer D.
33. Wash two times with 1X T4 ligase buffer.
34. Transfer beads to a new tube.

1.11. dsDNA adapter ligation

In this step the Illumina TruSeq Universal adapter is ligated to the ds-cDNA. The dsDNA adapter fragment has a 5' overhang to ensure direction dependent ligation to the bead bound ds-cDNA.

35. Prepare ligation reaction mix by adding 1X T4 DNA ligase buffer, 0.4 µM dsDNA Universal adapter, 2000U of T4 DNA ligase (NEB) and diluting to a volume of 100µL with water.
36. Remove buffer from ds-cDNA bound beads and add the ligation reaction mix.
37. Incubate the ligation reaction overnight at 16 °C on a thermomixer using a shaking cycle of 1400 rpm for 15 s followed by 2 min pause.

1.12. Final beads clean up

38. Wash the beads one time with 1X buffer C.
39. Remove buffer C and add 100 µL of cleaning solution (make fresh solution).
40. Incubate the cleaning reaction for 15 min at 37 °C to inactivate enzymes on a thermomixer using interval mixing of 1400 rpm for 15 s followed by 2 min pause.
41. Wash three times with 1X buffer D.
42. Wash two times with 1X Phusion HF buffer (Thermo Scientific).
43. Transfer beads to a new tube.

1.13. Enrichment PCR

A PCR amplification step is carried out to get the library off the beads and add the remainder of the adapter sequences to the library. Each library should be amplified using a different barcoded forward primer (sequences provided in Table 6.1), choosing barcode combinations according to Illumina guidelines.

44. Prepare PCR reaction mix by adding 1X HF PCR buffer, 6.25 µM each of universal primer and barcoded primer, 1 mM dNTPs and 1U Phusion High-Fidelity DNA polymerase (Thermo Scientific) and diluting to a volume of 50 µL with water.
45. Remove buffer from the beads and add the PCR reaction mix.
46. Perform PCR amplification using this program:

STEP 1: 98 °C 30 s

STEP 2: 98 °C 10 s

STEP 3: 63 °C 30 s

STEP 4: 72 °C 15 s

Cycle to STEP 2 15 times

STEP 5: 72 °C 10 min

STEP 6: 4 °C HOLD

47. Place amplified library on a magnet and transfer supernatant to a new tube.

Possible stop point. Store at -20 °C.

1.14. Library size selection

In this final step, the library is size selected on a non-denaturing polyacrylamide gel such that a significant portion of the library contains the junction between the polyA tail and the 3' end of the transcript upon sequencing the library in 1X 50bp sequencing mode. If using longer reads, the library should be sized accordingly to the increased read length.

48. Precipitate the amplified library with 3X volume of 100% ethanol and 10 µg of glycogen.
49. Incubate for a minimum of 2 h at -20 °C.
50. Pellet the library by spinning in a bench top centrifuge at 20,800 rcf for 20 min at 4 °C.

51. Remove supernatant and dry pellet for 2-5 min to remove traces of ethanol avoiding excessive drying of the pellet.
52. Re-suspend pellet in 10 µL of 1X Ficoll loading buffer.
53. Load sample flanked by 1 µg GeneRuler Low Range DNA ladder (Thermo Scientific), 1X Ficoll loading buffer in an 8% TBE polyacrylamide minigel (Invitrogen) and run gel at 200V for 35 min.
54. Stain gel for 10 min in 1X SYBR Gold (Invitrogen) in TBE buffer.
55. Wash two times in 1X TBE buffer for 10 min.
56. Image gel prior to excision of the library and cut the library using a razor blade in the size range between 175bp and 250bp using flanking ladders as size standards (Fig. 6.4).
57. Place the gel piece in 400µL Lonza elution buffer, rotate overnight at 4 °C.
58. Transfer the eluted library in a new tube.
59. Add 2X volume of 100% ethanol, 10 µg glycogen and precipitate at -20 °C for at least 2 h.
60. Pellet the library by spinning in a bench top centrifuge at 20,800 rcf for 20 min at 4 °C.
61. Wash pellet with 70% ethanol and pellet at 20,800 rcf for 5 min.
62. Remove the ethanol solution and let the pellet dry for ~5 min.
63. Re-suspend pellet in 6 µL nuclease-free water and store at -20 °C.

Note: Avoid excision below 150 bp, as adapter-adapter products are present at around 120-130 bp and should not contaminate the final library preparation.

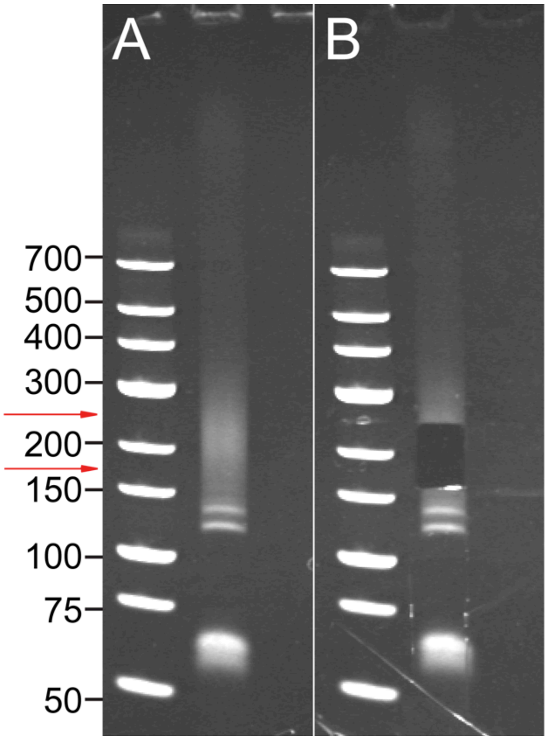


Figure 5.4 3'-seq library size selection.
3'-seq amplified library from step 1.13 was run on an 8% polyacrylamide TBE gel as outlined in 1.14. (A) Image of the library before excision of the relevant range. Bands caused by amplification of adapter-adapter sequences are observed below 150 bp (~120 bp). Care should be taken to avoid excision of these bands. (B) Image of the library post excision. The library is excised between 175-250 bp. This range is optimized for sequencing 1 X 50 bp. The range could be extended if sequencing using longer reads. This size range enriches for reads that contain the 3' end/polyA tail junction.

1.15. Library QC

Validate the library size distribution and quantify library amount.

64. Check the quality of 3'-seq library preparation by running 1 µL of sample on a Bioanalyzer 2100 with an Agilent High Sensitivity DNA Kit. A size distribution that reflects the size range extracted from the gel should be obtained (Fig. 6.5). Quantify the library either by using the Bioanalyzer estimate or other methods, such as Qubit fluorometric quantitation.

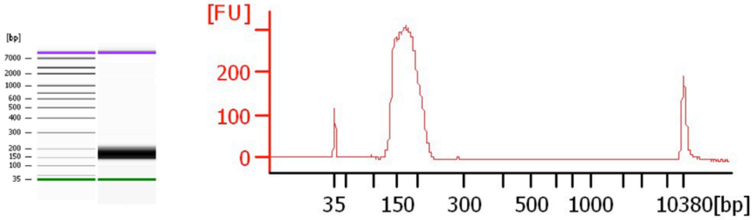


Figure 5.5 Example Agilent Bioanalyzer 2100 trace of 3'-seq library.
High quality library should result in a size range distribution as expected from the size range extracted from the gel. Shown is a 3'-seq library that falls in the 150-250 bp size range. Isolation of a fraction of reads below 175bp is to be expected and will still results in reads that can be mapped to the genome. The peak should be above 130 bp to avoid contamination of adapter-adapter reads (~120 bp).

Deep sequencing

Sequencing of the library can be done on any Illumina platform that supports TruSeq adapters. Libraries were successfully sequenced using Illumina HiSeq1000 and HiSeq2000 for single end 50 bp reads. Longer reads could be used to increase the proportion of the library that is sequenced into the polyA tail, however using 50 bp is sufficient to have 60-70% of trimmed reads spanning the junction between the polyA tail and the 3' end of the transcript map to a reference genome. Higher depth should be favored for discovery of rare events while lower depth and more replicates should be used for differential expression analysis of 3' UTR isoforms among samples. Barcoded adapters allow for libraries to be multiplexed and sequenced in the same lane. The library should be sequenced with the addition of 5% PhiX spike in control to provide increased complexity because of the large presence of polyA segments in the library.

Bioinformatics Analysis

3'-seq provides single nucleotide resolution of the 3' ends of polyadenylated RNA transcripts. This is achieved by mapping the reads that encompass the 3' end/polyA junction after trimming of the untemplated polyA. The 3' end nucleotide of those reads represents the nucleotide upstream to the cleavage site generated by the cleavage and polyadenylation machinery. The bioinformatics approach reported here provides commands to use available software as well as custom made java utilities (available on github) to map reads, filter internal priming and cluster adjacent events to call final 3' end positions. The analysis is designed to run in a Unix environment with >8 Gb of available RAM.

2.1. De-multiplexing and QC

De-multiplexing of reads can be performed with the Illumina CASAVA utility, which splits the fastQ file by barcode. Standard quality assessment on the fastQ files can be done using FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Over representation of polyA stretches that result from sequencing of polyA tail should be expected.

2.2. Mapping to the reference genome

The reads generated with this protocol will map to 3' ends of transcripts. Depending on downstream applications, the reads that map just upstream of the cleavage site can also be used to increase library depth and these reads will be labeled *untrimmed*, and will map to the genome without trimming. The remaining reads require trimming of untemplated polyA stretches to map onto the reference genome. The last nucleotide on the 3' end of these reads represents the nucleotide just upstream to the cleavage site of nascent transcripts. These reads will be labeled *trimmed* and provide us with the nucleotide level resolution of 3' end formation.

2.2.1. Mapping the untrimmed reads

The mapping to the reference genome is achieved using GSNAP (v. 2013-03-31) (Wu and Nacu, 2010).

1. Build a genomic index file for a specific reference genome. The reference genome sequence can be downloaded either from UCSC genome browser (http://genome.ucsc.edu/) or from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/).

Input:

Reference genome sequence in FASTA format – *genome.fasta*

Name of reference genome – *genome_name*

```
$ gmap_build -D /path/to/genome_index_folder \
 -d genome_name genome.fasta
```

Output:

/path/to/genome_index_folder/index_files

2. Align the untrimmed reads to the genome. After mapping the reads with GSNAP, the script below will generate a BAM file of mapped untrimmed reads using SAMtools (Li et al., 2009). The folder structure used below will be important for the final step.

Input:

De-multiplexed fastq.gz sequenced reads file – *sample.fastq.gz*

Name of the sample for output – *sample*

```
$ mkdir -p mapped/untrimmed/
$ cd mapped/untrimmed
$ gsnap -D /path/to/genome_index_folder -N 0 --gzip2 \
-A sam -d genome_name sample.fastq.gz | \
samtools view -bS - | samtools sort - sample
```

Output:

sample.bam

3. Index BAM file.

Input:

BAM file generated above – *sample.bam*

```
$ samtools index sample.bam
```

Output:

sample.bam.bai

2.2.2. Trimming of unmapped reads

The unmapped reads include the polyA/3' end junction reads. A fraction of these will also contain parts of adapter sequence. A java tool is provided to trim both the polyA tail and the adapter sequences off the 3' end of the unmapped reads.

Input:

BAM file generated above - */path/to/mapped/untrimmed/sample.bam*

```
$ java -Xmx4g -jar TrimUnmapped.jar trim3p -bam \
sample.bam -out /dev/stdout | gzip - > \
sample_trimmed.fastq.gz
```

Output:

sample_trimmed.fastq.gz

2.2.3. Mapping of trimmed reads

The untrimmed reads can be mapped as in steps 2 and 3 of 2.2.1 using as input the *sample_trimmed.fastq.gz* file. The resulting bam and bam.bai files should be placed in /path/to/mapped/trimmed folder for downstream analysis.

Output:

/path/to/mapped/trimmed/sample.bam

/path/to/mapped/trimmed/sample.bam.bai

2.3 Internal priming mask

An inherent problem with 3' end sequencing approaches is that stretches of genomically encoded adenosine (A) nucleotides can serve as templates for priming first strand synthesis. This is called ‘internal priming’ as 3' end calls are erroneously made upstream of bona fide termini. A filtering strategy is employed to minimize these erroneous annotations. Provided is a tool that uses a sliding window approach to identify genomic stretches with a certain percentage of A nucleotides provided by the user. The window used below is 16 nt wide and requires at least 9 As in the window. These parameters were derived to filter internal priming events occurring during 3'-seq of *Drosophila* total RNA. The width of window and number of As should be optimized for the particular organism being studied. This optimization can be performed using known alternative 3' ends if a confident set already exists.

Input:

Reference genome sequence in FASTA format – *genome.fasta*

Name of the file for output - *primingMask.bed*

Parameters:

-window – window in which to assess A-richness. The window starts one nucleotide off the 3' of the read (since the first nucleotide after the cleavage site is most often an A)

-maxAdenosine – max number of As in the window. Below this number, in this case 9, a read passes the filter. Every position with equal or above maxAdenosine in the provided window will be reported in the mask.

```
$ java -Xmx2g -jar ThreeSeqPipeline.jar \
IdentifyInternalPriming -fa genome.fasta \
-maxAdenosine 9 -primingMask primingMask.bed \
-window 16
```

Output:

primingMask.bed

The internal priming mask will be used in the next step to minimize false positive calls due to internal priming.

Note: An alternative method to identify internally primed events is to add 2 or 3 As to the 3' end of the trimmed reads and map these requiring no mismatches to the reference genome. The reads that map to the genome are likely to come from internal priming events, as the As are template in the genome, and can be omitted from the downstream analysis.

2.4. Clustering and 3' end calling

To call the predominant cleavage site, events occurring within a short distance of each other can be clustered. In this step the utility will cluster events occurring in a 25 nt window and call as the cluster end the most abundant position. The utility takes as input any number of libraries and the final called event is the most abundant amongst all libraries. Parameters for clustering window size as well as read number thresholds can be provided by the user and are outlined in detail online as below. Additionally, the events will be filtered using the internal priming mask generated in the previous step. Finally, each called event will be quantified in each sample provided. Additional parameters and details are outlined in the appendix online (https://github.com/piotrsan/3seq_javaUtilities).

Input:

Reference genome sequence in FASTA format – *genome.fasta*

Priming mask - */path/to/primingMask.bed*

Directory for gtf output – *gtf_files*

Text file with sample names without file extension, one per line - *sample_name.txt*

Parameters:

-minDistinctReads – minimum number of unique reads which are needed to call a specific 3' end.

The script should be run in the folder containing the *mapped* folder created above.

```
$ mkdir htsjdk_tmp
$ java -Xmx4g -Djava.io.tmpdir=htsjdk_tmp \
-jar ThreeSeqPipeline.jar DefineClusters \
-minDistinctReads 3 -inDir mapped -trimmed trimmed \
-untrimmed untrimmed -primingMask /path/to/primingMask.bed\
-outDir gtf_files -baseNames sample_name.txt
```

Output:

atlas.gtf - the 3' end positions are reported in this file. A detailed description is provided online (https://github.com/piotrsan/3seq_javaUtilities)

gtf_files/ - the output in this folder is described in detail online as above

Note: Existing annotation of 3' ends, if available, can be used to quantify events by using the mapped reads from step 2.2., using the annotated 3' ends as features to count the reads obtained using 3'-seq.

2.5. Basic analysis of 3'-seq data

The analysis steps provided in 2.1.-2.4. allow the user to annotate all predominant 3' ends of polyadenylated transcripts. Further analysis steps depend on the specific user study application. For example, one can compare changes in 3' UTR expression by taking the ratio of the two most dominant isoforms of 3' UTRs with multiple ends. Comparison between sample will indicate if a certain gene expresses longer or shorter 3' UTR isoforms. Sequence upstream and downstream of the derived 3' ends can be analyzed for the presence of enriched PAS sequences that might be specific to processing in certain tissues or organisms. Furthermore, the output of 3'-seq can be used to quantify gene expression by deriving gene counts and using gene expression analysis software such as DeSeq2 (Love et al., 2014).

Troubleshooting

Total RNA QC: If isolating good quality total RNA proves difficult, revisit the extraction method to minimize the time the sample is not in Trizol®. Smaller batches of dissected tissue can be placed in Trizol® and subsequently pooled to minimize degradation of RNA.

1st strand synthesis: If the resulting library results in excessive amounts of reads mapping to polyA tail, make sure that the reaction was not allowed to cool down below 55 °C as that will increase priming of the RT primer onto just polyA tail. A certain degree of internal priming and only polyA sequencing is to be expected.

PCR amplification: If PCR duplicates appear to be a problem, the number of cycles can be adjusted or a random k-mer sequence could be added to the barcode to control for PCR duplication.

3'-seq reads visualization: The reads can be visualized using a genome browser such as IGV (Thorvaldsdóttir et al., 2013) or the UCSC Genome Browser (http://www.genome.ucsc.edu). Visualization of the trimmed reads is sufficient to visualize the 3' end events. The atlas.gtf output can be uploaded to visualize the dominant 3' ends obtained in step 2.4. Visual analysis of these files as well as the mapped reads can be useful to determine success of the sequencing run, with reads clustering onto ends of annotated transcript models.

Concluding remarks

The recent realization that the majority of genes in eukaryotic organisms undergo APA adds an additional layer of complexity to post-transcriptional gene regulation. The complex pattern of 3' UTR isoforms expression appears to be under substantial regulation as seen in both normal and pathological biological settings. This 3'-seq protocol provides an additional tool to probe this diverse aspect of gene expression.

The mechanisms that lead to the steady state accumulation of different 3' UTR isoforms are just emerging. The genome-wide assessment of 3' UTR isoforms expression during normal as well as perturbed conditions will be an important tool to further our understanding of the regulatory mechanisms that control the expression of different 3' UTR isoforms.

Materials

Enzymes and chemicals

TRIzol® Reagent (Ambion, cat. No. 15596026)

SuperScript III Reverse Transcriptase (Invitrogen, cat. No. 18080-093) including:

5X First Strand Buffer (FS buffer)

0.1 M DTT

100 mM dNTPs (Invitrogen, dGTP – cat. No. 10218014, dCTP – cat. No. 10217016, dATP – cat. No. 10216018, dTTP – cat. No. 10219012)

RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen, cat. No. 10777019)

100% Ethanol (Decon, cat. No. 2716G)

DNase/RNase free water

Agencourt Ampure XP (BECKMAN COULTER, cat. No. A63880)

1 M UltraPure Tris-HCl pH 8.0 (Invitrogen, cat. No. 15568025)

DNA polymerase I (E. coli) (NEB, cat. No. M0209S) including:

10X NEB2 buffer

RNaseH (Thermo Scientific, cat. No. EN0201)

Dynabeads M-280 Streptavidin (Invitrogen, cat. No. 11205D)

DNA Polymerase I, Large (Klenow) Fragment (NEB, cat. No. M0210S)

T4 DNA Ligase (NED, cat. No. M0202L)

Phusion High-Fidelity DNA Polymerase (Thermo Scientific, cat. No. F530S)

SYBR Gold Nucleic Acid Gel Stain (Invitrogen, cat. No. S11494)

GeneRuler Low Range DNA Ladder (Thermo Scientific, cat. No. SM1191)

Pronase (Roche, cat. No. 10165921001)

Supplies and Equipment

Benchtop Centrifuge

1.5 mL tubes rotor

Thermocycler

Thermomixer C (Eppendorf, cat. No. 2231000269)

Sterile filter tips

Nonstick, RNase-free Microfuge Tubes, 1.5 mL (Applied Biosystems, cat. No. AM12450)

Dynamag-2 Magnet (Thermo Scientific, cat. No. 12321D)

Novex® TBE Gels, 8%, 10 well (Invitrogen, cat. No. EC6215BOX)

For processing more than 16 samples in parallel

Magnetic Stand-96 (Thermo Scientific, cat. No. AM10027)

SmartBlock™ PCR 96 (Eppendorf, cat. No. 5306000006) This can be omitted by processing samples in batches using the Dynamag-2 magnet

PCR Plate, 96-well, segmented, semi-skirted (Thermo Scientific, cat. No. AB-0900)

Solutions

2X B&W buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA pH 8.0, 2 M NaCl)

Buffer C (1X PBS, 0.01% Tween 20)

Buffer D (10 mM Tris-HCl pH 8.0, 2 mM EDTA pH 8.0, 0.01% Tween 20)

Pronase stock (10 mg/mL in water – stable 6 mo. at 4 °C)

Cleaning Solution (1X PBS, 1mM $CaCl_2$, 15 µg pronase)

5X Ficoll loading buffer (18 mM Tris Base, 10 mM Boric Acid, 0.4 mM EDTA, 3% Ficoll Type 400, 0.02% Bromo Blue, 0.02% Xylene Cyanol – stable 6 mo. at 4 °C)

Lonza elution buffer (0.5 mM Ammonium Acetate, 15 μM Magnesium Acetate, 1 mM EDTA pH 8.0, 1% SDS)

10X Annealing buffer (0.1 M Tris-HCl pH 7.5, 0.5 M NaCl, 10 mM EDTA pH 8.0)

dsDNA Universal adapter (50 μM Truseq_Universal_Adapter_F (Table 6.1), 50 μM Truseq_Universal_Adapter_R (Table 6.1), 1X Annealing buffer. Heat at 95 °C for 5 min. Remove heat block and let cool down to room temperature for annealing of the two primers)

Oligos

A list of the oligos required to generate libraries is provided in Table 6.1.

Java tools and output files description (Java utilities were generated by Sol Shenker)

Java tools can be downloaded from github where a detailed description of the utilities is also provided: https://github.com/piotrsan/3seq_javaUtilities.

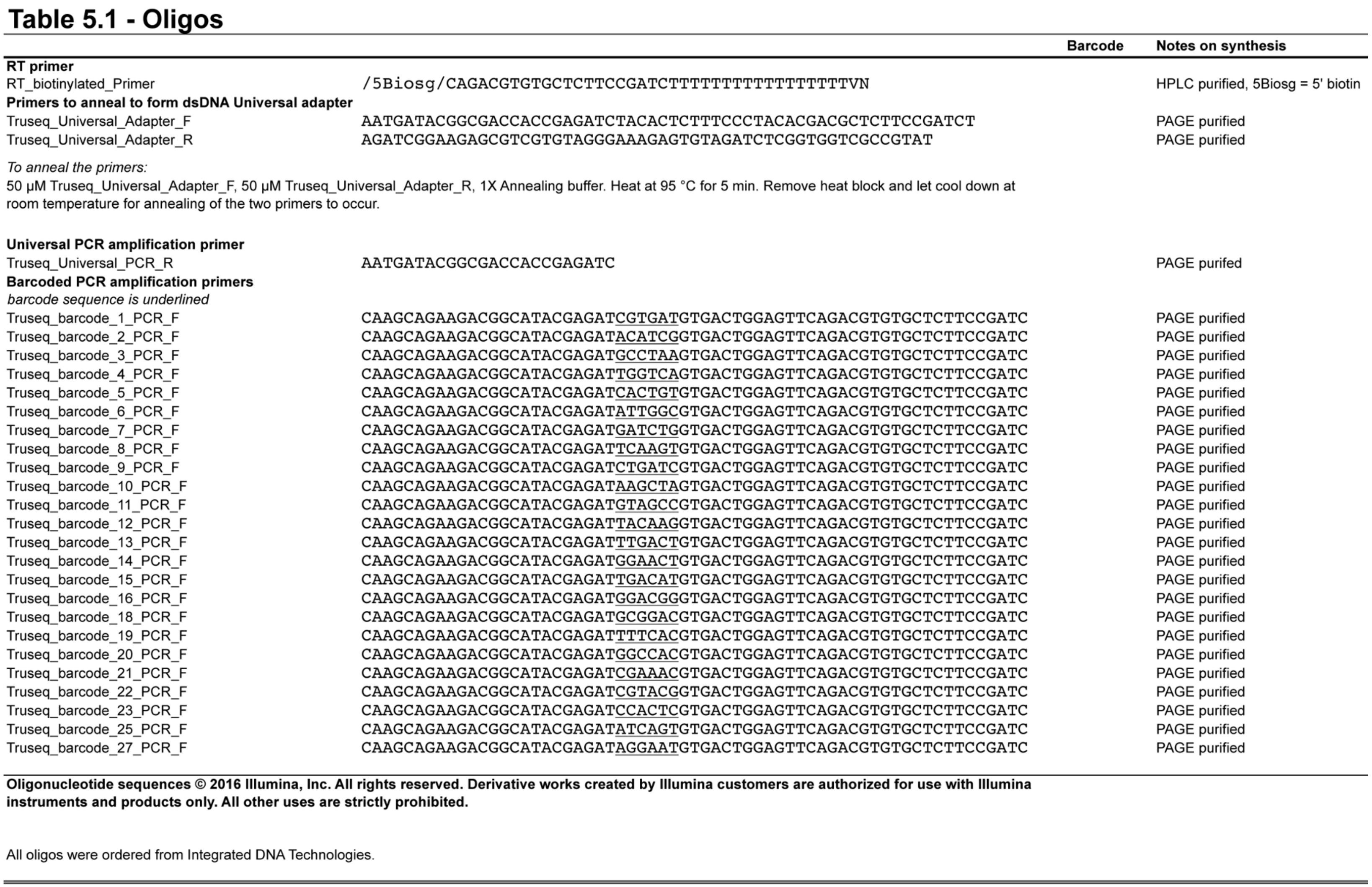


Table 5.1 - Oligos

		Barcode	Notes on synthesis
RT primer			
RT_biotinylated_Primer	/5Biosg/CAGACGTGTGCTCTTCCGATCTTTTTTTTTTTTTTTTTVN		HPLC purified, 5Biosg = 5' biotin
Primers to anneal to form dsDNA Universal adapter			
Truseq_Universal_Adapter_F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT		PAGE purified
Truseq_Universal_Adapter_R	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTAT		PAGE purified
To anneal the primers: 50 μM Truseq_Universal_Adapter_F, 50 μM Truseq_Universal_Adapter_R, 1X Annealing buffer. Heat at 95 °C for 5 min. Remove heat block and let cool down at room temperature for annealing of the two primers to occur.			
Universal PCR amplification primer			
Truseq_Universal_PCR_R	AATGATACGGCGACCACCGAGATC		PAGE purifed
Barcoded PCR amplification primers			
barcode sequence is underlined			
Truseq_barcode_1_PCR_F	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_2_PCR_F	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_3_PCR_F	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_4_PCR_F	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_5_PCR_F	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_6_PCR_F	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_7_PCR_F	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_8_PCR_F	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_9_PCR_F	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_10_PCR_F	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_11_PCR_F	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_12_PCR_F	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_13_PCR_F	CAAGCAGAAGACGGCATACGAGATTTGACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_14_PCR_F	CAAGCAGAAGACGGCATACGAGATGGAACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_15_PCR_F	CAAGCAGAAGACGGCATACGAGATTGACATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_16_PCR_F	CAAGCAGAAGACGGCATACGAGATGGACGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_18_PCR_F	CAAGCAGAAGACGGCATACGAGATGCGGACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_19_PCR_F	CAAGCAGAAGACGGCATACGAGATTTTCACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_20_PCR_F	CAAGCAGAAGACGGCATACGAGATGGCCACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_21_PCR_F	CAAGCAGAAGACGGCATACGAGATCGAAACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_22_PCR_F	CAAGCAGAAGACGGCATACGAGATCGTACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_23_PCR_F	CAAGCAGAAGACGGCATACGAGATCCACTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_25_PCR_F	CAAGCAGAAGACGGCATACGAGATATCAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified
Truseq_barcode_27_PCR_F	CAAGCAGAAGACGGCATACGAGATAGGAATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		PAGE purified

All oligos were ordered from Integrated DNA Technologies.

Table 5.1 Oligos for 3'-seq

CHAPTER 6: DISCUSSION

Arguably one of the most surprising outcomes from the sequencing of the human genome has been the realization of the small number of protein coding genes encoded in the genome relative to the number of protein coding genes found in lower organisms, such as the nematode *C. elegans* (C. elegans Sequencing Consortium, 1998). Some of the highest gene number estimates shortly prior to the release of the human genome sequence were in excess of 100,000 protein coding genes (Liang et al., 2000), exceedingly more than the as few as 19,000 protein coding genes predicted today (Ezkurdia et al., 2014). The failure to identify a higher number of protein coding genes in higher organisms, has shifted the attention of researchers onto non-coding DNA elements as well as the alternative processing of higher eukaryotic transcriptomes as more likely players in specifying the increased complexity of higher organisms. Indeed, alternative RNA processing, with the ability to both increase protein diversity as well as the regulatory potential of a particular locus, holds great promise as an important player in the complex biology observed with the progress of evolution (Licatalosi and Darnell, 2010).

Despite early insights into the formation of polyadenylated 3' ends (Proudfoot and Brownlee, 1976) it took several decades to appreciate the complexity of alternative cleavage and polyadenylation (Zhang et al., 2005). The advent of the genomics revolution has helped uncover APA as a common feature of transcriptomes and how this process works to diversify the 3' ends of mRNA, both impacting protein coding as well as regulatory potential of a gene locus (Tian and Manley, 2017). The last few years have seen the annotation of these events for humans and all major model organisms(Jan et al., 2011; Lianoglou et al., 2013; Mangone et al., 2010; Ulitsky et al., 2012; Zhang et al., 2005), with the exception of *D. melanogaster*. Our work fills this gap, providing a comprehensive annotation of 3' end events in *D. melanogaster* as well as the related *D. yakuba* and *D. virilis*. This resource is an important contribution to the annotation of the *Drosophila* transcriptome and provides a platform to investigate evolutionary dynamics that contribute to the conservation and divergence of cis-elements involved in regulating tissue specific patterns of 3' end isoform expression. Furthermore, we demonstrate the benefit of using *Drosophila* to investigate both the mechanism of CNS 3' UTR extension by Elav family RBPs as well as the role of the 3' UTR in specifying protein expression at the organismal level as showcased in our study on the role of the *elav* 3' UTR in specifying Elav neural expression.

The annotation of 3' ends of *Drosophila*, being the last major model organism missing annotation, closes the initial phase of APA research, which focused on cataloguing 3' end diversity as well as investigating biological settings in which 3' end isoforms undergo differential expression. Few recent studies have begun the next phase in APA investigation, understanding the biological output of such diversity and the consequence and role of regulating differential length 3' UTR isoforms. Important examples are the role of the long 3' UTR of *BDNF* in localizing transcript to dendrites for localized translation (An et al., 2008a) as well as the implication of the long 3' UTR isoform of *α-synuclein* in promoting higher protein levels that contribute to the pathogenicity of this protein in Parkinson's disease (Rhinn et al., 2012). Despite these studies, it is currently not clear what and if there is a general role of expressing different 3' UTR isoforms, particularly the long 3' UTR isoforms that are predominantly expressed in the CNS. Dissection of the role of such events at specific loci will be instrumental to generate a more general model of the role of neural APA in neurobiology, as well as the effect of the absence of such isoform in non neural tissues. We believe *Drosophila*, with the exceptional genetic toolkit available as well as the tremendous genomic resources compiled over the years, will constitute an important model for the studies of APA at the organismal level. Our organism wide 3' UTR GFP reporter studies demonstrate this, underscoring how only the observation of the function of these genetic elements in the context of the whole organism, can begin to explain the role of such important patterns of isoform regulation at the tissue level, and how 3' UTRs can affect processes such as translation in a tissue specific manner. Investigation of the role that APA plays in specifying cell fate and particular tissue physiology, tied back to the molecular consequences of differential 3' UTR expression, will enormously benefit of studies in the *Drosophila* system and will certainly provide us with important discoveries in years to come.

CHAPTER 7: MATERIALS AND METHODS

Cell and tissue samples preparation

Samples were prepared as detailed in Chapter 5 (section 1.1). Samples for head were isolated from 2-5 days old *D. melanogaster* (Canton S), *D. yakuba* or *D. virilis* mated female flies by cutting precisely at the head, which mostly excluded the thoracic VNC. Ovary and testis samples were also isolated from 2-5 days old *D. melanogaster* (Canton S), *D. yakuba* or *D. virilis* flies. For embryo collection, 2-5 day old flies were raised on apple juice agar plates on yeast paste for the required time windows (D. melanogaster: 0-45", 45"-90", 90"-6h, 6-12h, 12-18h, 18-24h, *D. yakuba*: 0-12h, 12-24h, *D. virilis*: 0-12h, 12-24h, 24-36h, 36-48h). The whole span of embryogenesis was represented for each species. Next, embryos were either collected and placed in TRIZOL or incubated for additional time to age the embryos as required. Whole flies were 2-5 days old mated males or females. Carcass samples of 2-5 days old males or females were obtained by cutting off the head and removing either the ovaries or the testes. Total RNA samples were obtained as part of the modENCODE project (Cherbas et al., 2011).

RNA-seq libraries preparation

RNAseq libraries were prepared using the Illumina TruSeq Stranded Total RNA Library Prep Kit (cat. No. RS-122-2201) starting with 1 µg of total RNA in water. The protocol was followed exactly as per manufacturer's instructions. The libraries were sequenced on a hiseq-1000 sequencer at the genomics core facility at MSKCC.

3'-seq libraries preparation, mapping and atlas generation

3'-seq libraries were prepared and sequenced as outlined in detail in Chapter 5, using 2 µg of total RNA as starting material. The libraries were sequenced using a hiseq-1000 with SE-50 mode at the genomics core facility at MSKCC. Raw FASTQ files were mapped onto the genome assemblies and 3' end clusters were derived and quantified as described in Chapter 5 (section 2). The internal priming filter parameter (9/16 As) was derived empirically by maximixing the identification of annotation encoded 3' ends. Atlas of 3' ends were de-novo derived for L1 CNS samples as well as for the sets of samples of S2 cells reported in Chapter 3. 3'-seq of S2 samples with the over-expression of WT or MT-3X Elav family RBPs were generated by Sonali Majumdar)

Bioinformatics

The entire analysis was performed using R if not otherwise specified and bioconductor packages (https://bioconductor.org/).

Reference Genomes, annotations and chain files

The following reference genome files were downloaded from FlyBase (Gramates et al., 2017): *D. melanogaster* (dm6) (Hoskins et al., 2015), D. *yakuba* (droYak3) and *D. virilis* (droYak3). The following annotations were downloaded from

FlyBase and used in our analysis: *D. melanogaster* (r6.12) (Hoskins et al., 2015), *D. yakuba* (r1.05) and *D. virilis* (r1.06). Chain files for comparative genome analysis were downloaded from the UCSC genome browser (Tyner et al., 2017) (http://hgdownload-test.soe.ucsc.edu/goldenPath/dm6/). Gene orthology tables for r6.12 were downloaded from FlyBase.]

modENCODE RNA-seq data mapping (performed by Jiayu Wen)

Total RNA-seq was mapped from data from modENCODE samples available on GEO (SRR1197403, SRR1197401, SRR1197396, SRR1197280, SRR1197410, SRR1197406, SRR1197398, SRR1197402, SRR1197397, SRR1197399, SRR500470, SRR1197428, SRR1197427, SRR1197471, SRR1197470, SRR1197465, SRR1197464, SRR1197433, SRR1197434, SRR1197432, SRR1197431, SRR1197436, SRR1197435, SRR1197430, SRR1197429, SRR1197469, SRR1197468, SRR1197370, SRR1197337, SRR1197367, SRR1197334, SRR1197369, SRR1197332, SRR1197331, SRR1197365, SRR1197330, SRR1197363, SRR1197327, SRR1197368, SRR1197336, SRR1197364, SRR1197329, SRR1197366, SRR1197328, SRR1197338, SRR1197333, SRR1197335, SRR1197475, SRR1197292, SRR1197372) to UCSC *Drosophila melanogaster* (dm6) genome assembly. Hisat2 aligner was used for the alignment with default parameters (Kim et al., 2015).

Comparison of expression quantification with of RNA-seq or 3'-seq (performed by Jiayu Wen)

RNA-seq and 3'-seq by the trimmed mean of M-values (TMM) normalization method in the edgeR/Limma Bioconductor library (Oshlack et al., 2010). The voom method of Limma was used (Ritchie et al., 2015) to correct for the Poisson noise due to the discrete counts of RNA-seq.

Gene expression analysis

Gene counts of either RNA-seq or 3'-seq, using our extended annotation, were computed using FeatureCounts (Liao et al., 2014). Differential gene expression analysis was performed using standard workflow using the DESeq2 (Love et al., 2014) package in bioconductor. For the L1 CNS analysis, triplicates where used for each condition. For the analysis of Elav family RBP over-expression, the 3 different WT or 3X-MT samples were each used as pseudo-replicates to assess the effect on gene expression of over-expression of an Elav family member RBP.

Assignment of 3' end clusters to genomic features

First different isoform annotations (5' UTR, intron, CDS) were collapsed prior to assignment of 3' end positions to identify regions that could be strictly defined (in the order intron<5' UTR<CDS<3' UTR). For genes that had more than one 3' UTR annotated, the 3' UTR with most distal 5' start coupled with the longest 3' end annotation for the gene was classified as terminal 3' UTR. The internal 3' UTRs annotated, which are derived through recognition of intronic or CDS pAs, were retained to account for the location of the 3' ends, however our analysis of 3' UTR isoform expression does not take into account those isoforms. To attribute our 3' ends to genomic features, we assigned them base on overlapping the annotated features with the following hierarchy, 3' UTR, CDS, 5' UTR and intron. Given this, for example an end falling in both annotated intron and 3' UTR of a gene will be assigned to the 3' UTR. If the end did not overlap with existing annotation it was initially flagged as intergenic. We attributed ends falling 3' of annotation within a 5 kb window if a matched RNAseq sample had continuous coverage in the window between the annotation and the 3' end. The coverage was assessed using isoSCM (Shenker et al., 2015) which creates annotation of 3' UTRs based on RNAseq if the coverage contains gaps that are less than 100 nt. For this we used our head, ovary and testis RNAseq and additionally, modENCODE RNAseq of an embryogenesis time course, whole fly and carcass as well as the matched cell line samples (Brown et al., 2014).

Analysis of predominant PAS or DSE around putative 3' ends

To assess sequence composition around cleavage sites, nucleotide distributions were computed around the cleavage site. To identify the most common PAS upstream of the cleavage site, we looked for the most represented hexamer in a 50 nt window upstream of the cleavage site. Once this was identified, we removed those ends and repeated the process again. The same was done to identify putative DSE elements, by looking in the 50 nt window downstream of cleavage sites.

Identification of genes that express alternative 3' UTR isoforms

For this and all downstream analyses, we looked at 3' ends in the 3' most terminal 3' UTR, as described above. Additionally we excluded all genes that contain introns in the 3' UTR from our analysis for simplicity. To assess the pattern of 3' UTR expression for a given gene, we calculated a weighted 3' UTR length, as previously reported (Bao et al., 2016). Briefly, we took the average of all 3' UTR isoforms length per gene weighted by the contribution of each isoform. We called as changed between two samples genes that changed weighted 3' UTR length between two tissues by at least 100 bp. A gene was considered in the analysis only if it was expressed at or above 5 RPM in both tissues.

Analysis of 3' end relative strength

To test the hypothesis that intrinsic pA strength can contribute to the pattern of 3' UTR isoform expression observed for a given gene, we calculate strength score for each signal. The expression as quantified by 3'-seq at each end in the terminal 3' UTR was normalized to total expression, giving the ratio of total expression of each 3' UTR isoform. The strength at each site was calculated as normalized expression of the end divided by the normalized expression of the end and the ones downstream. This gave us a measure of PAS strength if assuming that only strength of pAs was responsible for the patterns of 3' UTR isoform expression observed. A strength of 1 means that no isoforms are detected beyond that site (max strength) and a score of 0 means that the site is not recognized while downstream ones are. The strength score is not calculated for ends beyond the one end with evidence of recognition. We calculated a difference in strength between the same pA site in two different samples by taking the difference of the strength scores.

3' UTR expression pattern conservation

To identify the genes that show a similar pattern of 3' UTR expression between tissues, we considered the genes that are orthologous in all three species and that are expressed at 5 RPM minimum in both tissues and in all three species.

Orthologous 3' end conservation

For this analysis we only considered ends that had as evidence at least 3 reads in the head, ovary and testis libraries, as these have been sequenced in common between the three species. To identify the syntenic sites between *D. melanogaster* and *D. yakuba* or *D. virilis* we used liftOver. pA sites that were +/- 25 nt from the lifted *D. melanogaster* end were reciprocal best matches when the proposed syntenic site from the other species was lifted onto the *D. melanogaster* genome. Additionally, we only analyzed ends on terminal 3' UTRs that have defined orthologs in all three species.

Molecular cloning of Elav paralogs (Performed by Sonali Majumdar)

Elav, Fne and Rbp9 were cloned into pPAS-5C, an insect cloning vector driven by *Drosophila* actin promoter. 3X-RRM point mutations were engineered as previously reported (Lisbin et al., 2000).

Northern blotting (performed by Sonali Majumdar)

Northern blotting was performed as previously described (Lianoglou et al., 2013). Northern blot probes were generated using the following primers: goα (F: TGGCAAACACACAAACACG; R: AGAGCAAGAGCACAAGTGAGG), AcCoAs (F: CGAGGTATTCGACCAGAAGC; R: TGGTGAGCATGTCAACTACG) and Rpl32 (F: AGCATACAGGCCCAAGATCG ; R: CGCTTCTTGGAGGAGACG)

Isolation of L1 CNS (performed by Alex Panzarino)

For L1 CNS samples, the *elav*^{5} mutant embryos were identified by the lack of fluorescent balancer and picked 24 hours after egg laying, ensuring that the larva inside the case exhibited movement of the head. WT and elav^5-rescue larvae were dissected immediately after eclosion 24 h hours after egg laying. L1 CNS was immediately placed in a small aliquot of PBS on ice and shortly afterward in TRIZOL LS. Several dissections were pulled together to obtain ~100 L1 CNS per sample. Samples were collected in triplicates.

Fly stocks and *Drosophila* genetics

To generate 3' UTR sensor transgenes, we used recombineering to insert the entire *elav* or *tubulin* 3' UTRs, including ~1 kb downstream of the most distal cleavage site, downstream of a *tubulin-GFP* cassette (see recombineering section). To generate the *elav* rescue construct we recombineered a 40kb segment around the *elav* locus in p[acman]-KO (Chan et al., 2011). The resulting attB-p[acman]GFP-3' UTR sensors and *elav* rescue constructs were integrated into attP2 site (BDSC#8622) by BestGene, Inc. (Chino Hills).

We used published alleles of miRNA pathway factors on FRT backgrounds: *FRT42D drosha[21K11]* (Smibert et al., 2011), *FRT82B Dcr-1[Q1147X]* (Lee et al., 2004), *FRT82B pasha[KO]* (Martin et al., 2009); a recombinant *FRT82B, Dcr-1, pasha* chromosome was recombined. RNAi lines were from the Vienna *Drosophila* RNAi Center: *UAS-GW182-RNAi* (VDRC 103581), *UAS-elav-RNAi* (VDRC 37915) and *UAS-mei-P26-RNAi* (VDRC 101060). Other published mutants and transgenes included *elav[4]* (Campos et al., 1985), *elav-lacZ* (Yao and White, 1994), *mir-7[delta1]* (Li and Carthew, 2005), *mir-8[delta1]* (Shcherbata et al., 2007), *mir-279/996[15C]* (Sun et al., 2015), *16.6kb mir-279/996-GFP* (Sun et al., 2015), *UAS-elav[2e2]*, *UAS-elav[3e3]*, *ubi-GFP-UgGH* (Toba et al., 2002) and *UAS-mei-P26* (Page et al., 2000). *Drosophila virilis* and *Drosophila yakuba* were obtained from the *Drosophila* Species Stock Center.

Generation of constructs transgenes by recombineering

The whole *elav* or *tubulin* 3' UTRs, including ~1 kb downstream of the most distal cleavage site, were cloned by recombineering downstream of a *tubulin-GFP* cassette.

P[acman]-CmR-tub-GFP. To insert tubulin driven GFP into attB-P[acman]-CmR-F-2-attB first the annealed oligos RF1/RR1 were cloned into AscI/PacI digested attB-P[acman]-CmR-F-2-attB to create P[acman]-CmR-tub-GFP-LARA.

P[acman]-CmR-tub-GFP-LARA was digested with BamHI and used to retrieve tub-GFP from SW102 containing JB26 (Brennecke et al., 2005) by conventional recombineering.

P[acman]-ApR-tub-GFP. To be able to retrieve genomic sequence from p[acman]-CmR BAC genomic libraries we changed the resistance cassette to ApR by recombineering the tub-GFP segment into attB-P[acman]-ApR-F-2-5-attB. First we cloned the same RF1/RR1 into AscI/PacI digested attB-P[acman]-ApR-F-2-5-attB to obtain P[acman]-ApR-tub-GFP-LARA. P[acman]-ApR-tub-GFP-LARA was digested with BamHI to retrieve tub-GFP from SW102 containing P[acman]-CmR-tub-GFP by conventional recombineering.

P[acman]-ApR-tub-GFP-elav-3' UTR. To retrieve the whole elav 3' UTR and ~1000 bp of downstream sequence left arm (primers elav.LA.F/elav.LA.R) and right arm (elav.RA.F/elav.RA.R) were amplified from genomic DNA and stitched together using primers elav.LA.F/elav.RA.R in a PCR reaction containing left and right arm. The resulting left-right arm amplicon (with BamHI between the two arms) was cloned by *in vitro* recombineering using cold-fusion (SBI, cat. no. MC010B) into NotI digested P[acman]-ApR-tub-GFP following manufacturer guidelines to obtain P[acman]-ApR-tub-GFP-elav3UTR-LARA. P[acman]-ApR-tub-GFP-elav3UTR-LARA was linearized using BamHI and electroporated into SW102 cells containing the BAC CH321-35G14 (Venken et al., 2009) to retrieve the *elav* 3' UTR by conventional recombineering.

P[acman]-ApR-tub-GFP-tubulin-3' UTR. The tubulin 3' UTR and ~1000 bp of downstream sequence was amplified from genomic DNA using tub.F/tub.R. The resulting amplicon was cloned by *in vitro* recombineering using cold-fusion into NotI digested P[acman]-ApR-tub-GFP.

P[acman]-KO-elav-rescue

The genomic region around *elav* of coordinates chrX:497510-537452 (dm6) was inserted into attB-p[acman]-KO (Chan et al., 2011) according to conventional recombineering.

Immunohistochemistry

Imaginal-disc clones were induced 72 h AEL with either 60-min (mitotic or MARCM) or 15-min (FLP-out) 37°C heat shock and fixed 72 h later. Primary antibodies were chicken anti-GFP (1:1,000, Abcam cat. no. ab13970), mouse anti-beta-galactosidase (1:50, 40-1a, DSHB), rabbit anti−Mei-P26 (1:1,000, gift of P. Lasko), rat anti-Elav (1:50, 7E8A10, DSHB), mouse anti-Orb (1:50, 4H8, DSHB), rabbit anti-Dpn (1:1000, gift of Y. Jan), rabbit anti-Scrt (1:1000, our laboratory), rabbit anti-Shep (1:50, gift of E. Lei), mouse anti-Fas2 (1:50, 1D4, DSHB), mouse anti-Fas3 (1:50, 7G10, DSHB), mouse anti-Futsch (1:100, 22C10, DSHB), mouse-anti-Cut (1:100, DSHB), BP102 (1:100 DSHB) and rabbit anti-cleaved caspase-3 (1:250, Cell Signaling cat. no. 9661). Secondary antibodies were made in donkey and conjugated to Alexa-488, -568, or -647 (Jackson ImmunoResearch).

Cell culture and Western Blotting (performed by Sonali Majumdar in Chapter 3)

S2 cells were obtained from the *Drosophila* Genomics Resource Center and were confirmed to have the appropriate morphology of growth characteristics of S2 cells. RNAi-mediated knockdowns using *GFP*, *Dcr-1* and *Dcr-2* dsRNAs and subsequent western blotting are as described (Smibert et al., 2013). For Western blot we used rat anti-Elav (1:50, 7E8A10, DSHB) or anti-HA (12CA5). goat anti-Actin(1:500, SC-1616, Santa Cruz) was used as loading control in Chapter 4.

List of primers

RF1-CGCGCCGAATTCGATATCAAGCTTGCACAGGTCCTGTTCGATAACGTCGTACTCGGGAAGGATCCCACTCTCGGCATGGACGAGCTGTACAAGTAAAGCGGCCGCATAGGCCACTTTAAT

RR1-TAAAGTGGCCTATGCGGCCGCTTTACTTGTACAGCTCGTCCATGCCGAGAGTGGGATCCTTCCCGAGTACGACGTTATCGAACAGGACCTGTGCAAGCTTGATATCGAATTCGG

elav.LA.F-GACGAGCTGTACAAGTAAAGCGGCCCAAATGGAAGTGGAC elav.LA.R-AGGTCATCTGgctagcGATCGACTGTGCCAACCTTT

elav.RA.F-ACAGTCGATCgctagcCAGATGACCTTGATCCTGGC

elav.RA.R-GATCCACTAGTGGCCTATgcggccgcAATCACAGCCAACAACAGCA

tub.F-GACGAGCTGTACAAGTAAGCGTCACGCCACTTCAACGCTC

tub.R-GATCCACTAGTGGCCTATgcggccgcAAAGGCGCCAGTCTCTACCGGT

Genotypes analyzed

Mitotic clones

y, w, hsFLP ;; FRT82B / FRT82B, ubi-GFP, M(3)

y, w, hsFLP ;; FRT82B, Dcr-1[Q1147X] / FRT82B, ubi-GFP, M(3)

y, w, hsFLP ;; FRT82B, pasha[KO] / FRT82B, ubi-GFP, M(3)

y, w, hsFLP ;; FRT82B, Dcr-1[Q1147X], pasha[KO] / FRT82B, ubi-GFP, M(3)

y, w, hsFLP ;; FRT82B, mir-279/996[15C] / FRT82B, ubi-GFP, M(3)

y, w, hsFLP ; FRT42D, drosha[21K11] / FRT42D, arm-lacZ

y, w, ubx-FLP ; FRT42D, mir-7[delta1] / FRT42D, ubi-GFP, cell lethal

y, w, hsFLP ; FRT42D, mir-8[delta1] / FRT42D, ubi-GFP

y, w, hsFLP ;; FRT82B, pasha[KO] / P[tub-GFP-tubulin-3' UTR]attP2, FRT82B, arm-lacZ

y, w, hsFLP ;; FRT82B, mir-279/996[15C] / P[tub-GFP-tubulin-3' UTR]attP2, FRT82B, arm-lacZ

y, w, hsFLP/+ ; UgGH/+ ; FRT82B, Dcr-1[Q1147X] / FRT82B, M(3), arm-lacZ

y, w, hsFLP/+ ; UgGH/+ ; FRT82B, pasha[KO] / FRT82B, M(3), arm-lacZ *w, elav[4], FRT19A / armlacZ, FRT19 ; ; 70FLP/+*

MARCM clones

y, w, hsFLP, UAS-GFP ; ; FRT82 / tub-gal4, FRT82B, tub-gal80

y, w, hsFLP, UAS-GFP ; ; FRT82, Dcr-1[Q1147X] / tub-gal4, FRT82B, tub-gal80

y, w, hsFLP, UAS-GFP ; ; FRT82, pasha[KO] / tub-gal4, FRT82B, tub-gal80

y, w, hsFLP, UAS-GFP ; UAS-GW182-RNAi / + ; FRT82 / tub-gal4, FRT82B, tub-gal80

y, w, hsFLP, UAS-GFP ; mei-P26 RNAi / + ; FRT82, pasha[KO] / tub-gal4, FRT82B, tub-gal80

y, w, hsFLP, UAS-GFP ; UAS-GW182-RNAi / + ; FRT82 / tub-gal4, FRT82B, tub-gal80

y, w, hsFLP, UAS-GFP ; + / CyO, elav-lacZ / + ; FRT82, Dcr-1[Q1147X] / tub-gal4, FRT82B, tub-gal80

FLP-out clones

y,w, hsFLP; ; act>CD2>GAL4, UAS-GFPnls / UAS-elav-RNAi

y,w, hsFLP; UAS-mei-P26 / + ; act>CD2>GAL4, UAS-GFPnls / +

y,w, hsFLP; UAS-mei-P26-RNAi / + ; act>CD2>GAL4, UAS-GFPnls / +

y,w, hsFLP; UAS-mei-P26 / act>CD2>GAL4, UAS-lacZ ; P[tub-GFP-tubulin-3' UTR]attP2 / +

y, w, hsFLP/+ ; UgGH / act>CD2>GAL4, UAS-lacZ ; UAS-elav[3e3] / +

y,w, hsFLP; UAS-elav[2e2] / act>CD2>GAL4, UAS-lacZ

y,w, hsFLP; UAS-GFP / act>CD2>GAL4, UAS-lacZ

BIBLIOGRAPHY

Aboobaker, A. A., Tomancak, P., Patel, N., Rubin, G. M. and Lai, E. C. (2005). Drosophila microRNAs exhibit diverse spatial expression patterns during embryonic development. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 18017–18022.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., et al. (2000). The genome sequence of Drosophila melanogaster. *Science* **287**, 2185–2195.

Albert, M. L. and Darnell, R. B. (2004). Paraneoplastic neurological degenerations: keys to tumour immunity. *Nat Rev Cancer* **4**, 36–44.

Alt, F. W., Bothwell, A. L., Knapp, M., Siden, E., Mather, E., Koshland, M. and Baltimore, D. (1980). Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. *Cell* **20**, 293–301.

Amara, S. G., Evans, R. M. and Rosenfeld, M. G. (1984). Calcitonin/calcitonin gene-related peptide transcription unit: tissue-specific expression involves selective use of alternative polyadenylation sites. *Mol Cell Biol* **4**, 2151–2160.

Amara, S. G., Jonas, V., Rosenfeld, M. G., Ong, E. S. and Evans, R. M. (1982). Alternative RNA processing in calcitonin gene expression generates mRNAs encoding different polypeptide products. *Nature* **298**, 240–244.

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

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

Andersson, T., Rahman, S., Sansom, S. N., Alsio, J. M., Kaneda, M., Smith, J., O'Carroll, D., Tarakhovsky, A. and Livesey, F. J. (2010). Reversible block of mouse neural stem cell differentiation in the absence of dicer and microRNAs. *PLoS ONE* **5**, e13453.

Avendaño-Vázquez, S. E., Dhir, A., Bembich, S., Buratti, E., Proudfoot, N. and Baralle, F. E. (2012). Autoregulation of TDP-43 mRNA levels involves interplay between transcription, splicing, and alternative polyA site selection. *Genes Dev* **26**, 1679–1684.

Aviv, H. and Leder, P. (1972). Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid-cellulose. *Proceedings of the National Academy of Sciences of the United States of America* **69**, 1408–1412.

Bank, A., Terada, M., Metafora, S., Dow, L. and Marks, P. A. (1972). In vitro synthesis of DNA components of human genes for globins. *Nature New Biol.* **235**, 167–169.

Bao, J., Vitting-Seerup, K., Waage, J., Tang, C., Ge, Y., Porse, B. T. and Yan, W. (2016). UPF2-Dependent Nonsense-Mediated mRNA Decay Pathway Is Essential for Spermatogenesis by Selectively Eliminating Longer 3'UTR Transcripts. *PLoS Genet* **12**, e1005863.

Barolo, S., Walker, R., Polyanovsky, A., Freschi, G., Keil, T. and Posakony, J. W. (2000). A Notch-independent activity of Suppressor of Hairless is required for normal mechanoreceptor physiology. *Cell* **103**, 957–969.

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233.

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

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.

Bellen, H. J., Tong, C. and Tsuda, H. (2010). 100 years of Drosophila research and its impact on vertebrate neuroscience: a history lesson for the future. *Nat. Rev. Neurosci.* **11**, 514–522.

Berezikov, E., Robine, N., Samsonova, A., Westholm, J. O., Naqvi, A., Hung, J.-H., Okamura, K., Dai, Q., Bortolamiol-Becet, D., Martin, R., et al. (2011). Deep annotation of Drosophila melanogaster microRNAs yields insights into their processing, modification, and emergence. *Genome Res* **21**, 203–215.

Berger, C., Renner, S., Luer, K. and Technau, G. M. (2007). The commonly used marker ELAV is transiently expressed in neuroblasts and glial cells in the Drosophila embryonic CNS. *Dev. Dyn.* **236**, 3562–3568.

Bergsten, S. E. and Gavis, E. R. (1999). Role for mRNA localization in translational activation but not spatial restriction of nanos RNA. *Development* **126**, 659–669.

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.

Blair, S. S. (2003). Genetic mosaic techniques for studying Drosophila development. *Development* **130**, 5065–5072.

Boelens, W. C., Jansen, E. J., van Venrooij, W. J., Stripecke, R., Mattaj, I. W. and Gunderson, S. I. (1993). The human U1 snRNP-specific U1A protein inhibits polyadenylation of its own pre-mRNA. *Cell* **72**, 881–892.

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

Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., Booth, B. W., Wen, J., Park, S., Suzuki, A. M., et al. (2014). Diversity and dynamics of the Drosophila transcriptome. *Nature* **512**, 393–399.

Brownlee, G. G., Cartwright, E. M., Cowan, N. J., Jarvis, J. M. and Milstein, C. (1973). Purification and sequence of messenger RNA for immunoglobulin light chains. *Nature New Biol.* **244**, 236–240.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012–2018.

Calvo, O. and Manley, J. L. (2001). Evolutionarily conserved interaction between CstF-64 and PC4 links transcription, polyadenylation, and termination. *Mol Cell* **7**, 1013–1023.

Campos, A. R., Grossman, D. and White, K. (1985). Mutant alleles at the locus elav in Drosophila melanogaster lead to nervous system defects. A developmental-genetic analysis. *Journal of neurogenetics* **2**, 197–218.

Campos, A. R., Rosen, D. R., Robinow, S. N. and White, K. (1987). Molecular analysis of the locus elav in Drosophila melanogaster: a gene whose embryonic expression is neural specific. *EMBO J* **6**, 425–431.

Cannavò, E., Koelling, N., Harnett, D., Garfield, D., Casale, F. P., Ciglar, L., Gustafson, H. E., Viales, R. R., Marco-Ferreres, R., Degner, J. F., et al. (2017). Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature* **541**, 402–406.

Cayirlioglu, P., Kadow, I. G., Zhan, X., Okamura, K., Suh, G. S. B., Gunning, D., Lai, E. C. and Zipursky, S. L. (2008). Hybrid neurons in a microRNA mutant are putative evolutionary intermediates in insect CO2 sensory systems. *Science* **319**, 1256–1260.

Chalfie, M., Horvitz, H. R. and Sulston, J. E. (1981). Mutations that lead to reiterations in the cell lineages of C. elegans. *Cell* **24**, 59–69.

Chan, C.-C., Scoggin, S., Wang, D., Cherry, S., Dembo, T., Greenberg, B., Jin, E. J., Kuey, C., Lopez, A., Mehta, S. Q., et al. (2011). Systematic discovery of Rab GTPases with synaptic functions in Drosophila. *Curr Biol* **21**, 1704–1715.

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

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

Cherbas, L., Willingham, A., Zhang, D., Yang, L., Zou, Y., Eads, B. D., Carlson, J. W., Landolin, J. M., Kapranov, P., Dumais, J., et al. (2011). The transcriptional diversity of 25 Drosophila cell lines. *Genome Res* **21**, 301–314.

Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., et al. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**, 203–218.

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

Colombrita, C., Silani, V. and Ratti, A. (2013). ELAV proteins along evolution: back to the nucleus? *Mol. Cell. Neurosci.* **56**, 447–455.

Conne, B., Stutz, A. and Vassalli, J. D. (2000). The 3' untranslated region of messenger RNA: A molecular 'hotspot' for pathology? *Nature Medicine* **6**, 637–641.

Conway, L. and Wickens, M. (1985). A sequence downstream of A-A-U-A-A-A is required for formation of simian virus 40 late mRNA 3' termini in frog oocytes. *Proceedings of the National Academy of Sciences of the United States of America* **82**, 3949–3953.

Cramer, P., Cáceres, J. F., Cazalla, D., Kadener, S., Muro, A. F., Baralle, F. E. and Kornblihtt, A. R. (1999). Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer. *Mol Cell* **4**, 251–258.

Crawford, R. J., Scott, A. C. and Wells, J. R. (1977). Organization of sequences of avian globin mRNA. *Eur. J. Biochem.* **72**, 291–299.

Creemers, E. E., Bawazeer, A., Ugalde, A. P., van Deutekom, H. W. M., van der Made, I., de Groot, N. E., Adriaens, M. E., Cook, S. A., Bezzina, C. R., Hubner, N., et al. (2016). Genome-Wide Polyadenylation Maps Reveal Dynamic mRNA 3'-End Formation in the Failing Human Heart. *Circulation Research* **118**, 433–438.

Dai, W., Li, W., Hoque, M., Li, Z., Tian, B. and Makeyev, E. V. (2015). A post-transcriptional mechanism pacing expression of neural genes with precursor cell differentiation status. *Nat Commun* **6**, 7576.

Dai, W., Zhang, G. and Makeyev, E. V. (2012). RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage. *Nucleic Acids Res* **40**, 787–800.

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

Darnell, R. B. (1996). Onconeural antigens and the paraneoplastic neurologic disorders: at the intersection of cancer, immunity, and the brain. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 4529–4536.

Davidson, L., Muniz, L. and West, S. (2014). 3' end formation of pre-mRNA and phosphorylation of Ser2 on the RNA polymerase II CTD are reciprocally coupled in human cells. *Genes Dev* **28**, 342–356.

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

Di Giammartino, D. C., Li, W., Ogami, K., Yashinskie, J. J., Hoque, M., Tian, B. and Manley, J. L. (2014). RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3' UTRs. *Genes Dev* **28**, 2248–2260.

Duff, M. O., Olson, S., Wei, X., Garrett, S. C., Osman, A., Bolisetty, M., Plocik, A., Celniker, S. E. and Graveley, B. R. (2015). Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature* **521**, 376–379.

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R. and Hood, L. (1980). Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* **20**, 313–319.

Edmonds, M. and Abrams, R. (1960). Polynucleotide biosynthesis: formation of a sequence of adenylate units from adenosine triphosphate by an enzyme from thymus nuclei. *J Biol Chem* **235**, 1142–1149.

Edmonds, M. and Winters, M. A. (1976). Polyadenylate polymerases. *Prog. Nucleic Acid Res. Mol. Biol.* **17**, 149–179.

Edmonds, M., Vaughan, M. H. and Nakazato, H. (1971). Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly-labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship. *Proceedings of the National Academy of Sciences of the United States of America* **68**, 1336–1340.

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

Elkon, R., Ugalde, A. P. and Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. *Nat Rev Genet* **14**, 496–506.

Emery, J. F. and Bier, E. (1995). Specificity of CNS and PNS regulatory subelements comprising pan-neural enhancers of the deadpan and scratch genes is achieved by repression. *Development* **121**, 3549–3560.

Erson-Bensan, A. E. (2016). Alternative polyadenylation and RNA-binding proteins. *J. Mol. Endocrinol.* **57**, F29–34.

Erson-Bensan, A. E. and Can, T. (2016). Alternative Polyadenylation: Another Foe in Cancer. *Mol. Cancer Res.* **14**, 507–517.

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A. and Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum. Mol. Genet.* **23**, 5866–5878.

Fan, X. C. and Steitz, J. A. (1998). Overexpression of HuR, a nuclear-cytoplasmic shuttling protein, increases the in vivo stability of ARE-containing mRNAs. *EMBO J* **17**, 3448–3460.

Fitzgerald, M. and Shenk, T. (1981). The sequence 5“-AAUAAA-3”forms parts of the recognition site for polyadenylation of late SV40 mRNAs. *Cell* **24**, 251–260.

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

Flynt, A. S. and Lai, E. C. (2008). Biological principles of microRNA-mediated regulation: shared themes amid diversity. *Nat Rev Genet* **9**, 831–842.

Galli, G., Guise, J. W., McDevitt, M. A., Tucker, P. W. and Nevins, J. R. (1987). Relative position and strengths of poly(A) sites as well as transcription termination are critical to membrane versus secreted mu-chain expression during B-cell development. *Genes Dev* **1**, 471–481.

Galli, G., Guise, J., Tucker, P. W. and Nevins, J. R. (1988). Poly(A) site choice rather than splice site choice governs the regulated production of IgM heavy-chain RNAs. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 2439–2443.

Gautheret, D., Poirot, O., Lopez, F., Audic, S. and Claverie, J. M. (1998). Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res* **8**, 524–530.

Gawande, B., Robida, M. D., Rahn, A. and Singh, R. (2006). Drosophila Sex-lethal protein mediates polyadenylation switching in the female germline. *EMBO J* **25**, 1263–1272.

Gerstberger, S., Hafner, M. and Tuschl, T. (2014). A census of human RNA-binding proteins. *Nat Rev Genet* **15**, 829–845.

Gho, M., Lecourtois, M., Geraud, G., Posakony, J. W. and Schweisguth, F. (1996). Subcellular localization of Suppressor of Hairless inDrosophilasense organ cells during Notch signalling. *Development* **122**, 1673–1682.

Gil, A. and Proudfoot, N. J. (1984). A sequence downstream of AAUAAA is required for rabbit beta-globin mRNA 3'-end formation. *Nature* **312**, 473–474.

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

Gilmartin, G. M. and Nevins, J. R. (1991). Molecular analyses of two poly(A) site-processing factors that determine the recognition and efficiency of cleavage of the pre-mRNA. *Mol Cell Biol* **11**, 2432–2438.

Gonzalez, C. (2013). Drosophila melanogaster: a model and a tool to investigate malignancy and identify new therapeutics. *Nat Rev Cancer* **13**, 172–183.

Graber, J. H., Cantor, C. R., Mohr, S. C. and Smith, T. F. (1999). In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 14055–14060.

Gramates, L. S., Marygold, S. J., Santos, G. D., Urbano, J.-M., Antonazzo, G., Matthews, B. B., Rey, A. J., Tabone, C. J., Crosby, M. A., Emmert, D. B., et al. (2017). FlyBase at 25: looking to the future. *Nucleic Acids Res* **45**, D663–D671.

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., et al. (2011). The developmental transcriptome of Drosophila melanogaster. *Nature* **471**, 473–479.

Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P. and Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27**, 91–105.

Gruber, A. R., Martin, G., Keller, W. and Zavolan, M. (2014a). Means to an end: mechanisms of alternative polyadenylation of messenger RNA precursors. *WIREs RNA* **5**, 183–196.

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

Han, J., Pedersen, J. S., Kwon, S. C., Belair, C. D., Kim, Y.-K., Yeom, K.-H., Yang, W.-Y., Haussler, D., Blelloch, R. and Kim, V. N. (2009). Posttranscriptional crossregulation between Drosha and DGCR8. *Cell* **136**, 75–84.

Hardy, J. G. and Norbury, C. J. (2016). Cleavage factor Im (CFIm) as a regulator of alternative polyadenylation. *Biochem. Soc. Trans.* **44**, 1051–1057.

Hart, R. P., McDevitt, M. A., Ali, H. and Nevins, J. R. (1985). Definition of essential sequences and functional equivalence of elements downstream of the adenovirus E2A and the early simian virus 40 polyadenylation sites. *Mol Cell Biol* **5**, 2975–2983.

Hartenstein, V. (1993). *Atlas of Drosophila development*.

Herranz, H., Hong, X., Perez, L., Ferreira, A., Olivieri, D., Cohen, S. M. and Milan, M. (2010). The miRNA machinery targets Mei-P26 and regulates Myc protein levels in the Drosophila wing. *EMBO J* **29**, 1688–1698.

Higgs, D. R., Goodbourn, S. E., Lamb, J., Clegg, J. B., Weatherall, D. J. and Proudfoot, N. J. (1983). Alpha-thalassaemia caused by a polyadenylation signal mutation. *Nature* **306**, 398–400.

Hild, M., Beckmann, B., Haas, S. A., Koch, B., Solovyev, V., Busold, C., Fellenberg, K., Boutros, M., Vingron, M., Sauer, F., et al. (2003). An integrated gene annotation and transcriptional profiling approach towards the full gene content of the Drosophila genome. *Genome Biol* **5**, R3.

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

Hilgers, V., Perry, M. W., Hendrix, D., Stark, A., Levine, M. and Haley, B. (2011a). Neural-specific elongation of 3' UTRs during Drosophila development. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 15864–15869.

Hilgers, V., Perry, M. and Hendrix, D. (2011b). Neural-specific elongation of 3′ UTRs during Drosophila development.

Hirose, Y. and Manley, J. L. (2000). RNA polymerase II and the integration of nuclear events. *Genes Dev* **14**, 1415–1429.

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

Hollerer, I., Curk, T., Haase, B., Benes, V., Hauer, C., Neu-Yilik, G., Bhuvanagiri, M., Hentze, M. W. and Kulozik, A. E. (2016). The differential expression of alternatively polyadenylated transcripts is a common stress-induced response mechanism that modulates mammalian mRNA expression in a quantitative and qualitative fashion. *RNA* **22**, 1441–1453.

Hollerer, I., Grund, K., Hentze, M. W. and Kulozik, A. E. (2014). mRNA 3'end processing: A tale of the tail reaches the clinic. *EMBO Mol Med* **6**, 16–26.

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

Hoskins, R. A., Carlson, J. W., Wan, K. H., Park, S., Mendez, I., Galle, S. E., Booth, B. W., Pfeiffer, B. D., George, R. A., Svirskas, R., et al. (2015). The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome Res* **25**, 445–458.

Hu, W., Li, S., Park, J. Y., Boppana, S., Ni, T., Li, M., Zhu, J., Tian, B., Xie, Z. and Xiang, M. (2017). Dynamic landscape of alternative polyadenylation during retinal development. *Cell Mol Life Sci* **74**, 1721–1739.

Hummel, T., Krukkert, K., Roos, J., Davis, G. and Klämbt, C. (2000). Drosophila Futsch/22C10 is a MAP1B-like protein required for dendritic and axonal development. *Neuron* **26**, 357–370.

Hwang, H.-W., Park, C. Y., Goodarzi, H., Fak, J. J., Mele, A., Moore, M. J., Saito, Y. and Darnell, R. B. (2016). PAPERCLIP Identifies MicroRNA Targets and a Role of CstF64/64tau in Promoting Non-canonical poly(A) Site Usage. *Cell Rep* **15**, 423–435.

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

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

Ji, X., Wan, J., Vishnu, M., Xing, Y. and Liebhaber, S. A. (2013). αCP Poly(C) binding proteins act as global regulators of alternative polyadenylation. *Mol Cell Biol* **33**, 2560–2573.

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

Ji, Z., Lee, J. Y., Pan, Z., Jiang, B. and Tian, B. (2009). Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 7028–7033.

Ji, Z., Luo, W., Li, W., Hoque, M., Pan, Z., Zhao, Y. and Bin Tian (2011). Transcriptional activity regulates alternative cleavage and polyadenylation. *Molecular Systems Biology* **7**, 1–13.

Jia, X., Yuan, S., Wang, Y., Fu, Y., Ge, Y., Ge, Y., Lan, X., Feng, Y., Qiu, F., Li, P., et al. (2017). The role of alternative polyadenylation in the antiviral innate immune response. *Nat Commun* **8**, 14605.

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

Karginov, F. V., Cheloufi, S., Chong, M. M., Stark, A., Smith, A. D. and Hannon, G. J. (2010). Diverse endonucleolytic cleavage sites in the mammalian transcriptome depend upon microRNAs, Drosha, and additional nucleases. *Mol Cell* **38**, 781–788.

Kawase-Koga, Y., Low, R., Otaegi, G., Pollock, A., Deng, H., Eisenhaber, F., Maurer-Stroh, S. and Sun, T. (2010). RNAase-III enzyme Dicer maintains signaling pathways for differentiation and survival in mouse cortical neural stem cells. *Journal of Cell Science* **123**, 586–594.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res* **12**, 996–1006.

Kim, D., Langmead, B. and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360.

Kim-Ha, J., Kim, J. and Kim, Y. J. (1999). Requirement of RBP9, a Drosophila Hu homolog, for regulation of cystocyte differentiation and oocyte determination during oogenesis. *Mol Cell Biol* **19**, 2505–2514.

Kornblihtt, A. (2005). Promoter usage and alternative splicing. *Curr Opin Cell Biol* **17**, 262–268.

Koushika, S. P., Soller, M. and White, K. (2000). The Neuron-Enriched Splicing Pattern of Drosophila erect wing Is Dependent on the Presence of ELAV Protein. *Mol Cell Biol* **20**, 1836–1845.

La Mata, de, M., Alonso, C. R., Kadener, S., Fededa, J. P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D. and Kornblihtt, A. R. (2003). A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* **12**, 525–532.

Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W. and Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. *Curr Biol* **12**, 735–739.

Lai, S. L., Miller, M. R., Robinson, K. J. and Doe, C. Q. (2012). The Snail family member Worniu is continuously required in neuroblasts to prevent Elav-induced premature differentiation. *Dev Cell* **23**, 849–857.

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 Res* **36**, 5581–5590.

Lee, R. C., Feinbaum, R. L. and Ambros, V. (1993). TheC. elegansheterochronic genelin-4encodes small RNAs with antisense complementarity tolin-14. *Cell* **75**, 843–854.

Lee, Y. S., Nakahara, K., Pham, J. W., Kim, K., He, Z., Sontheimer, E. J. and Carthew, R. W. (2004). Distinct Roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA Silencing Pathways. *Cell* **117**, 69–81.

Letsou, A. and Bohmann, D. (2005). Small flies--big discoveries: nearly a century of Drosophila genetics and development. *Dev. Dyn.* **232**, 526–528.

Lécuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T. R., Tomancak, P. and Krause, H. M. (2007). Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* **131**, 174–187.

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

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

Li, W., You, B., Hoque, M., Zheng, D., Luo, W., Ji, Z., Park, J. Y., Gunderson, S. I., Kalsotra, A., Manley, J. L., et al. (2015). Systematic Profiling of Poly(A)+ Transcripts Modulated by Core 3' End Processing and Splicing Factors Reveals Regulatory Rules of Alternative Cleavage and Polyadenylation. *PLoS Genet* **11**, e1005166.

Li, X. and Carthew, R. W. (2005). A microRNA Mediates EGF Receptor Signaling and Promotes Photoreceptor Differentiation in the Drosophila Eye. *Cell* **123**, 1267–1277.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L. and Quackenbush, J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat Genet* **25**, 239–240.

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

Liao, Y., Smyth, G. K. and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930.

Licatalosi, D. D. and Darnell, R. B. (2010). RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet* **11**, 75–87.

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464–469.

Lim, L. and Canellakis, E. S. (1970). Adenine-rich polymer associated with rabbit reticulocyte messenger RNA. *Nature* **227**, 710–712.

Lisbin, M. J., Gordon, M., Yannoni, Y. M. and White, K. (2000). Function of RRM domains of Drosophila melanogaster ELAV: Rnp1 mutations and rrm domain replacements with ELAV family proteins and SXL. *Genetics* **155**, 1789–1798.

Lisbin, M. J., Qiu, J. and White, K. (2001). The neuron-specific RNA-binding protein ELAV regulates neuroglian alternative splicing in neurons and binds directly to its pre-mRNA. *Genes Dev* **15**, 2546–2561.

Liu, D., Brockman, J. M., Dass, B., Hutchins, L. N., Singh, P., McCarrey, J. R., MacDonald, C. C. and Graber, J. H. (2007). Systematic variation in mRNA 3'-processing signals during mouse spermatogenesis. *Nucleic Acids Res* **35**, 234–246.

Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., Jia, G., Yu, M., Lu, Z., Deng, X., et al. (2014). A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat. Chem. Biol.* **10**, 93–95.

Loedige, I., Stotz, M., Qamar, S., Kramer, K., Hennig, J., Schubert, T., Loffler, P., Langst, G., Merkl, R., Urlaub, H., et al. (2014). The NHL domain of BRAT is an RNA-binding domain that directly contacts the hunchback mRNA for regulation. *Genes Dev* **28**, 749–764.

Love, M. I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550.

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A., et al. (2005). MicroRNA expression profiles classify human cancers. *Nature* **435**, 834–838.

Luo, W. and Sehgal, A. (2012). Regulation of Circadian Behavioral Output via a MicroRNA-JAK/STAT Circuit. *Cell* **148**, 765–779.

MacDonald, C. C. and McMahon, K. W. (2010). Tissue-specific mechanisms of alternative polyadenylation: testis, brain, and beyond. *WIREs RNA* **1**, 494–501.

MacDonald, C. C. and Redondo, J.-L. (2002). Reexamining the polyadenylation signal: were we wrong about AAUAAA? *Mol. Cell. Endocrinol.* **190**, 1–8.

Manak, J. R., Dike, S., Sementchenko, V., Kapranov, P., Biemar, F., Long, J., Cheng, J., Bell, I., Ghosh, S., Piccolboni, A., et al. (2006). Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nat Genet* **38**, 1151–1158.

Mangone, M., Manoharan, A. P., Thierry-Mieg, D., Thierry-Mieg, J., Han, T., Mackowiak, S. D., Mis, E., Zegar, C., Gutwein, M. R., Khivansara, V., et al. (2010). The landscape of C. elegans 3'UTRs. *Science* **329**, 432–435.

Manley, J. L. (1983). Accurate and specific polyadenylation of mRNA precursors in a soluble whole-cell lysate. *Cell* **33**, 595–605.

Martin, G., Gruber, A. R., Keller, W. and Zavolan, M. (2012). Genome-wide analysis of pre-mRNA 3“ end processing reveals a decisive role of human cleavage factor I in the regulation of 3” UTR length. *Cell Rep* **1**, 753–763.

Martin, R., Smibert, P., Yalcin, A., Tyler, D. M., Schaefer, U., Tuschl, T. and Lai, E. C. (2009). A Drosophila pasha mutant distinguishes the canonical miRNA and mirtron pathways. *Mol Cell Biol* **29**, 861–870.

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

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

Masuda, A., Takeda, J.-I. and Ohno, K. (2016). FUS-mediated regulation of alternative RNA processing in neurons: insights from global transcriptome analysis. *WIREs RNA* **7**, 330–340.

Masuda, A., Takeda, J.-I., Okuno, T., Okamoto, T., Ohkawara, B., Ito, M., Ishigaki, S., Sobue, G. and Ohno, K. (2015). Position-specific binding of FUS to nascent RNA regulates mRNA length. *Genes Dev* **29**, 1045–1057.

Mathews, M. B., Osbron, M. and Lingrel, J. B. (1971). Translation of globin messenger RNA in a heterologous cell-free system. *Nature New Biol.* **233**, 206–209.

Matoulkova, E., Michalova, E., Vojtesek, B. and Hrstka, R. (2014). The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA Biol* **9**, 563–576.

Matthews, B. B., Santos, Dos, G., Crosby, M. A., Emmert, D. B., St Pierre, S. E., Gramates, L. S., Zhou, P., Schroeder, A. J., Falls, K., Strelets, V., et al. (2015). Gene Model Annotations for Drosophila melanogaster: Impact of High-Throughput Data. *G3 (Bethesda)* **5**, 1721–1736.

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.

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

Mifflin, R. C. and Kellems, R. E. (1991). Coupled transcription-polyadenylation in a cell-free system. *J Biol Chem* **266**, 19593–19598.

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

Miura, P., Sanfilippo, P., Shenker, S. and Lai, E. C. (2014). Alternative polyadenylation in the nervous system: to what lengths will 3' UTR extensions take us? *Bioessays* **36**, 766–777.

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

modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., Landolin, J. M., Bristow, C. A., Ma, L., et al. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. **330**, 1787–1797.

Mohammed, J., Siepel, A. and Lai, E. C. (2014). Diverse modes of evolutionary emergence and flux of conserved microRNA clusters. *RNA* **20**, 1850–1863.

Molinie, B., Wang, J., Lim, K. S., Hillebrand, R., Lu, Z.-X., Van Wittenberghe, N., Howard, B. D., Daneshvar, K., Mullen, A. C., Dedon, P., et al. (2016). m(6)A-LAIC-seq reveals the census and complexity of the m(6)A epitranscriptome. *Nat Methods* **13**, 692–698.

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

Morgan, T. H. (1910). Sex limited inheritance in *Drosophila*. **32**, 120–122.

Mount, S. M. and Salz, H. K. (2000). Pre-messenger RNA processing factors in the Drosophila genome. *J. Cell Biol.* **150**, F37–44.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., et al. (2000). A whole-genome assembly of Drosophila. **287**, 2196–2204.

Naftelberg, S., Schor, I. E., Ast, G. and Kornblihtt, A. R. (2015). Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annu Rev Biochem* **84**, 165–198.

Nagaike, T., Logan, C., Hotta, I., Rozenblatt-Rosen, O., Meyerson, M. and Manley, J. L. (2011). Transcriptional Activators Enhance Polyadenylation of mRNA Precursors. *Mol Cell* **41**, 409–418.

Nazim, M., Masuda, A., Rahman, M. A., Nasrin, F., Takeda, J.-I., Ohe, K., Ohkawara, B., Ito, M. and Ohno, K. (2017). Competitive regulation of alternative splicing and alternative polyadenylation by hnRNP H and CstF64 determines acetylcholinesterase isoforms. *Nucleic Acids Res* **45**, 1455–1468.

Neumüller, R. A., Betschinger, J., Fischer, A., Bushati, N., Poernbacher, I., Mechtler, K., Cohen, S. M. and Knoblich, J. A. (2008). Mei-P26 regulates microRNAs and cell growth in the Drosophila ovarian stem cell lineage. *Nature* **454**, 241–245.

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.

O'Neill, E. M., Rebay, I., Tjian, R. and Rubin, G. M. (1994). The activities of two Ets-related transcription factors required for Drosophila eye development are modulated by the Ras/MAPK pathway. *Cell* **78**, 137–147.

Oktaba, K., Zhang, W., Lotz, T. S., Jun, D. J., Lemke, S. B., Ng, S. P., Esposito, E., Levine, M. and Hilgers, V. (2015). ELAV links paused Pol II to alternative polyadenylation in the Drosophila nervous system. *Mol Cell* **57**, 341–348.

Oliver, B. (2006). Tiling DNA microarrays for fly genome cartography. *Nat Genet* **38**, 1101–1102.

Oshlack, A., Robinson, M. D. and Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biol* **11**, 220.

Ozair, M. Z., Kintner, C. and Brivanlou, A. H. (2013). Neural induction and early patterning in vertebrates. *Wiley interdisciplinary reviews. Developmental biology* **2**, 479–498.

Ozsolak, F., Kapranov, P., Foissac, S., Kim, S. W., Fishilevich, E., Monaghan, A. P., John, B. and Milos, P. M. (2010). Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143**, 1018–1029.

Page, S. L., McKim, K. S., Deneen, B., Van Hook, T. L. and Hawley, R. S. (2000). Genetic studies of mei-P26 reveal a link between the processes that control germ cell proliferation in both sexes and those that control meiotic exchange in Drosophila. *Genetics* **155**, 1757–1772.

Pelechano, V., Wilkening, S., Jarvelin, A. I., Tekkedil, M. M. and Steinmetz, L. M. (2012). Genome-wide polyadenylation site mapping. *Methods in enzymology* **513**, 271–296.

Peng, S. S., Chen, C. Y., Xu, N. and Shyu, A. B. (1998). RNA stabilization by the AU-rich element binding protein, HuR, an ELAV protein. *EMBO J* **17**, 3461–3470.

Pinto, P. A. B., Henriques, T., Freitas, M. O., Martins, T., Domingues, R. G., Wyrzykowska, P. S., Coelho, P. A., Carmo, A. M., Sunkel, C. E., Proudfoot, N. J., et al. (2011). RNA polymerase II kinetics in polo polyadenylation signal selection. *EMBO J* 1–14.

Prescott, J. and Falck-Pedersen, E. (1994). Sequence elements upstream of the 3' cleavage site confer substrate strength to the adenovirus L1 and L3 polyadenylation sites. *Mol Cell Biol* **14**, 4682–4693.

Proudfoot, N. (2016). Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science* **352**.

Proudfoot, N. J. (1976). Sequence analysis of the 3' non-coding regions of rabbit alpha- and beta-globin messenger RNAs. *J. Mol. Biol.* **107**, 491–525.

Proudfoot, N. J. (2011). Ending the message: poly(A) signals then and now. *Genes Dev* **25**, 1770–1782.

Proudfoot, N. J. and Brownlee, G. G. (1976). 3' non-coding region sequences in eukaryotic messenger RNA. *Nature* **263**, 211–214.

Proudfoot, N. J. and Longley, J. I. (1976). The 3' terminal sequences of human alpha and beta globin messenger RNAs: comparison with rabbit globin messenger RNA. *Cell* **9**, 733–746.

Proudfoot, N. J., Cheng, C. C. and Brownlee, G. G. (1976). Sequence analysis of eukaryotic mRNA. *Prog. Nucleic Acid Res. Mol. Biol.* **19**, 123–134.

Retelska, D., Iseli, C., Bucher, P., Jongeneel, C. V. and Naef, F. (2006). Similarities and differences of polyadenylation signals in human and fly. *BMC Genomics* **7**, 176.

Rhinn, H., Qiang, L., Yamashita, T., Rhee, D., Zolin, A., Vanti, W. and Abeliovich, A. (2012). Alternative alpha-synuclein transcript usage as a convergent mechanism in Parkinson's disease pathology. *Nat Commun* **3**, 1084.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47–e47.

Robinow, S. and White, K. (1991). Characterization and spatial distribution of the ELAV protein during Drosophila melanogaster development. *J. Neurobiol.* **22**, 443–461.

Robinow, S., Campos, A. R., Yao, K. M. and White, K. (1988). The elav gene product of Drosophila, required in neurons, has three RNP consensus motifs. **242**, 1570–1572.

Rot, G., Wang, Z., Huppertz, I., Modic, M., Lenče, T., Hallegger, M., Haberman, N., Curk, T., Mering, von, C. and Ule, J. (2017). High-Resolution RNA Maps Suggest Common Principles of Splicing and Polyadenylation Regulation by TDP-43. *Cell Rep* **19**, 1056–1067.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., et al. (2000). Comparative genomics of the eukaryotes. **287**, 2204–2215.

Rüegsegger, U., Beyer, K. and Keller, W. (1996). Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors. *J Biol Chem* **271**, 6107–6113.

Ryan, K., Calvo, O. and Manley, J. L. (2004). Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. *RNA* **10**, 565–573.

Salisbury, J., Hutchison, K. W. and Graber, J. H. (2006). A multispecies comparison of the metazoan 3'-processing downstream elements and the CstF-64 RNA recognition motif. *BMC Genomics* **7**, 55.

Samson, M.-L. and Chalvet, F. (2003). found in neurons, a third member of the Drosophila elav gene family, encodes a neuronal protein and interacts with elav. *Mech. Dev.* **120**, 373–383.

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. **320**, 1643–1647.

Sanfilippo, P., Smibert, P., Duan, H. and Lai, E. C. (2016). Neural specificity of the RNA-binding protein Elav is achieved by post-transcriptional repression in non-neural tissues. *Development* **143**, 4474–4485.

Schönemann, L., Kühn, U., Martin, G., Schäfer, P., Gruber, A. R., Keller, W., Zavolan, M. and Wahle, E. (2014). Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. *Genes Dev* **28**, 2381–2393.

Scotto-Lavino, E., Du, G. and Frohman, M. A. (2006). 3' end cDNA amplification using classic RACE. *Nat Protoc* **1**, 2742–2745.

Setzer, D. R., McGrogan, M. and Schimke, R. T. (1982). Nucleotide sequence surrounding multiple polyadenylation sites in the mouse dihydrofolate reductase gene. *J Biol Chem* **257**, 5143–5147.

Shcherbata, H. R., Ward, E. J., Fischer, K. A., Yu, J. Y., Reynolds, S. H., Chen, C. H., Xu, P., Hay, B. A. and Ruohola-Baker, H. (2007). Stage-specific differences in the requirements for germline stem cell maintenance in the Drosophila ovary. *Cell stem cell* **1**, 698–709.

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

Shenker, S., Miura, P., Sanfilippo, P. and Lai, E. C. (2015). IsoSCM: improved and alternative 3' UTR annotation using multiple change-point inference. *RNA* **21**, 14–27.

Shenoy, A. and Blelloch, R. H. (2014). Regulation of microRNA function in somatic stem cell proliferation and differentiation. *Nat Rev Mol Cell Biol* **15**, 565–576.

Shepard, P. J., Choi, E.-A., Lu, J., Flanagan, L. A., Hertel, K. J. and Shi, Y. (2011). Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**, 761–772.

Shepherd, A., Wesley, U. and Wesley, C. (2010). Notch and delta mRNAs in early-stage and mid-stage drosophila embryos exhibit complementary patterns of protein-producing potentials. *Dev. Dyn.* **239**, 1220–1233.

Shi, Y. (2012). Alternative polyadenylation: new insights from global analyses. *RNA* **18**, 2105–2117.

Shi, Y. and Manley, J. L. (2015). The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site. *Genes Dev* **29**, 889–897.

Shi, Y., Di Giammartino, D. C., Taylor, D., Sarkeshik, A., Rice, W. J., Yates, J. R., Frank, J. and Manley, J. L. (2009). Molecular architecture of the human pre-mRNA 3' processing complex. *Mol Cell* **33**, 365–376.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034–1050.

Simionato, E., Barrios, N., Duloquin, L., Boissonneau, E., Lecorre, P. and Agnès, F. (2007). The Drosophila RNA-binding protein ELAV is required for commissural axon midline crossing via control of commissureless mRNA expression in neurons. *Dev Biol* **301**, 166–177.

Smibert, P., Bejarano, F., Wang, D., Garaulet, D. L., Yang, J. S., Martin, R., Bortolamiol-Becet, D., Robine, N., Hiesinger, P. R. and Lai, E. C. (2011). A Drosophila genetic screen yields allelic series of core microRNA biogenesis factors and reveals post-developmental roles for microRNAs. *RNA* **17**, 1997–2010.

Smibert, P., Miura, P., Westholm, J. O., Shenker, S., May, G., Duff, M. O., Zhang, D., Eads, B. D., Carlson, J., Brown, J. B., et al. (2012). Global patterns of tissue-specific alternative polyadenylation in Drosophila. *Cell Rep* **1**, 277–289.

Smibert, P., Yang, J.-S., Azzam, G., Liu, J.-L. and Lai, E. C. (2013). Homeostatic control of Argonaute stability by microRNA availability. *Nat Struct Mol Biol* **20**, 789–795.

Soller, M. (2003). ELAV inhibits 3'-end processing to promote neural splicing of ewg pre-mRNA. *Genes Dev* **17**, 2526–2538.

Spies, N., Burge, C. B. and Bartel, D. P. (2013). 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res* **23**, 2078–2090.

Stark, A., Brennecke, J., Bushati, N., Russell, R. B. and Cohen, S. M. (2005). Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell* **123**, 1133–1146.

Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Hubbard, T. J., Guigó, R., Harrow, J., Bertone, P.RGASP Consortium (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* **10**, 1177–1184.

Sun, K. and Lai, E. C. (2013). Adult-specific functions of animal microRNAs. *Nat Rev Genet* **14**, 535–548.

Sun, K., Jee, D., de Navas, L. F., Duan, H. and Lai, E. C. (2015). Multiple in vivo biological processes are mediated by functionally redundant activities ofDrosophila mir-279andmir-996. *PLoS Genet* **11**, e1005245.

Sun, Y., Fu, Y., Li, Y. and Xu, A. (2012). Genome-wide alternative polyadenylation in animals: insights from high-throughput technologies. *J Mol Cell Biol* **4**, 352–361.

Tadros, W. and Lipshitz, H. D. (2009). The maternal-to-zygotic transition: a play in two acts. *Development* **136**, 3033–3042.

Takagaki, Y., Manley, J. L., MacDonald, C. C., Wilusz, J. and Shenk, T. (1990). A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs. *Genes Dev* **4**, 2112–2120.

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., 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.

Takahashi, H., Kato, S., Murata, M. and Carninci, P. (2012). CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. *Methods Mol Biol* **786**, 181–200.

Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics* **14**, 178–192.

Tian, B. and Graber, J. H. (2012). Signals for pre-mRNA cleavage and polyadenylation. *WIREs RNA* **3**, 385–396.

Tian, B. and Manley, J. L. (2017). Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18**, 18–30.

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 Res* **33**, 201–212.

Toba, G., Qui, J., Koushika, S. P. and White, K. (2002). Ectopic expression of Drosophila ELAV and human HuD in Drosophila wing disc cells reveals functional distinctions and similarities. *Journal of Cell Science* **115**, 2413–2421.

Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., Fischer, C. M., Gibson, D., Gonzalez, J. N., Guruvadoo, L., et al. (2017). The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res* **45**, D626–D634.

Ulitsky, I., Shkumatava, A., Jan, C. H., Subtelny, A. O., Koppstein, D., Bell, G. W., Sive, H. and Bartel, D. P. (2012). Extensive alternative polyadenylation during zebrafish development. *Genome Res* **22**, 2054–2066.

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.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. **291**, 1304–1351.

Verma, I. M., Temple, G. F., Fan, H. and Baltimore, D. (1972). In vitro synthesis of DNA complementary to rabbit reticulocyte 10S RNA. *Nature New Biol.* **235**, 163–167.

Wang, E. T., Cody, N. A. L., Jog, S., Biancolella, M., Wang, T. T., Treacy, D. J., Luo, S., Schroth, G. P., Housman, D. E., Reddy, S., et al. (2012). Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell* **150**, 710–724.

Wang, Y., Medvid, R., Melton, C., Jaenisch, R. and Blelloch, R. (2007). DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. *Nat Genet* **39**, 380–385.

Wen, J., Mohammed, J., Bortolamiol-Becet, D., Tsai, H., Robine, N., Westholm, J. O., Ladewig, E., Dai, Q., Okamura, K., Flynt, A. S., et al. (2014). Diversity of miRNAs, siRNAs, and piRNAs across 25 Drosophila cell lines. *Genome Res* **24**, 1236–1250.

Westholm, J. O., Miura, P., Olson, S., Shenker, S., Joseph, B., Sanfilippo, P., Celniker, S. E., Graveley, B. R. and Lai, E. C. (2014). Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Rep* **9**, 1966–1980.

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

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

Wu, T. D. and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881.

Yan, J. and Marr, T. G. (2005). Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Res* **15**, 369–375.

Yang, Q. and Doublié, S. (2011). Structural biology of poly(A) site definition. *WIREs RNA* **2**, 732–747.

Yao, C., Choi, E.-A., Weng, L., Xie, X., Wan, J., Xing, Y., Moresco, J. J., Tu, P. G., Yates, J. R. and Shi, Y. (2013). Overlapping and distinct functions of CstF64 and CstF64τ in mammalian mRNA 3' processing. *RNA* **19**, 1781–1790.

Yao, K. M. and White, K. (1994). Neural specificity of elav expression: defining a Drosophila promoter for directing expression to the nervous system. *Journal of Neurochemistry* **63**, 41–51.

Yao, K. M., Samson, M. L., Reeves, R. and White, K. (1993). Gene elav of Drosophila melanogaster: a prototype for neuronal-specific RNA binding protein gene family that is conserved in flies and humans. *J. Neurobiol.* **24**, 723–739.

Yudin, D., Hanz, S., Yoo, S., Iavnilovitch, E., Willis, D., Gradus, T., Vuppalanchi, D., Segal-Ruder, Y., Ben-Yaakov, K., Hieda, M., et al. (2008). Localized regulation of axonal RanGTPase controls retrograde injury signaling in peripheral nerve. *Neuron* **59**, 241–252.

Zaharieva, E., Haussmann, I. U., Bräuer, U. and Soller, M. (2015). Concentration and Localization of Coexpressed ELAV/Hu Proteins Control Specificity of mRNA Processing. *Mol Cell Biol* **35**, 3104–3115.

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 Res* **31**, 1375–1386.

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

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

Zheng, D. and Tian, B. (2014). RNA-binding proteins in regulation of alternative cleavage and polyadenylation. *Adv. Exp. Med. Biol.* **825**, 97–127.

Zhu, H., Zhou, H.-L., Hasman, R. A. and Lou, H. (2007). Hu proteins regulate polyadenylation by blocking sites containing U-rich sequences. *J Biol Chem* **282**, 2203–2210.