The Landscape of *C. elegans* 3'UTRs

by

Arun Prasad Manoharan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2011

Doctoral Committee:

Assistant Professor John Kim, Chair
Professor Margit Burmeister
Professor Daniel M.Burns Jr.
Assistant Professor Jun Li
Assistant Professor Maureen A. Sartor

Dedication

To my parents,

Manoharan and Sellvi

Sister,

Sridevi.

Your constant

encouragement, patience, sacrifice, prayers and love

allowed me to selfishly

pursue

my dreams across the world

I could not be blessed with a better family

Acknowledgements

I would first like to thank my advisor Dr. John Kim who was brave enough to allow a computer science graduate with no previous biology experience to perform research in his lab. I am indebted to his patience, constant motivation and high standards, which have led me to succeed in my career. I also want to thank my collaborators Prof. Kristin Gunsalus and Prof. Fabio Piano (NYU) for their support and advice throughout the length of my graduate career. I would forever cherish the time I spent at their lab at NYU during the collaboration. My special thanks to fellow researcher, friend and now assistant professor, Marco Mangone (NYU) with whom, I spent so many months, day and night working on the 3'UTRome project. I would also like to thank Jean and Danielle Thierry-Mieg (NIH) who helped validate many of my analysis and for their help in painstakingly annotating thousands of gene models and 3'UTRs used in our analysis and consolidating the data from various sources.

I would also like to thank Prof. Daniel Burns and Prof. Margit Burmeister from the Bioinformatics department who were with me every step of the way and were phenomenal in helping me through the initial stages of my graduate career to get accustomed to the vast and interesting world of biological research. Without their many hours of advice and encouragement my transition to bioinformatics from computer science would not have been possible. I would also like to thank my other committee members Prof. Jun Li and Prof. Maureen Sartor who were always there to guide me and ask interesting questions during my meetings. I would also like to specially thank Ms.

Julia Eussen of the Bioinformatics department for her expert administrative support and friendly smile and discussions, which made my student life a lot easier. I would also like to thank Jeff de Wet from the bioinformatics department for his introductory bioinformatics classes and friendly disposition, which made me feel right at home. I would also like to thank Prof. Gil Omenn and Prof. Brian Athey of the Bioinformatics department who always had time to talk to students in spite of their busy schedules.

I also want to specially thank my fellow lab mate Ting Han for his healthy scientific discussions and biological insights, fruitful collaborations and occasional lessons in Mandarin and trips to the local Chinese restaurants. I have no doubt that he will be a successful faculty in the future. I am also thankful to have spent time with fellow lab members Vishal Khivansara, Allison Billi, Amanda Day, Amelia Alessi, Mallory Freeberg and Danny Yang and past lab members Matthew Avenarius, Dongping Wei and rotation student Cheryl Jacobs who made the past five years spent in the lab very enjoyable. I would also like to thank bay mate Da Fang for her tasty home cooked meals.

No acknowledgement would be complete without thanking my friends Huy Vuong, Yongsheng Huang, Patrick Harrington, Ryan Welch, Gang Su, Yu-hsuan Lin, Conner Sandefur, Lucy Huang, Avinash Shanmugam, Keerthi Sannareddy, Tanu Soni and Sunit Jain from the Bioinformatics department and Katie Dumas, Liz Adams, Taylor Eves, Lily Harizanova and Jon Whicher from the LSI who were always there to help me and to drag me out of lab to have a healthy social life.

I would also like to thank the beautiful city of Ann Arbor for being my second home and never making me feel homesick and showing beautiful summers and fall and letting me meet so many interesting friends who have forever changed my outlook on life.

Finally I would like to thank my family and all my relatives back in India for their love, sacrifice, patience, understanding and constant encouragement which has helped get through these five long years away from home and my home country.

Table of Contents

List of Figures

List of Tables

Abstract

The 3' untranslated regions (3'UTRs) of mRNAs have recently been shown to play major roles in gene regulation by interaction with small (21-26nt length) RNAs, such as microRNAs and small interfering RNAs (siRNAs). Recent studies highlight dynamic expression of small RNAs and provide limited evidence of utilization of alternative 3'UTR lengths (3'UTR isoforms) across development and specific cell types. However, a comprehensive catalogue of the 3'UTRome of an organism has been unavailable. By computational analysis of traditional and novel high-throughput sequence data, chapter two of this dissertation provides annotated 3'UTRs for more than 75% of the genes in the model organism *Caenorhabditis elegans* across the major stages of development. At the whole-transcriptome level, 3'UTRs express remarkable diversity in utilization of alternative poly-adenylation sites, which define the alternative 3'UTR isoforms, in ~40% of genes in the genome. We identified many isoforms that are developmental stage-specific, and many genes show evidence of length switching between longer and shorter isoforms over development, the goal of which may be to include or escape regulation from small RNAs or other RNA binding proteins. Thus, our study reveals the diversity and temporally regulated expression of 3'UTR isoforms as a complex mechanism in gene regulation at an unprecedented scope. The analysis provides large-scale evidence for multiple alternative 6nt sequence elements (PAS sites) near 3' ends of 3'UTRs that are enriched in shorter, alternative isoforms. Chapter three of my dissertation compares and combines the results of a parallel study of 3'UTRs in *C. elegans* further expanding transcriptome coverage of 3'UTRs. As an example of biological relevance for 3'UTR

isoform usage, in chapter four we identified a potential connection between the synaptogenesis pathway and specific 3’UTR isoform usage in *C. elegans*.

In related work described in chapter five, I analyzed small RNA sequences from isolated sperm and oocytes and identified and characterized a new class of germline-specific siRNAs, 26G-RNAs, which target coding regions and 3’UTRs of genes to regulate their target gene expression. These have been classified into two subclasses: 26G-RNAs generated in the male germline targeting genes involved in spermatogenesis, and maternally inherited 26G-RNAs targeting genes that function in zygotic development.

Chapter 1: Introduction

1.1: Introduction

Successful transcription results in transfer of information from the DNA to the RNA domain namely to messenger RNAs. In eukaryotes precursor messenger RNAs (pre-mRNA) are transcribed in the nucleus by RNA polymerase II (pol II) enzyme and the nascent mRNA is transported to the cytoplasm from the nucleus to be translated into protein. However before the transport the mRNA must undergo multiple posttranscriptional modifications including 5’ capping, 3’ cleavage and polyadenylation and alternative splicing. 5’ capping is an addition process where 7-methylguanosine (m7G) is added to the 5’end of the mRNA [1]. This protects the exported mRNA from the destructive nature of degrading exonuclease enzymes present in the cytoplasm [2, 3] and helps in recruiting the ribosomes to the mRNA aiding in translational initiation[4]. 3’end cleavage and polyadenylation is the process by which polyadenylation signals (PAS) are recognized in the 3’end of the mRNA resulting in cleavage of the 3’end followed by addition of long adenine (polyA) tails at the cleaved end [5] and release of the transcribing RNA polymerase. This polyA tail is thought to add stability to the mRNA and aid in transport and translation [6-8]. When a mRNA has more than one signal site marking the 3’ends, there can be more than one polyadenylation site, resulting in alternative polyadenylation [9]. RNA splicing is another post-transcriptional modification process where intermediate regions of the mRNA called introns are excised out from the mRNA leaving functional regions behind called exons [4, 10]. This process is important in deciding which regions of the mRNA are to be included in the final protein template. Alternate splicing and polyadenylation can result in different proteins from the same gene and/or different cellular localization or protein output. The mature mRNA contains the 5’cap, the spliced exons and the 3’polyA tail which is then exported to the cytoplasm. The mechanisms which regulate this process are termed post-transcriptional regulation and mainly occur through proteins that interact with the RNA namely RNA Binding Proteins (RBP). The proteins in the Cap Binding Complex (CBC) bind the 5’ cap and help in exporting the RNA to the cytoplasm and protect against degradation by harmful exonucleases that exist there [2, 3]. The factors in the spliceosome complex successfully remove the intron regions in the mRNA which do not code proteins[11]. The proteins in the 3’end processing machinery successfully recognize the polyA signal sites and cleave at the 3’end followed by recruitment of polyA polymerases to add the polyA tail. The polyA tail provides stability to the mRNA during export and also protects against exonucleases. Furthermore, the PolyA binding proteins (PABP) bind to the polyA tail and promote translation through their interaction with the translation initiation factor [12, 13]. Decapping or deadenylation of the polyA tail results in the degradation of the mRNA in the cytoplasm, effectively controlling final protein output even though the RNA output of transcription remains unchanged. Recently, small RNAs such as microRNAs have been shown to promote deadenylation of the polyA tail and translational inhibition [14-19]. Similarly, alternative splicing determines what exons are included or excluded in the exported mRNA, and based on the combination, the structure and function of the protein is modulated [13].

In this thesis we are focusing on the 3’end processing of mRNA namely cleavage and polyadenylation. 3’end processing is crucial for the transport of the mRNA into the cytoplasm, and preventing polyadenylation has been shown to result in decreased cytoplasmic mRNA and hence reduction in protein levels. [20, 21]. A variety of diseases including thalassemia, thrombophilia, IPEX syndrome, oculopharyngeal muscular dystrophy, cancer, and idiopathic hyperosinophilic syndrome occur due to disruption of 3'end processing [22]. The polyA binding protein (PABP), which binds to the polyA tail, has been shown to protect the mRNA against degradation in *Xenopus* oocytes. [23] Further the PABP has been shown to interact with the 5' cap of the mRNA and plays a role in translation [8, 24]. Finally the 3'end processing machinery has been shown to dynamically interact with the transcription machinery including the C-Terminal Domain of RNA polymerase II. [13, 25]

1.2: Mechanisms of 3'end processing

1.2.1: Sequence elements defining 3'end processing

3' end processing in eukaryotes occurs as a concerted reaction between the various RNA binding proteins (RBP) and specific sequence elements in the mRNA. Disruption of these sequence sites affects the 3'end processing. In eukaryotes there are three major sequence elements that define the 3'end of the mRNA.

PAS: The primary sequence element is the Polyadenylation Signal (PAS). It is a six nucleotide sequence element canonically represented by AAUAAA. An early study of human Expressed Sequence Tags (EST) found that 75% of the mRNAs contain the PAS signal. [26]. The second most abundant PAS signal in humans is AUUAAA[5]. Mutations in the PAS signal have been shown to have biological consequences resulting in thalassemia [27] and reduced processing of pre-mRNA in *Xenopus [28]*. In addition to the sequence of PAS its position in the mRNA is also important and occurs within 10-30 nt upstream of the processed 3'end cleavage site. Modifying this distance results in a new cleavage site maintaining the initial distance [29].

Cleavage site: The cleavage site marks the position where the mRNA is cleaved for addition of a polyA tail. The cleavage site is between the PAS and the DSE and predominantly a CA dinucleotide. The cleavage occurs between the cytosine and adenine [30].

Downstream Sequence Element (DSE): This refers to a U or GU rich region downstream of the 3'end of the mRNA. It is usually within 30 nt downstream of the cleavage site. While the sequence of the DSE is not as conserved as the PAS and can tolerate more point mutations [31], modifying the distance of the DSE to the cleavage site greatly affects 3'end processing efficiency [32]

Auxiliary elements: In addition to the major sequence elements discussed above there are auxiliary sequence elements either upstream or downstream of the cleavage site that aid in 3'end processing by recruiting proteins . Upstream auxiliary elements are generally U-rich and UGUA is a common upstream auxiliary element. It has been shown to aid in non-canonical 3'end processing [33]. A G-rich downstream auxiliary element has been identified, however neither its position nor sequence is conserved [34-36].

1.2.2: Proteins involved in 3'end processing

3'end processing in eukaryotes is a two-step process involving cleavage of the mRNA at the cleavage site and addition of the polyA tail at the cleaved end. The factors involved in effecting these processes have been studied in many organisms including yeast and mammals [37] using traditional biochemical assays of nuclear extracts and by modern approaches using mass spectrometry. These studies indicate a multi-subunit protein complex involved in the 3'end processing.

Mammalian 3'end processing: The major players in mammalian 3'end processing are cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor

(CSTF), cleavage factor I (CFIm), cleavage factor II (CFIIm), polyA polymerase (PAP), polyA binding protein (PABP), symplekin and pol II. [22, 38, 39] CPSF consists of at least five subunits, Cpsf160, Cpsf100, Cpsf73, Cpsf30 (numbers representing protein weight in KDa) and hFip1. CSTF consists of three subunits, Cstf77, Cstf64 and Cstf50. CFIm consists of at least three subunits CFIm72, CFIm68, CFIm59 and two CFIm25 subunits. CFIIm consists of two subunits, Pcf11 and Clp1. Additional proteins such as Pfs2p and PPI also play a role in 3'end processing. These subunits interact with various sequence elements in the mRNA to process the 3'ends. CPSF binds directly to the AAUAAA PAS upstream of the cleavage site in the mRNA [40] while the CSTF recognizes the U/GU rich DSE downstream of the cleavage site [41-43]. 3' end processing in the absence of the canonical PAS has been shown to depend on the upstream auxiliary element UGUA and its interaction with the CFIm [33].

Cpsf160 binds directly to the AAUAAA PAS element in the mRNA [40]. Mutating the PAS sequence abolishes this interaction [44]. The binding efficiency of Cpsf160 to the PAS also depends on interaction with Cstf64 mediated by Cstf77 [45]. It also interacts with transcription factor TFIID and the C-Terminal Domain of pol II and plays roles in transcriptional initiation and elongation [46, 47]. The cleavage at the CA cleavage site is catalyzed by binding of Cpsf73 to the cleavage site in a PAS dependent manner suggesting a positioning role for PAS [48] and could be mediated through Cpsf160 and Cstf64. hFip1 mediates interaction with PAP and may be involved in bringing the PAP close to the cleavage site. It also interacts with Cpsf30 and in turn is required for interaction between Cpsf160 and Cstf77 [49].

Cstf64 contains a RNA binding domain and binds to the G/U rich DSE [41-43]. Cstf64 also interacts with Cstf77 and symplekin. In humans there is a second isoform called tau-Cstf64 which is expressed in male germ cells and may play a role in germ cell specific polyadenylation [50-53]. Cstf77 is required for 3'end processing specificity. Mutation of Cstf77 in *Drosophila* results in usage of alternative polyA sites [54]. Cstf50 is required for cleavage in vitro and interacts with Cstf77 and binds to the CTD of Pol II[47].

CFIm is required for cleavage *in vitro* and the primary function of CFIm may be to aid in recognition of the pre-mRNA substrate. In the absence of a polyadenylation signal, 3'ends are recognized by the presence of a UGUA signal, which is recognized by Cleavage Factors I (CFIm) [33, 55]. The CFIm is known to be a tetrameric complex consisting of two CFIm25 and one CFIm59 and one CFIm68 [56]. Recent work including the crystal structure of the complex suggests an important role in selecting polyA sites during alternative polyadenylation [57-59]. The model proposes a looping of the RNA between two UGUA sites and based on the combination of the UGUA sites chosen, different 3'ends can be selected.

CFIIm has two subunits, hPcF11 and hClp1. Pcf11 contains a Pol II CTD interacting domain and mutation in this domain has been shown to cause termination defects. It is speculated that it is necessary for release of 3'end processing factors from Pol II [60]. Clp1 is highly conserved and its mutation abolishes cleavage but not polyadenylation [61].

PAP catalyzes the addition of the polyA tail at the cleavage site and interacts with the polyA binding protein (PABP). PABP stimulates PAP to add the polyA tail at the cleavage site and also controls the polyA tail length [62]. PABP binds to stretches of 11-14 adenines [63] and the binding continues until proper polyA tail length of ~200-300 As is achieved [64]. Symplekin interacts with Cstf64 and CPSF [62]. It may also function in 3'end processing of histone mRNAs [65]. RNA Polymerase II (Pol II) also plays an important role in 3'end processing. Truncation in the CTD of Pol II results in defects in polyadenylation [47]. The CTD of Pol II also interacts with CPSF and CSTF and binds Cpsf160 and Cstf50. Additional proteins such as Pfs2p also play roles in 3'end processing.

Yeast 3'end processing: The yeast 3'end processing machinery contains the cleavage and polyadenylation factor (CPF), cleavage factor IA (CFIA) and cleavage factor IB (CFIB) [38]. CPF is homologous to mammalian CPSF but it is further subdivided into CFII and PFI. The factors involved in 3'end processing in mammals and yeast are generally conserved, however, there are differences in the organization of the subunits and sequence elements recognized. There are three sequence elements in yeast mRNA, AU-rich efficiency element (EE), A-rich positioning element (PE) and U-rich upstream element (UUE) or downstream element (DUE). The cleavage site is defined by a pyrimidine followed by multiple adenosines (Y(A)n). Cpsf160 (Yhh1) does not bind to the A-rich PE element but to the A-rich cleavage site [66]. Similarly, Cstf64 (Rna15) that binds G/U rich DSE in mammals binds to the A-rich PE upstream of the cleavage site [67]. Yeast Cstf (CFI) is involved in both cleavage and polyadenylation, which is different from mammals where it is only involved in cleavage. Symplekin (Pta1) is part of CFII and acts as a scaffold. Cstf50 is missing in yeast and CFIA bears homology to mammalian CFIIm and Cstf. Pol II CTD is not necessary for 3'end processing in yeast [68]. Cleavage is mediated by CFIA, CFIB and CFII while CPF, CFIA, CFIB and Pap1 mediate polyadenylation. There are also additional factors such as Pfs2p, Ssu72, Mpe1, Glc7, Ref2 and Hrp1, which are not present in mammals. Yeast specific Hrp1 of CFIB functions similar to mammalian CFIm. Pfs2p interacts with Fip1 and mutating Pfs2p results in cell death [69] and affects cleavage and polyadenylation.

1.2.3: Histone gene 3’end processing

The histone genes seem to vary from other genes in eukaryotes in terms of their 3’end processing [3, 70-73]. Instead of a PAS based recruitment of factors, a conserved stem loop upstream of the cleavage site that is recognized by a Stem Loop Binding Protein (SLBP) and a purine rich Histone Downstream Element (HDE) within 100nt of the stop codon, signals the end of histone mRNAs. HDE is recognized by the 5’ end of U7 snRNA, which recruits the snRNP (ribonucleoprotein complex). Lsm11, Lsm10, FLASH and ZPF100 are also recruited in the complex [74, 75]. The cleavage site is a CA dinucleotide similar to other mRNAs. Recent studies have shown that the cleavage is still mediated by recruitment of CPSF73 and CPSF 100 and CPSF 73 is the cleavage factor [76]. The 3’ end processing of histone mRNAs is different from the other mRNAs since it is a one-step process and only depends on signal elements and is incompatible with splicing. Furthermore, histone mRNAs are generally non-polyadenylated and the binding of the SLBP to the stem loop provides stability and translational efficiency the same way as the polyA tail [71].

1.2.4: Polyadenylation in Operons

Polycistronic transcription is when multiple genes in a loci are transcribed into a single transcript and later the individual genes are spliced out by trans-splicing[77-79], similar to how introns are spliced out of pre-mRNA. It is commonly seen in *C. elegans*, flatworms, hydra and trypanosomes. .Such polycistronic transcription units are called operons and in *C. elegans* there are about 1000 operons and each operon contains 2-8 genes [79]. The trans-spliced mRNAs were seen to begin with a 22nt sequence, which was not seen in the gene sequence. Later this sequence was found to be derived from a SLRNA, which acts as a splice donor. Full length cDNA sequencing of these genes shows that the first gene in the operon is spliced by SL1 RNA while the remaining downstream genes are spliced by SL2-12 RNAs and the proteins of the snRNP[80]. After splicing, each individual mRNA is polyadenylated. This is especially interesting since regulation occurs post-transcription and the 3' end processing machinery regulates conversion from polycistronic to monocistronic transcripts. Most of these genes occur very close to each other with ~100nt between the polyA site of one gene and the 5' end of the next gene. This places steric constraints on the polyadenylation machinery and it is interesting to note the level of diversity in the sequence determinants and the high alternative polyadenylation events in such a confined space [81]. Work in trypanosomes has now shown a two-step mechanism where transcription and trans-splicing are uncoupled and delaying trans-splicing results in gene regulation [82]. A recent global study of trans-splicing in *C. elegans* showed that 70% of the genes in *C. elegans* are trans-spliced with either SL1 or SL2 and use different underlying mechanisms [83]. They also show that the usage of the SL1 or SL2 depends on the promoter utilized for generating either a polycistronic or monocistronic transcript. If the promoter of the previous gene is used then it results in a polycistronic transcript and if the intermediate promoter is used then it is a monocistronic transcript coupling promoter selection with trans-splicing. Another interesting study highlighted the advantages of polycistronic transcription in efficient utilization of transcriptional resources for increased upregulation especially during recovery [84].

1.3: Alternative polyadenylation- regulation of 3'UTR length

Alternative polyadenylation refers to the variability in the length of the 3'UTRs defined by 3'end processing machinery, such that for the same transcript there can be different 3'UTR lengths. The first evidence was seen in alternative processing of IgM mRNA during B cell differentiation [85-87]. It was seen that switching from the membrane bound form to the secreted form in plasma cells occurs due to alternative polyadenylation of the 3'UTRs and the mechanism was regulated by concentration of CSTF 64. Genome wide studies now show that this mechanism occurs in many organisms including *C. elegans* [81], *Arabidopsis* [88], rice [89], *Chlamydomonas* [90], trypanosomes [82], mouse [51, 91], humans [91-93] and yeast [94]. In these studies it has been shown that alternative polyadenylation is pervasive and affects 40-70% of the total number of genes in an organism. This shows that the alternative polyadenylation mechanism is more global, complex and widespread than previously appreciated and could be an important mechanism of gene regulation yet to be explored. Immediate effects of this varying 3'UTR length could be differential localization, protein function, stability, protein quantity and post-transcriptional regulation [9, 95, 96]. In many cases the alternative polyadenylation sites have valid PAS and they are evolutionarily conserved [81, 97].

Polyadenylation sequencing of individual tissues in humans including heart, brain, colon, liver, breast, lymph node, skeletal muscle, retina, testis, and uterus has shown that alternative polyadenylation produces tissue-specific 3'UTRs [93, 98]. Developmental and environmental cues also seem to employ alternative polyadenylation as seen in embryogenesis [99], differentiation of stem cells [100], development of neurons [101, 102], immune response [103] and spermatogenesis [50-53, 58]. These examples show that alternative polyadenylation is a highly regulated process that is controlled both spatially across different tissues and spatially across development. Furthermore, cancer studies show that activation of oncogenes can occur due to shortening of 3'UTRs [104] and high-throughput studies on cancer cells show wide spread evidence of alternative polyadenylation [105].

Even though we now see the prominence of this mechanism, we still do not know much about the mechanisms underlying this regulation. An example from Opitz syndrome shows that the tissue specific expression of the causal MID1 occurs due to interaction between the promoter region and polyadenylation signal linking alternative polyadenylation with transcription [106]. Few studies identify CFIm subunit of the 3'end processing machinery to be the agent for alternative polyadenylation [56-58]. Cstf64 subunit of CSTF has also been implied to play a role in alternative polyadenylation [107] and a recent study in induced pluripotency reports major polyA machinery genes regulated differently between differentiated and undifferentiated cells, especially in the CSTF [100].

The fact that both the miRNAs and siRNAs have target regions in the 3'UTR makes it an important region in the mRNA in terms of posttranscriptional regulation. Recent studies also highlight the role of 3'UTRs in localization of an mRNA by presence of sequence signals [108]. The interactions between the 3'UTR region of the mRNA and small RNAs such as microRNAs highlight the importance of this mechanism in post-transcriptional control of gene expression. The biological impact of the 3'UTRs comes from the fact that varying the length of the 3'UTR, i.e defining its 3'end, may in turn effect gene regulation by inclusion/exclusion of target sites of the small RNAs. The next section discusses the various small RNAs that interact with the 3'UTRs.

1.4: RNA interference and small RNAs

The past decade has seen a rapid change in the study of gene regulation mainly due to the discovery of RNA interference (RNAi). RNA interference is a mechanism in which small RNA molecules interact with the mRNA resulting in post-transcriptional gene regulation through degradation of the target molecule in the cytoplasm. Incidentally this mechanism had been reported during silencing of introduced transgenes where not only the introduced transgenes, but also the endogenous genes were silenced [109-115]. Initially named co-suppression, it was assumed to be a defense mechanism in the cell against viral infections as seen in plants [116, 117] or against repetitive DNA as seen in transgene studies. Studies have shown the effect to be triggered by RNA, introducing a new role for RNA in gene regulation besides transcription and translation. The application of this mechanism was shown by work on *C.elegans* by Fire and Mello in 1998. They showed that introduction of double stranded RNA (dsRNA) targeting a gene seems to mediate silencing of the gene. Injecting dsRNA is more potent than injecting single stranded RNA and the silencing was seen to be passed to future generations and only a few molecules of the dsRNA was sufficient to trigger target gene silencing. They called this phenomenon RNA interference. RNAi seems to operate in the cytoplasm suggesting that the target is processed mRNA, and the silencing occurs irrespective of the proximity of the silenced genes, suggesting a trans-acting mechanism. Fire and Mello also showed that the silencing can cross cell boundaries in *C.elegans* since the injection of the dsRNA in to the intestine can integrate into the germline and be passed onto future generations. The effect was also seen in insects [118, 119], trypanosomes[120], zebrafish [121] and mouse[122, 123] making it an effective tool to control gene expression across organisms. This triggered the whole new field of RNA mediating gene silencing, promising tremendous potential in genetics, medicine and disease control.

There are many questions that arise at the prospect of such a mechanism. Why does dsRNA, but not single stranded RNA, have an effect? If only a small amount of dsRNA is enough to silence an mRNA many times more abundant, is this trigger just a catalyst or is there an amplification mechanism? What are the key players regulating this effect and is this mechanism conserved? What are the RNA molecules that mediate this suppression?

The first answers came from plants where an uniform class of 25nt small RNAs sense and antisense to the targeted gene was seen only during gene silencing [124]. It had been known previously in *C. elegans* that *lin-4* encodes a small RNA 22nt that is antisense to lin-14 and regulates its expression [16]. These along with *in vitro* studies in *Drosophila* [125, 126] suggest that small RNAs are the major players in the silencing mechanism. Based on their origin and function, these small RNAs are classified into three major classes – micro RNAs (miRNAs), small interfering RNAs (siRNAs) and piwi interacting RNAs (piRNAs). Depending on the class of small RNAs, the key players responsible for the biogenesis, transport and functionality vary. For this thesis we will discuss miRNAs and siRNAs since they have been shown to interact with 3'UTRs of genes.

1.4.1: MicroRNAs

MicroRNAs define the first class of small RNAs to be identified. In 1993 Ambros and colleagues showed that the gene lin-4 in *C.elegans,* known to control developmental timing, did not code for a protein but generated two small RNAs 61nt and 22nt in length [16]. They showed that the longer RNA folds into a hairpin and serves as the precursor of the shorter RNA. The 22nt sequence was shown to contain sequence antisense to the 3'UTR of lin-14 and regulated its expression, post-transcriptionally affecting the protein level without affecting the mRNA level [16, 17]. The Ruvkun lab later showed that another miRNA, 21nt let-7, functioned in controlling developmental timing through complementary regions in the 3'UTR of its target genes and was evolutionarily conserved all the way from worms, fruit flies, molluscs, sea urchins, zebra fish, frogs, chicken, and mouse to humans, suggesting regulation by miRNAs is an evolutionarily conserved mechanism [127, 128]. This spurred a rapid interest in identifying more species of small RNAs, now making it a major class.[15, 129-144]. Many of these miRNAs have also shown to be evolutionarily conserved in related species. MiRNAs are now known to control a variety of biological processes including cell proliferation [145], fat metabolism [146], cell death [145], neuronal patterning [147], developmental regulation [148], flowering and plant development [148], and brain morphogenesis [149].

The biogenesis of miRNAs in eukaryotes is a multi-step process. Hairpin structures in long Pol II transcripts (~1-2kb in length) called primary miRNA [150] are recognized by the RNAse III enzyme called Drosha in the nucleus which cleaves them into short hairpins called precursor miRNA[151]. These are ~60-100 nt in length and this association is assisted by DGCR8 [152]. This cleavage results in a 5'phosphate and a 3' 2nt overhang at the base of the hairpin [151, 153]. Some precursor miRNAs are also processed directly from introns without processing by Drosha [154]. The precursor miRNA is then exported to the cytoplasm by exportin-5 [155-157]. Once in the cytoplasm, a RNAse called Dicer processes the hairpin resulting in a short double strand duplex ~21-23 nt in length [158-161]. Dicer cleavage also results in 5' phosphate and 2nt 3' overhangs [162]. From this short duplex, one strand is called the mature miRNA and it is loaded into the miRNA RNA induced silencing complex (miRISC) through the Ago protein [14, 15, 126, 163, 164]. The other complementary strand is called miRNA* (star)[129, 130, 135, 138] . Once the mature miRNA is loaded into miRISC it guides the RISC to the corresponding targets. Target identification ocurrs through the complementarity to the 2-7 bp seed regions of the miRNA in the target 3'UTR [16, 17, 19, 165]. Once the target is identified, the miRNA can regulate the gene expression in one of two ways: mRNA cleavage [14, 166] or translational repression [16, 17, 19]. This can result in regulation of the protein output [167]. Studies have shown that this interaction between the miRNAs and the target 3'UTRs is not exclusive. A single miRNA can regulate multiple mRNAs and multiple miRNAs can have target sites in the same

3'UTR. A miRNA can regulate the target either alone or in a combinatorial fashion [19, 168, 169]. This makes prediction of these interactions extremely complex. However, few computational attempts have been made to generate this interaction map [170-178].

1.4.2: Small interfering RNAs

The first mammalian endo-siRNAs were reported to target LINE-1 retrotransposons in human cells [179]. Further studies in *Drosophila* somatic and germ cells identified an abundant class of small RNAs [180-183] and have been shown to be indispensable for germline maintenance, defense against transposons, and heterochromatic silencing [184]. Soon they were also seen in other organisms including *S. Pombeii* [185] and mouse [186, 187], opening up a new class of abundant small RNAs. Small interfering RNAs or siRNAs in *C.elegans* are of length ~22nt [188] and have perfect complementarity to the target mRNA (3'UTR and coding sequence). They require Dicer, ERI-1 endonuclease and RRF-3 RNA dependent RNA polymerase activity for their biogenesis [189, 190]. Recently they have also been discovered in other organisms with roles in gene regulation [144, 182, 186, 187, 191], suppression of transposons [180, 192], spermatogenesis [193, 194], genome surveillance [195] and chromosome segregation [196]. The siRNAs share many similarities with the miRNAs in terms of their biogenesis and targeting mechanism (reviewed in [197, 198]). Both of them require processing of dsRNA by Dicer and are loaded into RISC. However there are a few differences. First, while miRNAs are transcribed from distinct loci or from introns, the siRNAs are derived from existing loci such as the mRNA itself, transposons, or from external sources such as viruses. Second, miRNAs are processed from hairpin folding of a single RNA while siRNAs are generated from long RNA duplexes which may occur by bidirectional transcription, through RNA directed RNA polymerases, or from external sources such as viruses. Third, one hairpin generates one miRNA whereas one dsRNA duplex can generate multiple siRNA. Fourth, miRNAs target 3'UTR regions with imperfect complementarity while siRNAs have perfect complementarity with their target genes. Finally, different proteins are responsible for loading miRNAs and siRNAs into the RISC. For example miRNAs are loaded by the AGO2 protein while the siRNAs are loaded by a variety of other Argonautes.

My thesis work aimed to study the germline specific small RNA population in *C. elegans* namely to differentiate small RNA populations which are specific to sperm or oocyte and how they are inherited in the embryo. I analyzed high throughput sequencing from isolated germ cells and embryo and characterized a 26nt long class of siRNAs starting with a Guanine, hence named 26GRNAs that have unique populations specific to sperm and oocytes. These small RNAs regulate the expression of thousands of genes in *C. elegans*. We also saw that in addition to targeting coding regions of mRNAs, these 26GRNAs also target 3'UTRs of mRNAs, which adds another class of small RNA other than miRNAs that play a regulatory role in 3'UTRs.

1.5: Whole genome transcriptome annotation

Recent studies indicate that a vast part of the genome is transcribed, however current gene annotations only cover a small portion of this. Traditional reverse transcription and cDNA sequencing technologies have identified ribosomal, transfer and messenger RNA transcripts and in combination with current high throughput sequencing methods have added new classes of small RNAs, including micro RNAs [16], small interfering RNAs [124], piwi interacting RNAs [199], tiRNAs [200], and TASRs[201]. These new findings now reveal the fact that transcription is more pervasive than previously thought and covers more regions on the genome including both strands.

Our current understanding of genes arises from messenger RNAs. These mRNAs have been shown to generate multiple transcripts arising from the same genomic loci and the diversity is enforced by the choice of different transcriptional start sites or post transcriptional mechanisms including alternative splicing and polyadenylation. Based on the transcript generated, the functionality of the end protein can vary. Hence now the concept of the gene is moved from DNA to the RNA domain or the domain of the transcriptome. DNA elements in the promoter are still important in regulating transcription and deciding the start site to be used. Similarly, regions of the gene are also sites for generation of a number of non-coding RNAs. Studies have shown small RNAs, such as miRNAs, being generated from the spliced introns of mRNAs[154] . In addition, many of the small RNAs are Pol II transcripts and would require promoter regions for transcription before being processed into their mature form.

While traditional cDNA cloning methods (ESTs and full length cDNA) are responsible for identifying many of our current annotations and are still the gold standard [202], newer approaches are now catching up. One reason for the transition is the cost involved in cloning on a gene-by-gene basis. Furthermore, with the amount of transcript diversity now being realized there are still a lot more transcripts which may not have been identified due to low abundance, biological conditions, or masking by other abundant transcripts. One such example is the abundance of the rRNA transcripts in our polyA capture library. In the initial pilot study, 50% of our library was dominated by ~70 ribosomal genes. Subtracting these transcripts dramatically increased the number of newly identified low abundance transcripts. Hence to get a complete unbiased transcriptome and an increased discovery rate we have to use newer high throughput methods.

Whole genome transcriptome analysis gives evidence for transcription from both strands of the DNA and these antisense transcripts link neighboring genes to form transcriptional units. Antisense transcripts can provide the template for biogenesis of small RNAs resulting in sense strand cleavage, which forms the basis of RNAi[203, 204]. Disruption of the antisense transcription loci in mouse shows alteration in sense strand expression [205]. An important observation seen in Arabidopsis [88], yeast [92], mouse and human [92] 3'UTRomes was the presence of abundant antisense transcripts. Initial studies show 33% of *Arabidopsis* transcripts[88], 60% in yeast[92] and 30% of human transcripts[92] expressed antisense transcripts which could affect sense gene expression positively or negatively. A recent study in *C. elegans* postulates that transcription from both strands help in sharing transcriptional machinery factors and promote “genome compaction” [206].

With the recent advancement of high throughput technologies such asmicroarrays, tiling arrays and high throughput sequencing, the estimated number of transcribed loci is increasing every day. However, the field is still far from saturation and every experiment performed is identifying new transcripts. Our study identified ~1000 new loci not previously annotated and that are polyadenylated [81]. We also annotated ~26,000 3'ends of mRNAs and a parallel study identified ~9,800 new 3'ends that were not seen in our dataset [206]. A massive sequencing study in *C. elegans* identified ~28,000 new splice sites [207]. CAGE analysis in humans shows that there are ~67,000 transcription start sites in humans suggesting a much higher number of unique transcripts than currently estimated [208]. A whole genome tiling array study in humans demonstrated that a large portion of the genome (~25%) is transcribed into RNAs and that a large portion of them is cell type-specific [209]. Another study showed that there are as many polyadenylated RNAs as there are non-polyadenylated, and ~40% of these RNAs are confined to the nucleus [210]. Studies examining alternate polyadenylation have shown in organisms including *C. elegans*[81], *Arabidopsis*[88], rice [89], *Chlamydomonas* [90], trypanosomes [82], mouse [51, 91], humans [91-93] and yeast [94] that alternative polyadenylation is pervasive and affects 40-70% of the total number of genes in an organism. This gives an astonishing view of the complexity and diversity of the transcriptome and identifying the expression patterns, biogenesis and functionality of these transcripts will be a field of research for years to come.

1.6: Remaining questions:

There are many questions that come to our mind when we look at all the data presented to us. My thesis aims to answer some of these. Future research will further shed light on some or all of these.

First, looking at the global scale of alternative polyadenylation, there are many questions that face us about this mechanism. Some of the immediate ones are about the mechanism, such as how is alternative polyadenylation different from standard polyadenylation in terms of the protein factors used? Do the same proteins of the 3'end processing also mediate alternative polyadenylation or are there new proteins involved in the mechanism that are yet to be discovered? What is the contribution of factors from splicing and transcription machinery to alternative polyadenylation?

A second set of questions involves the sequence determinants defining alternative polyadenylation some of which were answered in chapter two. What are the sequence elements that drive alternative polyadenylation? Are PAS signals sufficient or are there auxiliary sequence elements that decide what PAS site to use? Is there any length bias towards utilization of a PAS?

Third, we have to analyze the effects of alternative polyadenylation in terms of localization in the cell. How does the localization profile vary between a regularly polyadenylated mRNA and an alternatively polyadenylated mRNA? Does alternative polyadenylation regulate where a mRNA is localized in a cell?

Fourth, we question the stability of the mRNA. To what extent and how does alternative polyadenylation affect stability of the mRNA? Does the change in length alone alter stability?

A fifth set of questions is regarding the interaction between the 3'UTRs and small RNAs, and how alternative polyadenylation can define this interaction or vice-versa. What are the small RNAs that interact with each 3'UTR in the organism and how does their expression change over different tissues, developmental timing and response to external stimulus. Does alternative polyadenylation really occur to escape regulation from small RNAs? If so, what are the interaction maps for the mRNA and the small RNA?

Finally, there are questions regarding the functional effect of alternative polyadenylation. How does the translational output vary with polyadenylation? Does alternative polyadenylation alter protein output and if so, what are the factors that mediate this? How do individual genes alter 3'UTR length for developmental and environmental cues? Is there any particular bias for long or short 3'UTRs in any specific tissue or developmental timing or both? We have shown variation of 3'UTR length to developmental timing in chapter two. What is the spatio-temporal map of 3'UTR lengths for all genes in an organism? What is the evolutionary need for alternative polyadenylation?

1.7: Thesis outline

The overall goal of this thesis is to utilize advances in high throughput sequencing methods to provide a comprehensive 3'UTRome of *C.elegans* on a whole genome scale.

Chapter 1 provides an introduction to the thesis discussing various factors and mechanisms involved in 3'end processing. It also gives a basic introduction to the various classes of small RNAs.

Chapter 2 presents the results of our collaborative 3'UTRome project, which was part of the modENCODE Consortium. The focus of our collaborative group was to provide insights into the mechanisms of 3'UTR biogenesis. The chapter highlights our novel work in cataloguing the 3'UTRs for more than 75% of genes in *C.elegans*, ~4,500 of which had no previous 3'UTR annotation. Ours results showed variable 3'UTR lengths for ~40% of these genes, which for the first time shed light on the scope of alternative polyadenylation. Our analysis also identified sequence elements other than the known canonical AAUAAA signal, which also seems to play a role in defining the 3'end of mRNA. We also showed 3'UTRs that are uniquely expressed in specific developmental stages of *C. elegans,* highlighting a temporal role of the alternative polyadenylation mechanism. Further analysis also identified examples of genes utilizing longer or shorter 3'UTRs in specific developmental stages, while switching to a different length with progression of development. This showed a regulatory role of the 3'end processing mechanism. In conclusion we defined the prominence of this 3'end defining mechanism which might provide a level of regulation in addition to post-transcriptional control by small RNAs.

Chapter 3 integrates our results with a parallel published study of sequencing 3'UTRs from different developmental stages of *C.elegans*. Comparing the results of the two studies provides additional validation for our data and analysis results. Identifying the differential 3'UTRs between the two independent datasets helped us to define stricter filters to effectively remove the inherent false priming artifacts in the sequencing protocol. This helped us achieve a higher quality 3'UTRome and benchmark for future sequencing using the polyA capture protocol.

Chapter 4 is extension of the 3'UTRome project where we apply the polyA capture protocol, designed to answer a specific biological question. Here we studied changes in the 3'UTRome of an organism when we disrupt components in one or both pathways in synapse and axon biogenesis. We specifically studied these pathways since previous research has shown evidence for relation between the components in the axon and synapse development pathways and 3'end processing of mRNAs. The results of this analysis generate a pathway specific 3'UTRome.

Chapter 5 presents related work where we aimed to study the small RNAs that potentially target the 3'UTR regions. We specifically chose to sequence the small RNAs from isolated sperm, oocytes, embryos, *glp-4* mutant defective in germline development and *eri-1* mutant defective in small RNA biogenesis. These choices helped us to study the gamete and germline specific small RNA distribution including microRNAs, 21U RNAs and endogenous small interfering RNAs. By analyzing the sequences, we characterized a new class of 26nt long RNAs, which start with a Guanine (hence named 26GRNAs). These RNAs were predominantly antisense to their mRNAs targeting 3'UTRs and coding regions of the mRNAs. Based on their targeting and source library they could be further classified into two sub-classes – one subclass from the sperm library targeting the spermatogenic genes and the other deriving from the oocytes to be inherited maternally into the embryos that function in zygotic development.

Chapter 6 summarizes the overall results of the thesis and its contribution to the field of research and presents possible future directions.

1.8: Contributions

The following journal publications represent the work detailed within this thesis.

- Mangone, M***., Manoharan, A.P***., Thierry-Mieg, D*., Thierry-Mieg, J*., Han, T*., Mackowiak, S., Mis, E., Zegar, C., Gutwein, M.R., Khivansara, V., Salehi-Ashtiani, K., Harkins, T. Bouffard, P., Suzuki, Y., Sugano, S., Kohara, Y., Rajewsky, N., Piano, F., Gunsalus, K.C., and Kim**,** J.K. The landscape of *C. elegans* 3' UTRs. *Science* 329: 432-5 (2010).
  *These authors contributed equally to this work.

- modENCODE Consortium. Unlocking the secrets of the genome. *Nature* 459: 927-30 (2009).

- Han, T., **Manoharan, A.P**., Harkins, T.T., Bouffard, P., Fitzpatrick, C., Chu, D.S., Thierry-Mieg, D., Thierry-Mieg, J., and Kim, J.K**.** 26G endo-siRNAs regulate spermatogenic and zygotic gene expression in *C. elegans*. *Proc. Natl. Acad. Sci. USA* 106:18674-9 (2009).

1.9: Reference

1. Shuman S: **Structure, mechanism, and evolution of the mRNA capping apparatus**. *Prog Nucleic Acid Res Mol Biol* 2001, **66**:1-40.
2. Hsu CL, Stevens A: **Yeast cells lacking 5'-->3' exoribonuclease 1 contain mRNA species that are poly(A) deficient and partially lack the 5' cap structure**. *Mol Cell Biol* 1993, **13**(8):4826-4835.
3. Walther TN, Wittop Koning TH, Schumperli D, Muller B: **A 5'-3' exonuclease activity involved in forming the 3' products of histone pre-mRNA processing in vitro**. *Rna* 1998, **4**(9):1034-1046.
4. Tarun SZ, Jr., Sachs AB: **Association of the yeast poly(A) tail binding protein with translation initiation factor eIF-4G**. *EMBO J* 1996, **15**(24):7168-7177.
5. Manley JL: **Polyadenylation of mRNA precursors**. *Biochim Biophys Acta* 1988, **950**(1):1-12.
6. Jacobson A, Peltz SW: **Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells**. *Annu Rev Biochem* 1996, **65**:693-739.
7. Lewis JD, Gunderson SI, Mattaj IW: **The influence of 5' and 3' end structures on pre-mRNA metabolism**. *J Cell Sci Suppl* 1995, **19**:13-19.
8. Wickens M, Anderson P, Jackson RJ: **Life and death in the cytoplasm: messages from the 3' end**. *Curr Opin Genet Dev* 1997, **7**(2):220-232.
9. Edwalds-Gilbert G, Veraldi KL, Milcarek C: **Alternative poly(A) site selection in complex transcription units: means to an end?** *Nucleic Acids Res* 1997, **25**(13):2547-2561.
10. Crick F: **Split genes and RNA splicing**. *Science* 1979, **204**(4390):264-271.
11. Stark H, Luhrmann R: **Cryo-electron microscopy of spliceosomal components**. *Annu Rev Biophys Biomol Struct* 2006, **35**:435-457.
12. Richter JD, Sonenberg N: **Regulation of cap-dependent translation by eIF4E inhibitory proteins**. *Nature* 2005, **433**(7025):477-480.
13. Moore MJ, Proudfoot NJ: **Pre-mRNA Processing Reaches Back toTranscription and Ahead to Translation**. *Cell* 2009, **136**(4):688-700.
14. Hutvagner G, Zamore PD: **A microRNA in a multiple-turnover RNAi enzyme complex**. *Science* 2002, **297**(5589):2056-2060.
15. Mourelatos Z, Dostie J, Paushkin S, Sharma A, Charroux B, Abel L, Rappsilber J, Mann M, Dreyfuss G: **miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs**. *Genes Dev* 2002, **16**(6):720-728.
16. Lee RC, Feinbaum RL, Ambros V: **The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14**. *Cell* 1993, **75**(5):843-854.
17. Wightman B, Ha I, Ruvkun G: **Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans***. *Cell* 1993, **75**(5):855-862.
18. Filipowicz W, Bhattacharyya SN, Sonenberg N: **Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?** *Nat Rev Genet* 2008, **9**(2):102-114.
19. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM: **Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs**. *Nature* 2005, **433**(7027):769-773.
20. Vinciguerra P, Stutz F: **mRNA export: an assembly line from genes to nuclear pores**. *Curr Opin Cell Biol* 2004, **16**(3):285-292.

21. Huang Y, Carmichael GG: **Nucleocytoplasmic mRNA transport**. *Results Probl Cell Differ* 2001, **34**:139-155.

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

23. Wormington M, Searfoss AM, Hurney CA: **Overexpression of poly(A) binding protein prevents maturation-specific deadenylation and translational inactivation in Xenopus oocytes**. *EMBO J* 1996, **15**(4):900-909.

24. Wilusz CJ, Wormington M, Peltz SW: **The cap-to-tail guide to mRNA turnover**. *Nat Rev Mol Cell Biol* 2001, **2**(4):237-246.

25. Proudfoot NJ, Furger A, Dye MJ: **Integrating mRNA processing with transcription**. *Cell* 2002, **108**(4):501-512.

26. Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D: **Patterns of variant polyadenylation signal usage in human genes**. *Genome Res* 2000, **10**(7):1001-1010.

27. Higgs DR, Goodbourn SE, Lamb J, Clegg JB, Weatherall DJ, Proudfoot NJ: **Alpha-thalassaemia caused by a polyadenylation signal mutation**. *Nature* 1983, **306**(5941):398-400.

28. Wickens M, Stephenson P: **Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3' end formation**. *Science* 1984, **226**(4678):1045-1051.

29. Fitzgerald M, Shenk T: **The site at which late mRNAs are polyadenylated is altered in SV40 mutant dl882**. *Ann N Y Acad Sci* 1980, **354**:53-59.

30. Sheets MD, Ogg SC, Wickens MP: **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* 1990, **18**(19):5799-5805.

31. Wahle E, Ruegsegger U: **3'-End processing of pre-mRNA in eukaryotes**. *FEMS Microbiol Rev* 1999, **23**(3):277-295.

32. Simonsen CC, Levinson AD: **Analysis of processing and polyadenylation signals of the hepatitis B virus surface antigen gene by using simian virus 40-hepatitis B virus chimeric plasmids**. *Mol Cell Biol* 1983, **3**(12):2250-2258.

33. Venkataraman K, Brown KM, Gilmartin GM: **Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition**. *Genes Dev* 2005, **19**(11):1315-1327.

34. Arhin GK, Boots M, Bagga PS, Milcarek C, Wilusz J: **Downstream sequence elements with different affinities for the hnRNP H/H' protein influence the processing efficiency of mammalian polyadenylation signals**. *Nucleic Acids Res* 2002, **30**(8):1842-1850.

35. Bagga PS, Ford LP, Chen F, Wilusz J: **The G-rich auxiliary downstream element has distinct sequence and position requirements and mediates efficient 3' end pre-mRNA processing through a trans-acting factor**. *Nucleic Acids Res* 1995, **23**(9):1625-1631.

36. Dalziel M, Nunes NM, Furger A: **Two G-rich regulatory elements located adjacent to and 440 nucleotides downstream of the core poly(A) site of the intronless melanocortin receptor 1 gene are critical for efficient 3' end processing**. *Mol Cell Biol* 2007, **27**(5):1568-1580.

37. Moore CL, Sharp PA: **Site-specific polyadenylation in a cell-free reaction**. *Cell* 1984, **36**(3):581-591.

38. Mandel CR, Bai Y, Tong L: **Protein factors in pre-mRNA 3'-end processing**. *Cell Mol Life Sci* 2008, **65**(7-8):1099-1122.

39. Millevoi S, Vagner S: **Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation**. *Nucleic Acids Research* 2009, **38**(9):2757-2774.
40. Keller W, Bienroth S, Lang KM, Christofori G: **Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal AAUAAA**. *EMBO J* 1991, **10**(13):4241-4249.
41. Murthy KG, Manley JL: **The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation**. *Genes Dev* 1995, **9**(21):2672-2683.
42. Takagaki Y, Manley JL: **RNA recognition by the human polyadenylation factor CstF**. *Mol Cell Biol* 1997, **17**(7):3907-3914.
43. Takagaki Y, Manley JL, MacDonald CC, Wilusz J, Shenk T: **A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs**. *Genes Dev* 1990, **4**(12A):2112-2120.
44. Moore CL, Chen J, Whoriskey J: **Two proteins crosslinked to RNA containing the adenovirus L3 poly(A) site require the AAUAAA sequence for binding**. *EMBO J* 1988, **7**(10):3159-3169.
45. Bai Y, Auperin TC, Chou CY, Chang GG, Manley JL, Tong L: **Crystal structure of murine CstF-77: dimeric association and implications for polyadenylation of mRNA precursors**. *Mol Cell* 2007, **25**(6):863-875.
46. Dantonel JC, Murthy KG, Manley JL, Tora L: **Transcription factor TFIID recruits factor CPSF for formation of 3' end of mRNA**. *Nature* 1997, **389**(6649):399-402.
47. McCracken S, Fong N, Yankulov K, Ballantyne S, Pan G, Greenblatt J, Patterson SD, Wickens M, Bentley DL: **The C-terminal domain of RNA polymerase II couples mRNA processing to transcription**. *Nature* 1997, **385**(6614):357-361.
48. Ryan K, Calvo O, Manley JL: **Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease**. *Rna* 2004, **10**(4):565-573.
49. Kaufmann I, Martin G, Friedlein A, Langen H, Keller W: **Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase**. *EMBO J* 2004, **23**(3):616-626.
50. Dass B, Tardif S, Park JY, Tian B, Weitlauf HM, Hess RA, Carnes K, Griswold MD, Small CL, Macdonald CC: **Loss of polyadenylation protein tauCstF-64 causes spermatogenic defects and male infertility**. *Proc Natl Acad Sci U S A* 2007, **104**(51):20374-20379.
51. Liu D, Brockman JM, Dass B, Hutchins LN, Singh P, McCarrey JR, MacDonald CC, Graber JH: **Systematic variation in mRNA 3'-processing signals during mouse spermatogenesis**. *Nucleic Acids Res* 2007, **35**(1):234-246.
52. McMahon KW, Hirsch BA, MacDonald CC: **Differences in polyadenylation site choice between somatic and male germ cells**. *BMC Mol Biol* 2006, **7**:35.
53. Wallace AM, Dass B, Ravnik SE, Tonk V, Jenkins NA, Gilbert DJ, Copeland NG, MacDonald CC: **Two distinct forms of the 64,000 Mr protein of the cleavage stimulation factor are expressed in mouse male germ cells**. *Proc Natl Acad Sci U S A* 1999, **96**(12):6763-6768.
54. Juge F, Audibert A, Benoit B, Simonelig M: **Tissue-specific autoregulation of Drosophila suppressor of forked by alternative poly(A) site utilization leads to accumulation of the suppressor of forked protein in mitotically active cells**. *Rna* 2000, **6**(11):1529-1538.
55. Brown KM, Gilmartin GM: **A mechanism for the regulation of pre-mRNA 3' processing by human cleavage factor Im**. *Mol Cell* 2003, **12**(6):1467-1476.

56. Kim S, Yamamoto J, Chen Y, Aida M, Wada T, Handa H, Yamaguchi Y: **Evidence that cleavage factor Im is a heterotetrameric protein complex controlling alternative polyadenylation**. *Genes Cells* 2010, **15**(9):1003-1013.
57. Kubo T, Wada T, Yamaguchi Y, Shimizu A, Handa H: **Knock-down of 25 kDa subunit of cleavage factor Im in Hela cells alters alternative polyadenylation within 3'-UTRs**. *Nucleic Acids Res* 2006, **34**(21):6264-6271.
58. Sartini BL, Wang H, Wang W, Millette CF, Kilpatrick DL: **Pre-messenger RNA cleavage factor I (CFIm): potential role in alternative polyadenylation during spermatogenesis**. *Biol Reprod* 2008, **78**(3):472-482.
59. Yang Q, Coseno M, Gilmartin GM, Doublié S: **Crystal Structure of a Human Cleavage Factor CFIm25/CFIm68/RNA Complex Provides an Insight into Poly(A) Site Recognition and RNA Looping**. *Structure* 2011, **19**(3):368-377.
60. Hollingworth D, Noble CG, Taylor IA, Ramos A: **RNA polymerase II CTD phosphopeptides compete with RNA for the interaction with Pcf11**. *Rna* 2006, **12**(4):555-560.
61. de Vries H, Ruegsegger U, Hubner W, Friedlein A, Langen H, Keller W: **Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors**. *EMBO J* 2000, **19**(21):5895-5904.
62. Kuhn U, Gundel M, Knoth A, Kerwitz Y, Rudel S, Wahle E: **Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor**. *J Biol Chem* 2009, **284**(34):22803-22814.
63. Meyer S, Urbanke C, Wahle E: **Equilibrium studies on the association of the nuclear poly(A) binding protein with poly(A) of different lengths**. *Biochemistry* 2002, **41**(19):6082-6089.
64. Keller RW, Kuhn U, Aragon M, Bornikova L, Wahle E, Bear DG: **The nuclear poly(A) binding protein, PABP2, forms an oligomeric particle covering the length of the poly(A) tail**. *J Mol Biol* 2000, **297**(3):569-583.
65. Kolev NG, Steitz JA: **Symplekin and multiple other polyadenylation factors participate in 3'-end maturation of histone mRNAs**. *Genes Dev* 2005, **19**(21):2583-2592.
66. Dichtl B, Blank D, Sadowski M, Hubner W, Weiser S, Keller W: **Yhh1p/Cft1p directly links poly(A) site recognition and RNA polymerase II transcription termination**. *EMBO J* 2002, **21**(15):4125-4135.
67. Gross S, Moore CL: **Rna15 interaction with the A-rich yeast polyadenylation signal is an essential step in mRNA 3'-end formation**. *Mol Cell Biol* 2001, **21**(23):8045-8055.
68. Licatalosi DD, Geiger G, Minet M, Schroeder S, Cilli K, McNeil JB, Bentley DL: **Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II**. *Mol Cell* 2002, **9**(5):1101-1111.
69. Wang SW, Asakawa K, Win TZ, Toda T, Norbury CJ: **Inactivation of the pre-mRNA cleavage and polyadenylation factor Pfs2 in fission yeast causes lethal cell cycle defects**. *Mol Cell Biol* 2005, **25**(6):2288-2296.
70. Dominski Z, Marzluff WF: **Formation of the 3′ end of histone mRNA: Getting closer to the end**. *Gene* 2007, **396**(2):373-390.
71. Marzluff WF, Wagner EJ, Duronio RJ: **Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail**. *Nat Rev Genet* 2008, **9**(11):843-854.
72. Pettitt J, Crombie C, Schumperli D, Muller B: **The *Caenorhabditis elegans* histone hairpin-binding protein is required for core histone gene**

expression and is essential for embryonic and postembryonic cell division. *J Cell Sci* 2002, **115**(Pt 4):857-866.

73. Williams AS, Marzluff WF: **The sequence of the stem and flanking sequences at the 3' end of histone mRNA are critical determinants for the binding of the stem-loop binding protein**. *Nucleic Acids Res* 1995, **23**(4):654-662.
74. Godfrey AC, White AE, Tatomer DC, Marzluff WF, Duronio RJ: **The Drosophila U7 snRNP proteins Lsm10 and Lsm11 are required for histone pre-mRNA processing and play an essential role in development**. *Rna* 2009, **15**(9):1661-1672.
75. Yang XC, Burch BD, Yan Y, Marzluff WF, Dominski Z: **FLASH, a proapoptotic protein involved in activation of caspase-8, is essential for 3' end processing of histone pre-mRNAs**. *Mol Cell* 2009, **36**(2):267-278.
76. Dominski Z, Yang XC, Marzluff WF: **The polyadenylation factor CPSF-73 is involved in histone-pre-mRNA processing**. *Cell* 2005, **123**(1):37-48.
77. Blumenthal T, Spieth J: **Gene structure and organization in Caenorhabditis elegans**. *Curr Opin Genet Dev* 1996, **6**(6):692-698.
78. Blumenthal T, Steward K: **RNA Processing and Gene Structure**. 1997.
79. Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu WL, Duke K, Kiraly M *et al*: **A global analysis of *Caenorhabditis elegans* operons**. *Nature* 2002, **417**(6891):851-854.
80. Denker JA, Zuckerman DM, Maroney PA, Nilsen TW: **New components of the spliced leader RNP required for nematode trans-splicing**. *Nature* 2002, **417**(6889):667-670.
81. Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V *et al*: **The Landscape of *C. elegans* 3'UTRs**. *Science* 2010, **329**(5990):432-435.
82. Jager AV, De Gaudenzi JG, Cassola A, D'Orso I, Frasch AC: **Inaugural Article: mRNA maturation by two-step trans-splicing/polyadenylation processing in trypanosomes**. *Proceedings of the National Academy of Sciences* 2007, **104**(7):2035-2042.
83. Allen MA, Hillier LW, Waterston RH, Blumenthal T: **A global analysis of *C. elegans* trans-splicing**. *Genome Res* 2011, **21**(2):255-264.
84. Zaslaver A, Baugh LR, Sternberg PW: **Metazoan operons accelerate recovery from growth-arrested states**. *Cell* 2011, **145**(6):981-992.
85. Alt FW, Bothwell AL, Knapp M, Siden E, Mather E, Koshland M, Baltimore D: **Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends**. *Cell* 1980, **20**(2):293-301.
86. Early P, Rogers J, Davis M, Calame K, Bond M, Wall R, Hood L: **Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways**. *Cell* 1980, **20**(2):313-319.
87. Rogers J, Early P, Carter C, Calame K, Bond M, Hood L, Wall R: **Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin mu chain**. *Cell* 1980, **20**(2):303-312.
88. Wu X, Liu M, Downie B, Liang C, Ji G, Li QQ, Hunt AG: **Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation**. *Proceedings of the National Academy of Sciences* 2011.
89. Shen Y, Ji G, Haas BJ, Wu X, Zheng J, Reese GJ, Li QQ: **Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation**. *Nucleic Acids Res* 2008, **36**(9):3150-3161.

90. Shen Y, Liu Y, Liu L, Liang C, Li QQ: **Unique features of nuclear mRNA poly(A) signals and alternative polyadenylation in Chlamydomonas reinhardtii**. *Genetics* 2008, **179**(1):167-176.

91. Tian B, Hu J, Zhang H, Lutz CS: **A large-scale analysis of mRNA polyadenylation of human and mouse genes**. *Nucleic Acids Res* 2005, **33**(1):201-212.

92. Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM: **Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation**. *Cell* 2010, **143**(6):1018-1029.

93. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470-476.

94. Keller W, Minvielle-Sebastia L: **A comparison of mammalian and yeast pre-mRNA 3'-end processing**. *Curr Opin Cell Biol* 1997, **9**(3):329-336.

95. Chabot B: **Directing alternative splicing: cast and scenarios**. *Trends Genet* 1996, **12**(11):472-478.

96. Lutz CS: **Alternative polyadenylation: a twist on mRNA 3' end formation**. *ACS Chem Biol* 2008, **3**(10):609-617.

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

98. Zhang H, Lee JY, Tian B: **Biased alternative polyadenylation in human tissues**. *Genome Biol* 2005, **6**(12):R100.

99. Ji Z, Lee JY, Pan Z, Jiang B, Tian B: **Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development**. *Proc Natl Acad Sci U S A* 2009, **106**(17):7028-7033.

100. Ji Z, Tian B: **Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types**. *PLoS One* 2009, **4**(12):e8419.

101. Didiano D, Cochella L, Tursun B, Hobert O: **Neuron-type specific regulation of a 3'UTR through redundant and combinatorially acting cis-regulatory elements**. *Rna* 2010, **16**(2):349-363.

102. Lau AG, Irier HA, Gu J, Tian D, Ku L, Liu G, Xia M, Fritsch B, Zheng JQ, Dingledine R *et al*: **Distinct 3'UTRs differentially regulate activity-dependent translation of brain-derived neurotrophic factor (BDNF)**. *Proceedings of the National Academy of Sciences* 2010, **107**(36):15945-15950.

103. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB: **Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites**. *Science* 2008, **320**(5883):1643-1647.

104. Mayr C, Bartel DP: **Widespread Shortening of 3′UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells**. *Cell* 2009, **138**(4):673-684.

105. Fu Y, Sun Y, Li Y, Li J, Rao X, Chen C, Xu A: **Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing**. *Genome Research* 2011, **21**(5):741-747.

106. Winter J, Kunath M, Roepcke S, Krause S, Schneider R, Schweiger S: **Alternative polyadenylation signals and promoters act in concert to control tissue-specific expression of the Opitz Syndrome gene MID1**. *BMC Mol Biol* 2007, **8**:105.

107. Takagaki Y, Seipelt RL, Peterson ML, Manley JL: **The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation**. *Cell* 1996, **87**(5):941-952.

108. Andreassi C, Riccio A: **To localize or not to localize: mRNA fate is in 3′UTR ends**. *Trends in Cell Biology* 2009, **19**(9):465-474.

109. Napoli C, Lemieux C, Jorgensen R: **Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans**. *Plant Cell* 1990, **2**(4):279-289.

110. Ruiz F, Vayssie L, Klotz C, Sperling L, Madeddu L: **Homology-dependent gene silencing in Paramecium**. *Mol Biol Cell* 1998, **9**(4):931-943.

111. Kelly WG, Fire A: **Chromatin silencing and the maintenance of a functional germline in *Caenorhabditis elegans***. *Development* 1998, **125**(13):2451-2456.

112. Pal-Bhadra M, Bhadra U, Birchler JA: **Cosuppression of nonhomologous transgenes in Drosophila involves mutually related endogenous sequences**. *Cell* 1999, **99**(1):35-46.

113. Romano N, Macino G: **Quelling: transient inactivation of gene expression in Neurospora crassa by transformation with homologous sequences**. *Mol Microbiol* 1992, **6**(22):3343-3353.

114. Wassenegger M, Heimes S, Riedel L, Sanger HL: **RNA-directed de novo methylation of genomic sequences in plants**. *Cell* 1994, **76**(3):567-576.

115. Carthew RW: **Gene silencing by double-stranded RNA**. *Curr Opin Cell Biol* 2001, **13**(2):244-248.

116. Ratcliff F, Harrison BD, Baulcombe DC: **A similarity between viral defense and gene silencing in plants**. *Science* 1997, **276**(5318):1558-1560.

117. Lindbo JA, Silva-Rosales L, Proebsting WM, Dougherty WG: **Induction of a Highly Specific Antiviral State in Transgenic Plants: Implications for Regulation of Gene Expression and Virus Resistance**. *Plant Cell* 1993, **5**(12):1749-1759.

118. Kennerdell JR, Carthew RW: **Use of dsRNA-mediated genetic interference to demonstrate that frizzled and frizzled 2 act in the wingless pathway**. *Cell* 1998, **95**(7):1017-1026.

119. Misquitta L, Paterson BM: **Targeted disruption of gene function in Drosophila by RNA interference (RNA-i): a role for nautilus in embryonic somatic muscle formation**. *Proc Natl Acad Sci U S A* 1999, **96**(4):1451-1456.

120. Ngo H, Tschudi C, Gull K, Ullu E: **Double-stranded RNA induces mRNA degradation in Trypanosoma brucei**. *Proc Natl Acad Sci U S A* 1998, **95**(25):14687-14692.

121. Wargelius A, Ellingsen S, Fjose A: **Double-stranded RNA induces specific developmental defects in zebrafish embryos**. *Biochem Biophys Res Commun* 1999, **263**(1):156-161.

122. Wianny F, Zernicka-Goetz M: **Specific interference with gene function by double-stranded RNA in early mouse development**. *Nat Cell Biol* 2000, **2**(2):70-75.

123. Svoboda P, Stein P, Hayashi H, Schultz RM: **Selective reduction of dormant maternal mRNAs in mouse oocytes by RNA interference**. *Development* 2000, **127**(19):4147-4156.

124. Hamilton AJ, Baulcombe DC: **A species of small antisense RNA in posttranscriptional gene silencing in plants**. *Science* 1999, **286**(5441):950-952.

125. Zamore PD, Tuschl T, Sharp PA, Bartel DP: **RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals**. *Cell* 2000, **101**(1):25-33.

126. Hammond SM, Bernstein E, Beach D, Hannon GJ: **An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells**. *Nature* 2000, **404**(6775):293-296.

127. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G: **The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans**. *Nature* 2000, **403**(6772):901-906.

128. Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Muller P *et al*: **Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA**. *Nature* 2000, **408**(6808):86-89.

129. Lau NC, Lim LP, Weinstein EG, Bartel DP: **An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans***. *Science* 2001, **294**(5543):858-862.

130. Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP: **Vertebrate microRNA genes**. *Science* 2003, **299**(5612):1540.

131. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP: **The microRNAs of *Caenorhabditis elegans***. *Genes Dev* 2003, **17**(8):991-1008.

132. Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP: **MicroRNAs in plants**. *Genes Dev* 2002, **16**(13):1616-1626.

133. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T: **Identification of novel genes coding for small expressed RNAs**. *Science* 2001, **294**(5543):853-858.

134. Lee RC, Ambros V: **An extensive class of small RNAs in *Caenorhabditis elegans***. *Science* 2001, **294**(5543):862-864.

135. Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T: **Identification of tissue-specific microRNAs from mouse**. *Curr Biol* 2002, **12**(9):735-739.

136. Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T: **New microRNAs from mouse and human**. *Rna* 2003, **9**(2):175-179.

137. Ambros V, Lee RC, Lavanway A, Williams PT, Jewell D: **MicroRNAs and Other Tiny Endogenous RNAs in *C. elegans***. *Current Biology* 2003, **13**(10):807-818.

138. Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, Gaasterland T, Meyer J, Tuschl T: **The small RNA profile during Drosophila melanogaster development**. *Dev Cell* 2003, **5**(2):337-350.

139. Dostie J, Mourelatos Z, Yang M, Sharma A, Dreyfuss G: **Numerous microRNPs in neuronal cells containing novel microRNAs**. *Rna* 2003, **9**(2):180-186.

140. Kim J, Krichevsky A, Grad Y, Hayes GD, Kosik KS, Church GM, Ruvkun G: **Identification of many microRNAs that copurify with polyribosomes in mammalian neurons**. *Proc Natl Acad Sci U S A* 2004, **101**(1):360-365.

141. Houbaviy HB, Murray MF, Sharp PA: **Embryonic stem cell-specific MicroRNAs**. *Dev Cell* 2003, **5**(2):351-358.

142. Michael MZ, SM OC, van Holst Pellekaan NG, Young GP, James RJ: **Reduced accumulation of specific microRNAs in colorectal neoplasia**. *Mol Cancer Res* 2003, **1**(12):882-891.

143. Kato M, de Lencastre A, Pincus Z, Slack FJ: **Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during *Caenorhabditis elegans* development**. *Genome Biol* 2009, **10**(5):R54.

144. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP: **Large-Scale Sequencing Reveals 21U-RNAs and Additional MicroRNAs and Endogenous siRNAs in *C. elegans***. *Cell* 2006, **127**(6):1193-1207.
145. Brennecke J, Hipfner DR, Stark A, Russell RB, Cohen SM: **bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila**. *Cell* 2003, **113**(1):25-36.
146. Xu P, Vernooy SY, Guo M, Hay BA: **The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism**. *Curr Biol* 2003, **13**(9):790-795.
147. Johnston RJ, Hobert O: **A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans**. *Nature* 2003, **426**(6968):845-849.
148. Kedde M, Agami R: **Interplay between microRNAs and RNA-binding proteins determines developmental processes**. *Cell Cycle* 2008, **7**(7):899-903.
149. Giraldez AJ, Cinalli RM, Glasner ME, Enright AJ, Thomson JM, Baskerville S, Hammond SM, Bartel DP, Schier AF: **MicroRNAs regulate brain morphogenesis in zebrafish**. *Science* 2005, **308**(5723):833-838.
150. Lee Y, Jeon K, Lee JT, Kim S, Kim VN: **MicroRNA maturation: stepwise processing and subcellular localization**. *EMBO J* 2002, **21**(17):4663-4670.
151. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S *et al*: **The nuclear RNase III Drosha initiates microRNA processing**. *Nature* 2003, **425**(6956):415-419.
152. Han J, Lee Y, Yeom KH, Kim YK, Jin H, Kim VN: **The Drosha-DGCR8 complex in primary microRNA processing**. *Genes Dev* 2004, **18**(24):3016-3027.
153. Basyuk E, Suavet F, Doglio A, Bordonne R, Bertrand E: **Human let-7 stem-loop precursors harbor features of RNase III cleavage products**. *Nucleic Acids Res* 2003, **31**(22):6593-6597.
154. Ruby JG, Jan CH, Bartel DP: **Intronic microRNA precursors that bypass Drosha processing**. *Nature* 2007, **448**(7149):83-86.
155. Bohnsack MT, Czaplinski K, Gorlich D: **Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs**. *Rna* 2004, **10**(2):185-191.
156. Lund E, Guttinger S, Calado A, Dahlberg JE, Kutay U: **Nuclear export of microRNA precursors**. *Science* 2004, **303**(5654):95-98.
157. Yi R, Qin Y, Macara IG, Cullen BR: **Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs**. *Genes Dev* 2003, **17**(24):3011-3016.
158. Hutvagner G, McLachlan J, Pasquinelli AE, Balint E, Tuschl T, Zamore PD: **A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA**. *Science* 2001, **293**(5531):834-838.
159. Grishok A, Pasquinelli AE, Conte D, Li N, Parrish S, Ha I, Baillie DL, Fire A, Ruvkun G, Mello CC: **Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing**. *Cell* 2001, **106**(1):23-34.
160. Lee YS, Nakahara K, Pham JW, Kim K, He Z, Sontheimer EJ, Carthew RW: **Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways**. *Cell* 2004, **117**(1):69-81.
161. Bernstein E, Caudy AA, Hammond SM, Hannon GJ: **Role for a bidentate ribonuclease in the initiation step of RNA interference**. *Nature* 2001, **409**(6818):363-366.
162. Elbashir SM, Lendeckel W, Tuschl T: **RNA interference is mediated by 21- and 22-nucleotide RNAs**. *Genes Dev* 2001, **15**(2):188-200.

163. Hammond SM, Boettcher S, Caudy AA, Kobayashi R, Hannon GJ: **Argonaute2, a link between genetic and biochemical analyses of RNAi**. *Science* 2001, **293**(5532):1146-1150.

164. Martinez J, Patkaniowska A, Urlaub H, Luhrmann R, Tuschl T: **Single-stranded antisense siRNAs guide target RNA cleavage in RNAi**. *Cell* 2002, **110**(5):563-574.

165. Bartel DP: **MicroRNAs: target recognition and regulatory functions**. *Cell* 2009, **136**(2):215-233.

166. Song JJ, Smith SK, Hannon GJ, Joshua-Tor L: **Crystal structure of Argonaute and its implications for RISC slicer activity**. *Science* 2004, **305**(5689):1434-1437.

167. Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP: **The impact of microRNAs on protein output**. *Nature* 2008, **455**(7209):64-71.

168. Baskerville S, Bartel DP: **Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes**. *Rna* 2005, **11**(3):241-247.

169. Friedman RC, Farh KK, Burge CB, Bartel DP: **Most mammalian mRNAs are conserved targets of microRNAs**. *Genome Res* 2009, **19**(1):92-105.

170. Lall S, Grun D, Krek A, Chen K, Wang YL, Dewey CN, Sood P, Colombo T, Bray N, Macmenamin P *et al*: **A genome-wide map of conserved microRNA targets in C. elegans**. *Curr Biol* 2006, **16**(5):460-471.

171. Grun D, Wang YL, Langenberger D, Gunsalus KC, Rajewsky N: **microRNA target predictions across seven Drosophila species and comparison to mammalian targets**. *PLoS Comput Biol* 2005, **1**(1):e13.

172. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M *et al*: **Combinatorial microRNA target predictions**. *Nat Genet* 2005, **37**(5):495-500.

173. Addo-Quaye C, Eshoo TW, Bartel DP, Axtell MJ: **Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome**. *Curr Biol* 2008, **18**(10):758-762.

174. Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC: **Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs**. *Genome Res* 2007, **17**(12):1850-1864.

175. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP: **MicroRNA targeting specificity in mammals: determinants beyond seed pairing**. *Mol Cell* 2007, **27**(1):91-105.

176. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets**. *Cell* 2005, **120**(1):15-20.

177. Jones-Rhoades MW, Bartel DP: **Computational identification of plant microRNAs and their targets, including a stress-induced miRNA**. *Mol Cell* 2004, **14**(6):787-799.

178. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB: **Prediction of mammalian microRNA targets**. *Cell* 2003, **115**(7):787-798.

179. Yang N, Kazazian HH, Jr.: **L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells**. *Nat Struct Mol Biol* 2006, **13**(9):763-771.

180. Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, Xu J, Kittler EL, Zapp ML, Weng Z *et al*: **Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells**. *Science* 2008, **320**(5879):1077-1081.

181. Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, Perrimon N, Kellis M, Wohlschlegel JA, Sachidanandam R *et al*: **An endogenous small interfering RNA pathway in Drosophila**. *Nature* 2008, **453**(7196):798-802.

182. Okamura K, Chung WJ, Ruby JG, Guo H, Bartel DP, Lai EC: **The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs**. *Nature* 2008, **453**(7196):803-806.

183. Kawamura Y, Saito K, Kin T, Ono Y, Asai K, Sunohara T, Okada TN, Siomi MC, Siomi H: **Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells**. *Nature* 2008, **453**(7196):793-797.

184. Volpe TA, Kidner C, Hall IM, Teng G, Grewal SI, Martienssen RA: **Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi**. *Science* 2002, **297**(5588):1833-1837.

185. Sigova A, Rhind N, Zamore PD: **A single Argonaute protein mediates both transcriptional and posttranscriptional silencing in Schizosaccharomyces pombe**. *Genes Dev* 2004, **18**(19):2359-2367.

186. Watanabe T: **Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes**. *Genes & Development* 2006, **20**(13):1732-1743.

187. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T *et al*: **Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes**. *Nature* 2008, **453**(7194):539-543.

188. Ambros V, Lee RC, Lavanway A, Williams PT, Jewell D: **MicroRNAs and other tiny endogenous RNAs in C. elegans**. *Curr Biol* 2003, **13**(10):807-818.

189. Duchaine TF, Wohlschlegel JA, Kennedy S, Bei Y, Conte D, Jr., Pang K, Brownell DR, Harding S, Mitani S, Ruvkun G *et al*: **Functional proteomics reveals the biochemical niche of C. elegans DCR-1 in multiple small-RNA-mediated pathways**. *Cell* 2006, **124**(2):343-354.

190. Lee RC: **Interacting endogenous and exogenous RNAi pathways in Caenorhabditis elegans**. *Rna* 2006, **12**(4):589-597.

191. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM *et al*: **Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes**. *Nature* 2008, **453**(7194):534-538.

192. Siomi MC, Sato K, Pezic D, Aravin AA: **PIWI-interacting small RNAs: the vanguard of genome defence**. *Nat Rev Mol Cell Biol* 2011, **12**(4):246-258.

193. Conine CC, Batista PJ, Gu W, Claycomb JM, Chaves DA, Shirayama M, Mello CC: **Argonautes ALG-3 and ALG-4 are required for spermatogenesis-specific 26G-RNAs and thermotolerant sperm in Caenorhabditis elegans**. *Proc Natl Acad Sci U S A* 2010, **107**(8):3588-3593.

194. Han T, Manoharan AP, Harkins TT, Bouffard P, Fitzpatrick C, Chu DS, Thierry-Mieg D, Thierry-Mieg J, Kim JK: **26G endo-siRNAs regulate spermatogenic and zygotic gene expression in Caenorhabditis elegans**. *Proceedings of the National Academy of Sciences* 2009, **106**(44):18674-18679.

195. Gu W, Shirayama M, Conte D, Vasale J, Batista PJ, Claycomb JM, Moresco JJ, Youngman EM, Keys J, Stoltz MJ *et al*: **Distinct Argonaute-Mediated 22G-RNA Pathways Direct Genome Surveillance in the C. elegans Germline**. *Molecular Cell* 2009, **36**(2):231-244.

196. Claycomb JM, Batista PJ, Pang KM, Gu W, Vasale JJ, van Wolfswinkel JC, Chaves DA, Shirayama M, Mitani S, Ketting RF *et al*: **The Argonaute CSR-1**

and Its 22G-RNA Cofactors Are Required for Holocentric Chromosome Segregation. *Cell* 2009, **139**(1):123-134.

197. Bartel B, Bartel DP: **MicroRNAs: at the root of plant development?** *Plant Physiol* 2003, **132**(2):709-717.

198. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function**. *Cell* 2004, **116**(2):281-297.

199. Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE: **Characterization of the piRNA complex from rat testes**. *Science* 2006, **313**(5785):363-367.

200. Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest AR, Grimmond SM, Schroder K *et al*: **Tiny RNAs associated with transcription start sites in animals**. *Nat Genet* 2009, **41**(5):572-578.

201. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL *et al*: **RNA maps reveal new RNA classes and a possible function for pervasive transcription**. *Science* 2007, **316**(5830):1484-1488.

202. Das M, Harvey I, Chu LL, Sinha M, Pelletier J: **Full-length cDNAs: more than just reaching the ends**. *Physiol Genomics* 2001, **6**(2):57-80.

203. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC: **Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans**. *Nature* 1998, **391**(6669):806-811.

204. Ambros V: **The functions of animal microRNAs**. *Nature* 2004, **431**(7006):350-355.

205. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J *et al*: **Antisense transcription in the mammalian transcriptome**. *Science* 2005, **309**(5740):1564-1566.

206. Jan CH, Friedman RC, Ruby JG, Bartel DP: **Formation, regulation and evolution of Caenorhabditis elegans 3′UTRs**. *Nature* 2010, **469**(7328):97-101.

207. Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH: **Massively parallel sequencing of the polyadenylated transcriptome of C. elegans**. *Genome Res* 2009, **19**(4):657-666.

208. Carninci P: **Tagging mammalian transcription complexity**. *Trends Genet* 2006, **22**(9):501-510.

209. Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, Ecker JR: **Applications of DNA tiling arrays for whole-genome analysis**. *Genomics* 2005, **85**(1):1-15.

210. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G *et al*: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution**. *Science* 2005, **308**(5725):1149-1154.

Chapter 2: The landscape of *C. elegans* 3'UTRs

2.1: Contribution

The primary aim for this chapter in the thesis was to generate a high quality 3'UTRome of an organism and to decipher the biological patterns hidden within. Being a highly collaborative project spanning across many labs and countries, it becomes imperative to specify the contributions and acknowledge the work performed by various researchers. The polyA capture protocol was idealized and conceived by John Kim and Ting Han of University of Michigan. Ting Han developed the protocol for capturing the 3'UTR ends and prepared the libraries to be submitted for sequencing. Pascal Bouffard and Tim Harkins of Roche 454 life sciences performed the pyrosequencing of the libraries. Fabio Piano and Kris Gunsalus of NYU spearheaded the 3' RACE (3' Rapid Amplification of cDNA Ends) capture project and Marco Mangone of Piano Lab performed sequencing of the libraries. The full-length cDNA library was provided by Yutaka Suzuki, Sumiyo Sugano and Yuji Kohara from Japan. While I processed the sequences for polyA captured 3'UTRs, Kris Gunsalus analyzed the sequences from 3'RACE libraries. The centralized consolidation of our sequence libraries and the publicly available RNAseq data was performed by Jean and Danielle Thierry-Mieg at NIH. They also performed the quality check of the various sequences and removal of the false primed sequences. They also provided us with updated gene models including newly identified genes and curated the 3'UTR ends to these new gene models.

Jean Thierry-Mieg performed independent validation of my work including the clustering algorithm and PAS motif analysis. Jean and Danielle Thierry-Mieg also analyzed the cDNA data to identify patterns in SL isoforms linking 5' and 3' end processing. Sebastian Mackowiak did MiRNA target identification and conservation analysis from Rajewsky Lab in MDC Berlin. The 3'UTRome website was maintained by Marco Mangone of NYU. Mitzi Morris of NYU loaded the datasets to modEncode. Nicole Washington of the DCC greatly helped us in making our datasets publicly available. John Kim, Kris Gunsalus, Jean and Danielle Thierry-Mieg, Niklaus Rajewsky and Fabio Piano wrote the manuscript submitted to *Science*, which is provided in sections below.

My contribution to this project was in the processing and analysis of the polyA capture sequence data that we generated. This includes designing an architecture to handle large amount of sequencing data, designing and maintaining databases for the storage of the sequence data, writing custom scripts for removal of linkers in the adapter regions in the sequence, perform quality control of the sequencing, mapping the sequences to the genome, clustering of the 3'UTR ends and annotation of the sequences to existing gene models. In addition to this I also processed the 454 sequencing of the 3' RACE amplicons from NYU. I also performed the PAS motif analysis of the 3'UTRs including identification of the PAS sites, sequence and position distribution and length dependent utilization of the PAS sites. I also performed the computational analysis including developmental stage analysis of 3'UTRs, identification of facultative introns in the 3'UTRs, bidirectional transcription analysis, polyadenylation in histone genes and PAS analysis of the genes in operons excluding the SL-specific PAS usage.

2.2: Abstract

Three-prime untranslated regions (3'UTRs) of metazoan mRNAs contain numerous regulatory elements, yet remain largely uncharacterized. Using polyA capture, 3'RACE, full-length cDNAs, and RNA-seq, we define ~26,000 distinct 3'UTRs in *Caenorhabditis elegans* for ~85% of the 18,328 experimentally supported protein coding genes and revise ~40% of gene models. Alternative 3'UTR isoforms are frequent, often differentially expressed during development. Average 3'UTR length decreases with animal age. Surprisingly, no polyadenylation signal (PAS) is detected for 13% of polyA sites, predominantly among shorter alternative isoforms. Trans-spliced (vs. non-trans-spliced) mRNAs possess longer 3'UTRs and frequently contain no PAS or variant PAS. We identify conserved 3' UTR motifs, isoform-specific predicted microRNA target sites, and polyadenylation of most histone genes. Our data reveal a rich complexity of 3'UTRs genome-wide and throughout development.

2.3: Introduction

The 3'UTRs of mRNAs contain cis-acting sequences that interact with RNA binding proteins and/or small non-coding RNAs (e.g. miRNAs) to influence mRNA stability, localization, and translational efficiency [1-3]. The differential processing of mRNA 3'ends has evident roles in development, metabolism, and disease [4, 5]. Despite these critical roles, genome-wide characterization of 3'UTRs lags far behind that of coding sequences (CDSs). Even in the well-annotated genome of *C. elegans*, nearly half (~47%) of the 20,191 genes annotated in WormBase (release WS190) [6, 7] lack an annotated 3'UTR, and only ~1,180 (~5%) are annotated with alternative 3'UTR isoforms (Fig. 2.S1A,B).

2.4: Results

We have taken a multifaceted, empirical approach to define the 3'UTR landscape in *C. elegans* [Figs. 2.S2-S5, Tables 2.S1-S4, 8]. We prepared developmentally staged cDNA libraries comprising mostly full-length clones spanning from 5'capped first base to polyadenylated (polyA) tail, and annotated 16,659 polyA addition sites in 11,180 genes by manually curating ~300,000 ABI traces in NCBI AceView [9]. We developed a method to capture the 3'ends of polyadenylated transcripts genome-wide by deep sampling and generated a comprehensive developmental profile comprising over 2.5 million sequence reads from Roche/454 (Fig. 2.S2-S5, Tables 2.S1-S4). We cloned 3'RACE products directly targeting 3'UTRs for 7,105 CDSs (6,741 genes) in both the Promoterome [10] and ORFeome [11] collections, and recovered one or more sequenced isoforms for 85% of targets [Figs. 2.S2, 2.S5, Tables 2.S1-S4, 8, 12]. Finally, we remapped and annotated polyA addition sites in published RNA-seq data [13, 14].

All datasets were mapped, cross-validated, consolidated and filtered to eliminate obvious experimental artifacts, including internal priming on A-rich stretches [Figure 1A, 8]. These datasets are not yet saturated: while for most genes (11,516 or 73%), at least one 3'UTR isoform is supported by two or more experimental approaches, 47% of transcripts are observed by only one method [in part due to limitations specific to each protocol, 8] (Fig. 2.1; Tables 2.S3, 2.S4). The resulting 130,090 distinct polyA sites, identified at single nucleotide resolution and supported by over 3 million independent polyA tags, were clustered into 26,967 representative polyA sites. Due to biological variation, 86% of tags occur within 4 nucleotides of representative sites, although individual polyA tags may spread over ~20 nucleotides (Fig. 2.S6).

Linking polyA sites to their parent genes proved to be a challenge, as many previous gene models were incomplete or incompatible with our new data. Using all available empirical evidence, we reannotated in AceView the *C. elegans* gene models [9]. Of the 15,683 protein-coding genes with both polyA sites and cDNA support, 57% confirm the structure of WormBase WS190 gene models. The remainder encode different proteins, usually representing different cDNA-supported splice patterns: ~25% share the same stop codon, ~12% use a different stop (hundreds of those correspond to fusions or splits of prior gene models), and ~6% are not yet annotated in WormBase (Datasets 2.S1, 2.S2).

This integrated collection, herein called the 3'UTRome (Fig. 2.S1, Dataset 2.S2), provides evidence supporting 3'UTR structures for ~74% of all *C. elegans* protein-coding genes in WormBase WS190, including previously unannotated isoforms for ~7,397 genes (Fig. 2.S1A-D). The length distribution of 3'UTRs parallels that in WormBase (Fig. 2.S1D), with a mean of 211 nt (median = 140 nt). The 3'UTRome matches 61% of WormBase 3'UTRs within ±10 nt (6,714 polyA ends for 6,563 genes), and contains thousands of longer or shorter isoforms (Fig. 2.S1A). We identified 6,177 polyA ends for 4,466 genes with no previous 3'UTR annotation and discovered 1,490 polyA ends for 1,031 genes not yet represented in WormBase (Fig. 2.S1A; Datasets 2.S1-S3).

We annotate more than one 3'UTR isoform for 43% of 3'UTRome genes (Figs. 2.S1, 2.S7). Of these, a majority (65%) reflects alternative 3'end formation at distinct locations in the same terminal exon for proteins using the same stop; the remainder use distinct stops in the same last exon or distinct last exons. Very rarely (79 examples), an intron within the 3'UTR is excised or retained (Fig. 2.S8), potentially affecting functional sequence content elements (Fig. 2.S8C). Indeed, putative binding sites for miRNAs (this study) or ALG-1 [15] were identified in the variable regions of some of these transcripts. About 2% of genes possess five or more 3’UTR isoforms (Figs. 2.1A, 2.S1B, 2.S7).

To identify putative cis-acting sequences that may play a role in 3’end formation, we scanned the 50 nt upstream of the cleavage and polyA addition sites for all possible 5- to 10-mers and assigned the most likely PAS motif to each 3’UTR using an iterative procedure based on enrichment and centering of the *k*-mers. The canonical PAS motif AAUAAA (seen in 39% of 3’ends) and many variants differing by 1-2 nt are detected, with distributions all peaking 19 nt upstream of the polyA site [Fig. 2.S9-S10, Table 2.S5, 8]. The canonical signal predominates in genes with unique 3'UTRs (57%). Strikingly, however, many high quality 3’UTRs (3,658) lack a detectable PAS motif altogether (Fig. 2.1B,C). All PAS variants are embedded within a T-rich region that spikes 5 nt downstream of the PAS motif and extends beyond the cleavage site about 20 nt (Fig. 2.1D). 3’UTRs with no PAS tend to be T-rich throughout, except for a very A-rich 8 nucleotide region just after the cleavage site (Fig. 2.1D). Thus, a functional PAS motif with strict sequence specificity appears dispensable for 3’end formation in *C. elegans*.

Among genes with alternative 3’UTRs, successive polyA sites show a striking asymmetry: the longest isoform prefers a PAS, whereas shorter isoforms more often show no PAS (Fig. 2.1C, Fig. 2.S11). The distance between alternative polyA sites peaks at ~40 nt, with resonances at ~80 nt and ~140 nt (Fig. 2.S11A). This regularity suggests that a physical constraint (possibly queuing transcription complexes) could contribute to cleavage and polyA addition at some upstream sites, which may therefore depend less on instructive cues from signal sequences.

Because many *C. elegans* genes undergo trans-splicing of a splice leader (SL) to the 5’end of a nascent transcript [16], we asked whether any properties of transcript 5’ and 3’ ends correlate (Fig. 2.2A,B). About 15% of *C. elegans* genes belong to transcriptional units called operons, each containing 2 to 8 genes that can be co-transcribed, cleaved into separate transcripts, polyadenylated, and trans-spliced with specific leaders (Fig. 2.2A,B). The first gene in an operon is trans-spliced only to SL1; downstream genes are usually trans-spliced to one of 11 other SLs (SL2 to SL12), although we observed that two thirds occasionally become trans-spliced to SL1. The processing of adjacent operon transcript ends (cleavage, polyA addition to the upstream transcript, and SL addition to the downstream transcript) is coupled mechanistically by machinery resembling the cis-splicing apparatus [17]. Comparing 3’UTRs within operons, we observe that the ‘first’ (SL1-spliced), ‘middle’ (any gene between first and last), and ‘last’ genes progressively decrease in average length (from 266 to 213 nt), number of 3’UTR isoforms per gene (from 2.64 to 2.51), and frequency of 3’UTRs with no PAS (from 23% to 18% in ~1,400 sites; Fig. 2.2B).

However, only a small fraction (13%) of the 7,026 mainly SL1-spliced genes clearly belongs to an operon, and these differ notably from non-operon SL1-spliced genes in their usage of the canonical AAUAAA hexamer (22% of 1,409 sites vs. 32% of 10,879 sites, respectively). Furthermore, we observed the canonical PAS motif much more frequently in non-trans-spliced than in SL-containing transcripts (43% of 5,131 sites vs. 30% of 14,873 sites; Fig. 2.2A). While “standard” non-trans-spliced genes have ~30% more 3’UTR isoforms per gene than “isolated” ones having no neighbor within 2 kb (2.4 vs 1.7), these are more similar to each other than to trans-spliced genes – having shorter and fewer 3’UTR isoforms, and higher canonical PAS usage. Thus, trans-splicing within operons appears to enhance (directly or indirectly) the activity of non-canonical PAS sequences upstream, and trans-splicing at the 5’end correlates with distinct properties at the 3’end of the same transcript, independent of 5’end processing downstream.

Unexpectedly, the 3’UTRome reveals polyadenylated transcripts for nearly all histone genes (Fig. 2.S12, Table 2.S6). The major class of replication-dependent histones (H2a, H2b, H3 and H4) are thought not to be polyadenylated in metazoans – instead, their 3'ends form a stem-loop structure that is recognized and cleaved several nucleotides downstream by U7 snRNP and factors such as stem–loop binding protein (SLBP) [18, 19]. *C. elegans* has 61 cDNA-supported histone genes [9] that all harbor conserved sequences with 3’stem-loop potential; however, they also contain conserved PAS elements downstream of the hairpin sequence [20]. Since *C. elegans* histone transcripts also terminate in the typical stem-loop structure and are depleted in successive rounds of polyA selection [20], we were surprised to recover polyadenylated transcripts for 57 histone genes in multiple, independent datasets (Fig. 2.S12, Table 2.S6). This suggests that, at least in *C. elegans* (and perhaps in higher metazoans), the usual route for histone mRNA 3’end processing may include initial cleavage and polyA addition at conserved PAS sites, followed by further processing to remove sequences downstream of the stem-loop.

We searched 3’UTRs for conserved sequence motifs and other potential functional elements. We updated our atlas of predicted conserved miRNA targets for the 3’UTRome, using the PicTar algorithm with new 3-way and 5-way multi-species alignments (Figs. 2.3, 2.S13; Table 2.S7). Roughly half of newly predicted sites match our previous predictions [21], but many sites are gained or lost (Fig. 2.S13A, Table 2.S7). These differences reflect improvements in both 3’UTR annotations and multi-species alignments, which increase the accuracy of conserved seed site identification and signal-to-noise ratios [8]. Over 3,000 PAS motifs are positionally conserved among

Caenorhabditis species, including within alternative 3'UTRs (Fig. 2.S13B). Thus, maintenance of multiple specific 3'termini may be functionally important for some genes. Thousands of unexplained conserved sequence blocks of varying lengths within 3'UTRs (Fig. 2.3B, Table 2.S7) may represent novel functional elements that await further characterization. In vivo Argonaute (ALG-1) binding sites [15] overlap significantly with predicted miRNA target sites but not with other conserved blocks (Table 2.S7), indicating that the latter are, overall, not directly related to microRNA function [8]. For 1,876 convergently transcribed neighboring genes, overlapping 3'regions could pair as dsRNA if co-expressed, potentially triggering endogenous siRNA production [22] that could down-regulate cognate mRNAs (Fig. 2.S14, Dataset S4).

We examined alternative 3'UTR isoforms in different developmental stages (Fig. 2.4). We found a downward trend in average length and number of 3'UTRs per gene from the embryonic through the adult stage (Fig. 2.4A,B). Among genes expressed in more than one developmental stage, embryos display the largest proportion of stage-specific 3'UTR isoforms, and these tend toward longer isoforms (Fig. 2.4B, 2.4C, Table 2.S8, 2.S9, Dataset 2.S5). Some genes switch 3'UTR length coincident with developmental transitions, most notably from embryo to L1, L1 to dauer entry, dauer exit to L4, and in adult hermaphrodites vs. males (Fig 2.4D, Table 2.S9; Datasets 2.S5-S6). Thus, 3'UTR-mediated gene regulation may be widespread in the *C. elegans* embryo, and differential expression of alternative isoforms may represent a mechanism to engage or bypass 3'UTR-mediated regulatory controls in specific developmental contexts [23, 24].

The 3'UTRome compendium evidences support for multiple mechanisms of transcript 3'end formation in *C. elegans*, including standard PAS-directed 3'end formation from a large collection of PAS variants, regularly spaced "shadow" polyA addition sites devoid of recognizable signals, and both operon-dependent and -independent correlations between features at the 5' and 3' ends of the same or of consecutive transcripts that are consistent with the possibility that trans-splicing and 3'end processing within a gene could occur by functionally linked mechanisms. We characterize thousands of new and alternative 3'UTR isoforms throughout development; we define a comprehensive catalog of PAS elements, and discover a surprising number of polyadenylated transcripts with no discernable PAS; and we definitively document polyadenylation of histone transcripts. We also identify conserved sequence elements in 3'UTRs that may interact with trans-acting factors such as miRNAs and RNA-binding proteins, some of which occur within variable regions of alternative 3'UTRs. A collection of cloned 3'UTRs for several thousand *C. elegans* genes is available to the research community for high-throughput downstream analyses and *in vivo* studies [Table 2.S10, Dataset 2.S6, 8].

2.5: Supplementary Materials and Methods

PolyA capture

Strains: Worms were grown on NGM plates seeded with *E. coli* OP50 to adulthood. For collection of staged samples, the wild-type N2 strain was used. Embryos were isolated from gravid worms by standard alkaline/hypochloride treatment [1]. A sample of embryos was frozen down in TriReagent (Ambion, Austin, TX), and the remainder hatched overnight in M9 buffer to yield synchronized L1 stage worms. Starved L1 larvae were plated and fed on NGM plates seeded with OP50 *E. coli* and raised at 20℃. Synchronized staged samples were collected at ~8 hr (L1), ~20 hr (L2), ~30 hr (L3), ~45 hr (L4), and ~70 hr (adult hermaphrodite). The developmental stage of each sample was verified by monitoring the seam cell lineage using Nomarski optics (Olympus, Center Valley, PA). For adult male isolation, the CB1489 *him-8 (e1489)* strain was used, which increases the percentage of XO males to ~37% of the population versus ~0.2% males in the N2 wild-type strain [2]. The *him-8 (e1489)* embryos were synchronized by bleaching and incubated overnight at room temperature. Hatched L1s were aliquoted onto NGM plates seeded with *E. coli* OP50 and grown at 20℃ for 4 days. Male adults were isolated by filtering through 35 µm nylon mesh, resulting in >95% males in the final sample. For dauer larvae preparation, CB1370 *daf-2 (e1370)*, CB1372 *daf-7 (e1372)*, DR47 *daf-11 (m47)*, DR2281 *daf-9 (m540)* mutants from starved plates were collected, resuspended in M9 buffer [1] containing 1% SDS, and incubated for 20 min at room temperature. The suspension was then washed with M9 buffer and worms were placed on a fresh unseeded plate at 20℃ for 1 2 h. Live worms that had crawled away from the dead worms were collected as dauer larvae. Worms were washed off plates with M9, washed 5 times with M9 to remove residual bacteria, and frozen in TriReagent.

RNA preparation: Total RNA was extracted using TriReagent following the vendor's protocol with the following modification: three freeze-thaw cycles (freeze in liquid nitrogen / thaw at room temperature / vortex 1 min) were included to increase worm lysis efficiency; RNA was precipitated with isopropanol at -80℃ for one hour. To subtract 72 most abundant ribosome subunit genes, 25µg total RNAs were mixed with antisense DNA oligos (IDT, Coralville, IA) targeting the last DpnII site of each of these genes and digested with RNaseH (Invitrogen, Carlsbad, CA), which only cleaves RNA in RNA: DNA duplex. After subtraction, PolyA$^+$- selected mRNAs were isolated from total RNA using oligo (dT) magnetic beads (Invitrogen, Carlsbad, CA) using the manufacturer's protocol.

cDNA synthesis: First-strand synthesis was carried out using Superscript III reverse transcription kit (Invitrogen, Carlsbad, CA) with ~20 ng of PolyA$^+$- selected mRNA and 10 pmol of biotinylated reverse primer at 50℃ for 30 min followed by incubation at 42℃ for 30 min. The following biotin -labeled primer was synthesized by Integrated DNA Technologies (Coralville, IA) and PAGE-purified: 5'Biotin-*TAATAC*-GGCGCGCCGCCTTGCCAGCCCGCTCAG-T_{20}-VN-3'. The poly (dT) and two-nucleotide anchor (VN) target the proximal end of the mRNA polyA tail. The second strand was synthesized using DNA polymerase I in the presence of RNase H for 2.5 hr. The double-stranded cDNA product was extracted twice with 200 µL phenol/chloroform/ isoamyl alcohol (25:24:1), ethanol precipitated, and dissolved in 20 µL H_2O.

DpnII digestion: The resulting cDNA was digested with *DpnII* restriction enzyme (New England Biolabs, Ipswich, MA) at 37℃ for 1 hr , extracted twice with 200 µL phenol/chloroform/isoamyl alcohol (25:24:1), and then ethanol precipitated and dissolved in 20 µL H_2O.

Binding biotinylated cDNA to magnetic beads: 100 µL of Streptavidin-Dynabeads M-280 (Invitrogen, Carlsbad, CA) were prepared in a 1.5 mL Eppendorf tube and then washed twice with 1 mL TE (10mM Tris-HCl, PH7.5, 1mM EDTA) and twice with 200 µL 1X B&W buffer (5mM Tris-HCl, PH7.5, 0.5mM EDTA, 1M NaCl). The beads were resuspended in 100-µL 2X B&W buffer (10mM Tris-HCl, PH7.5, 1mM EDTA, 2M NaCl). 10 µL of *DpnII*-digested cDNA fragments and 90 µL H_2O were added to the beads. The tube was rotated for 30 min at room temperature and then the beads were washed twice with 200 µL 1X B&W buffer and twice with 200 µL TE.

Ligation of barcoded linkers to the bound cDNA: Immediately after binding to Dynabeads, cDNAs were ligated to 5 µL Linker A (10 µM) using T4 DNA ligase (Invitrogen, Carlsbad, CA) (5 U/µL) for 2 hr at 16°C with intermittent gentle mixing. The beads were washed twice with 200 µL 1X B&W buffer, washed twice with 200 µL TE, and resuspended in 200 µL TE. Linker A was prepared by annealing the following two complementary oligonucleotides in TE plus 50 mM NaCl: 5'-GCCT-CCCTCGCGCCATCAG-XXXX-3' and 5'-phosphate-*GATC*-XXXX-CTGATGGCGCGAG

GGAGGC-3’, where *GATC* is the *DpnII* restriction sequence and XXXX represents a four-base barcode tag specific to each developmental stage: CATG (embryo), TAGT (L1), GATC (L2), CACT (L3), TACG (L4), or GAGC (adult hermaphrodite).

3’ cDNA recovery: 100 µL beads were mixed with 100-µL phenol/chloroform/isoamyl alcohol (25:24:1), incubated at 65℃ for 30min, vortexed at full speed for 5min, and centrifuged at 15,000 rpm for 5 min. The supernatant was collected using Phase Lock Gel (5PRIME Inc., Gaithersberg, MD). DNA was ethanol precipitated and resuspended in 20 µL H_2O.

PCR amplification: The ligation products from each developmental stage were used as template for two sequential rounds of PCR using 1 µL of DNA, the forward primer set 5’-GCCT-CCCTCGCGCCATCAG-XXXX-3’, and the reverse primer set 5’-GCCTTGCCAGCCCGCTCAG-X-TTTT-X-TTTT-X-TTTT-X-TTTT-3’, where the four Xs represent the four nucleotides of the stage-specific barcode tag distributed in order along a polyA tail. The periodic insertion of the X nucleotides improves reliability of Roche/454 sequencing by decreasing homopolymerization of Ts. Samples were extracted with phenol/chloroform/isoamyl alcohol (25:24:1), ethanol precipitated, and resuspended in 50 µL H_2O. DNA concentration was measured using a Nanodrop 1000 spectrophotometer (Thermo Scientific, Wilmington, DE).

454 GS FLX Sequencing: Deep sequencing was performed on the Genome Sequencer FLX system (Roche/454 Life Sciences, Branford, CT) following the manufacturer's protocol.

<u>3'RACE</u>

***RNA extraction*:** Total RNA from *C. elegans* N2 mixed developmental stages was prepared using an adaptation of the RNeasy Mini kit (Qiagen, Valencia, CA). Worms were grown on NGM plates seeded with *E. coli* OP50, washed with M9 buffer, transferred to an RNase-free Eppendorf tube, and dipped into liquid nitrogen. Worms were ground using RNase-free pestles and incubated with RLT buffer (Qiagen) and beta-mercaptoethanol. The lysate was homogenized by aspiration through a 20-gauge needle fitted to a syringe and centrifuged at 13,000 rpm for 3 min. The supernatant was transferred to RNAse-free tubes and treated as per the manufacturer's recommendations.

Primer Design: Forward primers were designed to target 7,077 CDS-specific regions from WormBase WS150 for CDSs also contained in the Promoterome [3] and the ORFeome [4, 5] collections. For each CDS, in-frame sequence just upstream of and including the STOP codon (based on spliced transcript models) was selected to achieve a T_m of 60℃ ± 5℃ during PCR amplification. Each CDS-s pecific sequence was preceded by the Gateway adaptor 5'-GGGGACAGCTTTCTTGTACAAAGTGGGA-3' to allow recombination into the pDONR P2R-P3 vector (Invitrogen, Carlsbad, CA). The primer list is available at http://www.utrome.org. A universal reverse primer was used, containing a Gateway adaptor (for recombination into pDONR P2R-P3) followed by poly(dT) and a two nucleotide anchor (VN) to target the proximal end of the mRNA polyA tail: 5'-GGGGACAAACTTTGTATAATAAAGTTG-T_{20}-VN-3'. Primers were obtained from Invitrogen.

RT-PCR: Total RNA was incubated at 55℃ for one hour with Superscript III reverse transcriptase (Invitrogen, Carlsbad, CA) and the universal reverse primer according to the manufacturer's specifications. PCR amplification of 3'UTRs from the single-stranded cDNA reaction was performed in 96-well plate format, using, in each well, the universal reverse primer and a different transcript-specific forward primer as follows: denaturation at 94℃ for 30 sec, annealing at 60℃ for 30 sec, extension at 72℃ for 3 min.

***Gateway BP recombination reaction and transformation*:** 3'UTRs were recombined into the pDONR P2R-P3 entry vector using the BP Clonase II Enzyme Mix kit (Invitrogen, Carlsbad, CA) following the manufacturer's specifications and transformed into MultiShot Stripwell TOP 10 plates (Invitrogen, Carlsbad, CA). The transformed bacteria were grown overnight at 37℃ u nder kanamycin selection.

Sanger Sequencing: Aliquots from overnight cultures of 3'UTR minipools were used as templates for PCR with the M13 primer set as follows: denaturation at 94℃ for 30 sec, annealing at 60℃ for 30 sec, extension at 72℃ for 3 min. 7,077 PCR amplicons were sequenced at Agencourt Bioscience Corporation (Beckman Coulter Genomics, Danvers, MA) using the ABI 3700 automated DNA sequencers.

Preparation of deconvolved 3'UTR libraries: 6,912 minipools containing 3'UTR isoforms were manually streaked onto LB kanamycin plates. From each minipool, eight single colonies were manually isolated and propagated as individual 3'UTR clonal isoforms in 96-well plates (for a total of 55, 296 colonies). Liquid aliquots of isolated clones were re-pooled into eight different super-pools using the Aquarius automated multi-channel pipetting system (Tecan Trading AG, Switzerland), resulting in eight libraries that should each contain zero (if no insert was cloned) or one unique 3'UTR isoform per targeted CDS. These deconvolved libraries (labeled A-H) were sequenced using Solexa/Illumina and FLX Roche/454 platforms.

Sample preparation and sequencing with Illumina Genome Analyzer II: Plasmid DNA was recovered using standard alkaline lysis from overnight cultures of the eight deconvolved libraries (A-H). Inserts from each library were amplified by PCR using common Forward (5'-GTTTCTCGTTCAACTTTCTTGTACAAAGTGGGA-3') and Reverse (5'-ATAATGCCAACTTTGTATAATAAAGTTGTTTTTTTTTTT-3') primers. The eight amplicon libraries were purified using MinElute columns (Qiagen), treated to create blunt ends using T4 DNA polymerase (New England Biolabs, Ipswich, MA) and T4 polynucleotide kinase (New England Biolabs), incubated overnight with DNA ligase (New England Biolabs), and then sonicated using the Bioruptor UCD-200 (Diagenode Inc., Sparta, NJ) for 30 min in cycles of 30 sec ON, 30 sec OFF. The resulting 8 fragmented libraries were prepared for Illumina sequencing according to manufacturer's recommendations, and six of the libraries were sequenced using the Illumina Genome Analyzer II system (Illumina, Inc., San Diego, CA) in the Sachidanandam laboratory at the Mount Sinai School of Medicine (New York, NY).

***Sample preparation and sequencing with 454 GS FLX*:** Plasmid DNA was recovered from overnight cultures of the eight deconvolved libraries (A-H) using the Wizard Plus miniprep kit (Promega, Madison, WI) and used as template for PCR amplification with eight barcode-matched primer pairs: AdaptorA::Barcode::Forward (5'-GCCTCCCTCGCGCCATCAG-XXXX-Forward-3') and AdaptorB::Barcode::Reverse (5'-GCCTTGCCAGCCCGCTCAG-XXXX-Reverse-3'), where Forward and Reverse are the same sequences used for Illumina above and barcode tags, XXXX, for libraries A-H are A: CATG, B: TAGT, C: GATC, D: CACT, E: TACG, F: GAGC, G: CTGC, H: ATCG. Barcoded PCR amplicons from all eight libraries were combined and purified using the MinElute PCR purification kit (Qiagen). Because the FLX platform output for samples of variable length is biased toward shorter reads, the combined sample was split into two equal batches: (i) untreated, and (ii) treated with the Agencourt AMPure SPRI PCR purification kit (Beckman Coulter Genomics) to enrich for longer fragments by removing fragments shorter than 100 bp. AMPure library DNA was evaluated for quality and quantified using a BioAnalyzer DNA 1000 lab chip (Agilent, Santa Clara, CA). DNA concentration in ng/µl was converted to molecules/µl and adjusted to 2x10^5 molecules/µl in TE buffer. The resulting fragments were prepared for 454 sequencing according to the manufacturer's recommendations and sequenced using the Genome Sequencer FLX system.

cDNA libraries

Two sets of polyA$^+$-selected cDNA libraries from the Kohara laboratory and prepared from various stages of *C. elegans* development were used (totaling 152,000 cDNA clones).

First, lambda-zap embryonic and *him-8* mixed stage libraries were prepared without any amplification or rationalization steps. These libraries are of very high quality, with ~10^{-4} mismatches per base relative to the genome (after removal of ~200 errors detected in the genome) and less than 3% structural defects or artifacts.

The second set consists of full-length L1, L2, L4 and mixed stage libraries prepared by S. Sugano Y. Suzuki and Y. Kohara using the oligo cap selection procedure [6]. These libraries were designed to include the entire transcript, from 5' capped first base to poly A, and are validated by the fact that >99% of the clones with a *trans*-spliced leader in this collection contain the entire leader sequence (21 to 23 bases long). These collections allowed identification of 12 varieties of SL as well as 3,953 genes that are not *trans*-spliced.

Sequencing traces from a polyA$^+$-selected library (n=14,811 cDNA clones), generously provided by Exelixis Inc. (San Francisco, CA), along other publicly available cDNAs and EST data obtained from the NCBI Trace and dbEST archives (in the form of either sequences or traces), were also manually curated at NCBI as part of the experimentally supported worm transcriptome project known as AceView [7].

The combined cDNA dataset provides experimental evidence for 16,659 distinct polyA sites in 11,180 genes. These data are all publicly available from http://www.aceview.org and http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly.

RNA-Seq datasets: Illumina data for staged samples (L2, L3, and L4 larvae and young adults) from the modENCODE transcriptome project, described in [8], were obtained from NCBI GEO (SRX001872-SRX001875). Additional published Roche/454 datasets for the L1 stage [9] were also analyzed. Together, these data provide support for 8,332 polyA sites for 7,461 genes.

Sequence analysis of primary datasets

Genome version: All data were aligned to *C. elegans* genome sequence version CE6 (on which WormBase WS190 gene annotations are also anchored).

PolyA capture libraries: 454 sequence data from three independent runs were pooled. Runs A and B comprised sequences from combined staged samples (Run A: embryo, L1-L4, adult hermaphrodite; Run B: embryo, L1-L4, adult hermaphrodite, adult male); Run C contained mixed sequences from four dauer mutants: *daf-2*, *daf-7*, *daf-9*, and *daf-11* (see Table S2 for read counts from each run). Forward reads were identified by the pattern 5'-XXXX-*GATC*-N_m-X´-AAAA-X´-AAAA-X´-AAAA-X´-AAAA-3', where *GATC* is the DpnII restriction site, N_m is a sequence of length *m* extending from the DpnII site to the end of the 3'UTR, and X´X´X´X´ is the reverse complement of the matching 3'end barcode. Reads that did not contain a decipherable barcode tag were discarded. Barcodes were used to identify the library of origin for the remaining reads, and sequences were processed to remove the 5' and 3' adaptor sequences and barcode tags. Sequences retaining length ≥15 nt were aligned to the genome using BLAT [10], with a maximum intron size of 1000, minimum window size of 5, and maximum gap of 6. Best matches were selected, and multiple alignments reported if present in more than one genomic location. Alignments in PSL format were converted to SAM format using the psl2sam.pl script provided with SAMtools [11]. Alignments for sequences that did not reach the polyA were set aside; the remaining alignments were further annotated.

3'RACE: RACE clones were sequenced by three different methods. Sequences from ABI or SCF files were trimmed of vector sequence and filtered for empty vectors and putative primer-dimer products. The remaining sequences were aligned to the genome using BLAT [10] and WU-BLAST 2.0 [12]. Aligned regions were scanned for the presence of detectable CDS-specific primer and terminal polyA sequences (defined as 10 or more consecutive As with zero or one intervening nucleotide).

For Illumina data, 50 million sequence reads from six independently sequenced libraries were aligned to both the genome and to AceView transcripts using the AceView aligner (http://www.ncbi.nlm.nih.gov/ IEB/Research/Acembly/Software). PolyA sites were identified by trimming reads beginning with at least 5 consecutive T's or ending in at least 5 consecutive A's, and then mapping either the full remaining tag sequence or a version lacking the last two nucleotides upstream of the polyA (since we had previously determined that the cloned RACE products contained a high proportion of T to C base changes at these positions, which pair with the two anchor nucleotides in the universal reverse primer; data not shown). Overlapping mapped reads were assembled into contigs, and these were used for further annotation.

From the two 454 runs, a total of ~170,000 reads corresponding to ~85,000 unique sequences were produced. Initial processing, BLAT alignment, and conversion to SAM format were the same as described above for polyA capture data.

Alignments from all three platforms were then considered together and, where possible, alignments were assigned to the putative plate-well of origin based on the identity of the corresponding primer; for deconvoluted libraries, the combination of primer and barcode, if detectable, was used to assign a putative location in the isolated clone library plates.

cDNAs and ESTs: cDNA clones from the yk collection were sequenced using the Sanger method. All cDNA and EST data from this collection and from other sources (as described above) were aligned to the genome and annotated using AceView tools; these were further hand-curated by visual inspection of multiply aligned ABI sequence traces, where available.

RNA-seq: Illumina and Roche/454 datasets (described above) were aligned to both the genome and AceView transcripts using the AceView software tools (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/Software). PolyA sites were identified trimming reads beginning with at least 5 consecutive T's or ending in at least 5 consecutive A's, and mapping the remaining sequence tag as above. Overlapping mapped reads were assembled into contigs, and these were used for further annotation.

cDNA and transcriptome annotation

Annotation of independent datasets: Sequences from RACE and polyA capture with best-hit alignments or assembled contigs near the last exon of a (targeted, for RACE clones) CDS were defined as candidate 3'USTs (UTR Sequence Tags). USTs were initially assigned to the overlapping or immediately adjacent upstream CDSs from WormBase WS190 gene models (http://www.wormbase.org); these assignments were later revised using AceView genes (http://www.aceview.org), which in some cases revealed that the combined data were incompatible with existing WS190 (or WS150) CDS models. In such cases, USTs from RACE experiments were retained as evidence of transcriptional activity but were removed from the final list of cloned 3'UTRs. USTs with a contiguous BLAT alignment extending through the STOP codon of a valid AceView CDS model and containing polyA sequences were considered to be bona fide complete 3'UTR isoforms with full-length coverage. Those with incomplete 3'UTR coverage and/or no detectable polyA sequence were annotated as partial 3'USTs and used to refine 3'UTR boundaries. Mapped tags from short read data were assembled into contigs and used together with cDNA, EST, UST data to define transcribed regions. The combined data were used to refine and extend existing AceView genes. Data mapping downstream of (but not overlapping) an existing gene were extended *in silico*, where possible, and assigned to the nearest gene upstream or else used to define new transcriptional units. All annotated 3'USTs and 3'UTRs were used for subsequent analyses.

Definition of representative polyA sites and 3'UTR isoforms: To define 3'UTR isoforms and assign a single representative polyA site per isoform, we combined evidence for polyA addition sites from all four independent data sources in the 3'UTR compendium into a single large dataset.

To define the 3'UTRs, I performed a custom written iterative local clustering procedure using the chromosomal coordinates and abundance of the 3'UTR ends. I used a 20 nucleotide window to scan across the genome from left to right and for each strand of the chromosome separately and looked for neighboring 3'UTR ends within the window. All 3'UTR ends within the window were annotated to a cluster and the most abundant 3'UTR end was made the representative end of the cluster and its abundance was calculated as the sum of the individual abundances. If multiple 3'UTR ends in a cluster had the same maximum abundance then one of them was chosen as the representative at random. This was performed iteratively and if the representatives of two neighboring clusters are within 20 nucleotides of each other they were clustered into one. This recursion continued till no further clustering was possible.

A parallel clustering implementation was performed by Jean-Thierry Mieg to validate my results and their clustering software is included in the AceView software, available from http://www.ncbi.nlm.nih.gov/IEB/Research/ Acembly/Software. When evidence sources were attached to a known gene model, clustering was performed on a per-gene basis. The local maximum for each cluster was computed and used as the position of the reported ("representative") polyA addition site for each 3'UTR isoform. The spread of the clusters extends from one up to around 20 nucleotides, with 86% of all individual data points falling within 4 nt of the representative polyA site (Fig. 2.S6).

Using this clustering procedure, each 3’UTR isoform was then defined as a unique sequence span that extends from a specific CDS end and terminates downstream at a distinct “canonical” polyA addition site: 3’UTR sequences that share the same CDS end and terminate within the same polyA cluster were defined as examples of the same isoform, whereas 3’UTR sequences that terminate within different polyA clusters (even if linked to the same CDS) were defined as distinct isoforms. Isoforms of a gene that were represented by less than 5% of the total polyA counts for that gene, isoforms that were not supported by two or more independent pieces of evidence, and those that were shorter than 20 nt (which mostly contained dubious cloning artefacts) were removed from the final dataset. For reporting purposes and all downstream analyses involving isoforms, we considered only the “representative” polyA coordinate for each reported 3’UTR isoform.

Identification of PAS sites: The 50 nt regions immediately upstream of all polyA sites were scanned in an unbiased way for all possible 5 to 10-mer sequences to identify any statistically over-represented motifs. The only motifs returned from this exercise were the canonical PAS sequence (AAUAAA) and several closely related sequences. The distribution of all over-represented hexamers peaked at a start position of -19 nt from the polyA site, which was taken as the most likely position of the PAS site. All of the 3'UTR isoforms in the compendium were then scanned for the canonical PAS sequence and any hexamer with an edit distance of 1 or 2 nt. Because it is not possible to definitively identify the "real" PAS site, we scanned for hexamers in a preferred order based on their observed frequency of occurrence in 3'UTRs between 10 and 30 nt upstream of the polyA site, and considered those occurring at a frequency of ≥1% as putative PAS motifs. We used the first occurrence of a putative motif in the ordered list as the most likely functional PAS sequence. UTRs that did not contain one of the resulting 26 putative PAS motifs within this interval were termed “no PAS”.

Analysis of genomic nucleotide frequencies in the 120 nt region spanning ±60nt of polyA sites showed that strongly supported PAS sites, which we consider the best candidates for recognition by CPSFs for 3'end-processing [13], also show an enrichment of T's that peaks at +5 nt downstream of the putative PAS site (Fig. 1D). These include nine principal motifs: AATAAA (the canonical PAS hexamer), AATgAA, tATAAA, cATAAA, gATAAA, AtTAAA, tATgAA, AgTAAA and cATgAA (where upper-case letters are identical with the canonical hexamer, and lower-case letters indicate substitutions).

Comparison of 3'UTRome and WormBase annotations: Operon, Gene, CDS, and 3'UTR annotations for WS190 were obtained from WormBase. For comparative purposes, any 3'UTR in our compendium whose 5'end matched a WS190 CDS and whose 3'end was within 10 nt of an annotated WS190 3'UTR was considered identical; all others were labeled as "longer" or "shorter" than the WS190 3'UTR, as appropriate. 3'UTRs in our dataset that matched a WS190 CDS end but had no corresponding WS190 3'UTR were annotated as "new 3'UTRs". 3'UTRs that did not match a WS190 gene model, but matched an alternate transcript model that could be generated from experimental data, were annotated as 3'UTRs of "new AceView genes". These data are summarized in Fig. 2.S1.

Intron analysis: Gapped sequence alignments were examined for the presence of putative splice signal consensus sequences, and introns were annotated as appropriate. Numerous gapped alignments of polyA capture data spanned bona fide splice junctions but were on the opposite strand and thus contained the reverse complement of known splice consensus signals. Such alignments were observed to occur most frequently within coding regions; these were determined most likely to represent mis-priming in A-rich regions and were discarded. A subset of gapped alignments for these data contained terminal segments <10 nt; these appeared to be alignment artifacts of degraded sequence data and were also discarded. A total of 363

3’UTRs for 192 genes were determined to contain bona fide introns, based on the presence of a strongly supported CDS upstream with no evidence for another CDS that could extend into the putative 3’UTR. The 3’UTRs with an intron that could also occur internally within the CDS of an alternative isoform were not counted in this set.

Operon and SL analysis: To compare the six categories of genes analyzed in Fig. 2, we selected a subset of trans-spliced and non-trans-spliced genes for which assignment to a unique category could be unambiguously determined. Among the SL1 trans-spliced genes, we identified 574 SL1 genes occupying the first position of an operon (genes fully supported from SL1 to polyA and separated by at most 300 bases from the next gene in cis, which is itself trans-spliced mostly to SL2) and 3,530 SL1-genes undoubtedly not in an operon (selected as followed either by another SL1-gene (n=1,749) or by a confirmed non-transspliced gene (n=1781)); these two subsets were found to be indistinguishable and were merged in Fig. 2.

Directed RT-PCR assay for retained 3’UTR introns: Total RNA was extracted from mixed-stage worms and RT-PCR was performed essentially as described above. 1 µg of total RNA from mixed-stage worms was used as template for a first strand reaction using the universal anchored poly(dT) reverse primer. PCR was performed using internal primer pairs flanking putative retained introns in the 3’UTRs of two genes: *par-5* (Forward: 5’-GAG GGA AAC CAG GAA GCT GGA AAC TAA-3’; Reverse: 5’-GAT GCT ATT GCG CAG TGT TGT ATG GAG TAT TGG) and *sams-1* (Forward: 5’-GCC ACA TCT GCT ATC GCT CAC TAA-3’; Reverse: 5’-CAA GAC AGC TCA GCG GGT AGC GGA AAC CG-3’). Products were separated on a 1% agarose gel and visualized with ethidium bromide.

Developmental stage analysis: The staged polyA capture dataset was used for this analysis, since this dataset can provide specific information on the abundance of alternative 3’ends expressed in different stages. Since the total polyA tag count differed between libraries, the total number of read counts from each stage was normalized to match the total counts in embryo, and counts for individual isoforms scaled proportionally to reflect the relative expression level in different life stages. The number of isoforms detected per gene was evaluated for each developmental stage and across all stages. To study the expression of long vs. short isoforms we identified genes showing exactly two distinct 3'UTR isoforms (2,295 in total) and restricted our analysis to a stringent subset of 1,960 genes showing at least 5 read counts for the most abundant isoform (Supplementary Dataset S5). To identify genes showing preferential isoform usage, we further selected a subset of genes that showed, in the cumulative dataset, at least twice as many total counts for one isoform as the other (915 genes for long>short; 615 genes for short>long). The per-stage relative expression of a particular isoform of a gene was calculated by dividing the counts for that isoform by the total counts for both isoforms expressed during that stage. The relative expression of an isoform across all stages was calculated as the ratio of the normalized counts of the isoform in a single stage to the total normalized counts of both isoforms of the gene across all developmental stages.

To identify genes that exhibit a differential preference for 3'UTR isoforms during development (i.e. 3'UTR isoform "switching"), we filtered the 1,960 genes described above using the following criteria: 1) isoform *'a'* was more abundant than the isoform '*b*' in one developmental stage, and isoform '*b*' was more abundant than isoform '*a*' in any other developmental stage; 2) the total abundance of all isoforms for the same was ≥ 20 counts (abundance was based on normalized polyA capture counts). We identified 612 genes exhibiting such 3'UTR isoform switching (see Supplementary Datasets S5, S6). To obtain a "high-confidence" subset of these genes, we imposed two additional criteria: 1) the ratio of counts for isoform '*a*' to counts for isoform '*b*' (a/b) was ≥2 fold in one stage, and the ratio of isoform ‘*b*’ to isoform ’*a*’ counts (b/a) was ≥2 fold in another stage; 2) the difference in support between isoform ‘*a*’ and ‘*b*’ was ≥5 counts within each developmental stage in which switching occurred. Of the 612 genes, 263 genes passed these filters (see Supplementary Datasets S5, S6).

miRNA target prediction and 3’UTR conservation analysis

***3'UTR alignments*:** We used the Galaxy server processing pipeline [14] and the UCSC Table Browser [15] to prepare a multiple alignment file (MAF) for *C. elegans* (WS190/CE6), *C. remanei, C. briggsae, C. brenneri*, and *C. japonica*. The MAF file did not contain overlapping blocks or gaps in the *C. elegans* sequence. We then extracted a MAF file for each of the initial 33,909 3'UTRs from the 3'UTRome. Overlapping 3’UTRs were fused to yield 15,685 unique 3'UTR regions that were used for subsequent analyses.

***miRNA sequences*:** We used for our analyses 174 *C. elegans* mature miRNA sequences downloaded from miRBase version 14 [16] and 9 novel miRNAs determined by miRDeep2 [17]. These miRNAs were grouped into 124 miRNA families sharing the same seed sequence at nucleotides 2-7 in each miRNA.

Identification of miRNA seeds in 3’UTRs: The PicTar algorithm [18, 19] was used to identify non-conserved and conserved miRNA seeds in mRNA sequences, which were defined as regions in mRNA sequences with perfect base complementarity to miRNA 6-mer seeds (nucleotides 1-6 or 2-7 at the miRNA 5' end). Seeds conserved in 3 species (*C. elegans, C. remanei*, *C. briggsae*) and those conserved in 5 species (*C. elegans, C. remanei, C. briggsae, C. brenneri*, *C. japonica*) were identified. PicTar was further used to predict and assign scores for full miRNA binding sites, as described [19]. The probabile number of conserved predicted miRNA target seed site being functional in 3-way or 5-way species comparisons is 2.7 and 3.1, respectively. The comprehensive list of PicTar predictions is available from the UTRome (http://www.utrome.org) and modENCODE (http:// www.modencode.org) websites.

Comparison with Lall et al., 2006: We compared our updated miRNA target predictions within our previous predictions for *C. elegans [19]*. For this comparison, we considered only those miRNAs that were analyzed in Lall et al. and the set of unique (non-overlapping) 3'UTRs contained in the UTRome to which the Lall et al. target site predictions map; thus, any predicted sites from either study that were not contained in UTRs considered in the other study were not included in this comparison. In addition, we excluded from the comparison the two miRNAs cel-miR-68 and cel-miR-69 used in the Lall et al. analysis (because they are currently annotated as siRNAs in WormBase), and the seven miRNAs cel-miR-42, cel-miR-239b, cel-miR-248, cel-miR-250, cel-miR-252, cel-miR-253 and cel-miR-358 (because the reported sequences of their seed regions, i.e. positions 1-7 or 2-8 in the mature miRNA, were different according to Rfam version 6 and miRBase version 14).

We then compared the number of predicted sites from this study with the previous set of predictions within the sequence space analyzed in both studies (summarized in Table S7). From our new prediction set, 5,943 predicted miRNA target sites fall in this intersecting sequence space, of which 580 sites (9.8%) were not identified in the Lall et al. study. We attribute the identification of these new sites to improved multi-species alignments and the inclusion of newly sequenced species in the alignments.

Of the 11,131 miRNA target sites predicted in the Lall et al. study, 6,474 sites were located in the intersecting sequence space. In the current study, we recovered 5,363 of those sites, or 82.8%; the remaining 1,111 sites from Lall et al. (17.2%) could not be recovered. The loss of these sites is explained by the fact that the Lall et al. study used some sequence regions outside the 3'UTRome for the initial predictions; if conserved sites were identified in these regions, then non-conserved sites falling within shorter 3'UTRs would also be designated as candidate target sites due to the presence of the initial conserved site. However, if this sequence region is not used for the initial identification, and no other conserved sites are identified within the sequence space analyzed, then non-conserved sites will not be considered by the algorithm as potential target sites, and previously predicted sites would then be lost.

We note that many previously predicted target sites from Lall et al. that fall outside the spans of our 3'UTR annotations (either because they targeted genes for which we have no 3'UTR annotation, or because we previously used up to 500nt spans downstream of any CDS if no 3'UTR was available) are not currently supported by empirically defined 3'UTR regions.

Conserved blocks not explained by miRNA seeds: To identify conserved sequence blocks that do not correspond to conserved miRNA seed sequences, all (reverse complemented) miRNA seeds were masked with Ns in the 3'UTR multiple alignment files (MAFs), and all remaining k-mers ($k \geq 6$) conserved in 3 species (*C. elegans, C. remanei, C. briggsae*) or in 5 species (*C. elegans, C. remanei, C. briggsae, C. brenneri, C. japonica*) were identified. The alignment of any conserved 6-mer was extended as far as possible in both directions.

Distribution of conserved PAS motifs and sequence blocks: We excluded from this analysis all 3'UTRs shorter than 10 nt and those contained within coding sequences of alternative CDSs, resulting in a final set of 24,858 3'UTRs, of which 8,319 genes have a single isoform, 3,320 genes have exactly two isoforms, and 2,616 have more than two isoforms. All conserved miRNA seeds in 3'UTRs, all 29 putative PAS motifs, and all conserved sequence blocks as defined above were investigated with respect to their positions relative to UTR ends. A PAS site was considered as "conserved" in this analysis if it was found in *C. elegans* and the same or another PAS motif was found within a window of ±5 nucleotides in aligned *C. briggsae* and *C. remanei* sequences. Only PAS sites in genes with one isoform or exactly two isoforms, where the longest isoform was at least 100 nt, were considered. The set of genes with 2 isoforms was further filtered to require a length difference of at least 50 nt between the short and long isoform; if this requirement was not met, the short isoform was discarded and the gene was treated as having a single long isoform for this analysis.

Analysis of overlaps between experimentally determined ALG-1 binding sites and conserved sequence motifs: We compared recently published in vivo Argonaute (ALG-1) binding sites [20] with our conserved sequence motifs (predicted miRNA target sites and conserved sequence blocks). For this analysis we considered only those 3'UTRs containing or overlapping at least one ALG-1 binding site. The probability of predicted miRNA target seed sites from 3-way species alignments (*C. elegans, C. briggsae, C. remanei*) occurring within an ALG-1 binding site was 0.75. As a control, we calculated the overlap between ALG-1 sites and 6-mers (the length of predicted miRNA seed sites) placed at random positions along the length of annotated 3'UTRs (p=0.43), which represents a lower bound to the resolution at which we could discern meaningful correlations with ALG-1 sites. The overlap was not significant for the thousands of other conserved blocks that are not explained by predicted miRNA target sites or by conserved PAS sites (0.54 vs. 0.48 for random controls). These results indicate that the overlap between ALG-1 sites and predicted miRNA target sites is highly significant, and that while other conserved sequence blocks are likely functional, they are not, overall, directly related to microRNA function.

Data Availability

Raw data from Roche/454 and Illumina sequencing were deposited at NCBI Short Read Archive (accession numbers: GSM443959-GSM443964, GSM446651-GSM446661, GSM469439, GSM469976) and GEO (accession number: GSE17781). ABI traces and UST sequences were deposited in the NCBI Trace (trace IDs: 2216286010-2216288816) and dbEST (dbEST IDs: 63366486-63366494) archives. Genome alignments and annotations for 3'UTRs, polyA sites, and PAS sites were deposited with the modENCODE DCC (accession numbers: 515, 896, 992, 2327-2337, 2455-2465, 2482, 2484, 2501 and 2745), along with metadata describing experimental and bioinformatic protocols and links to raw datasets in NCBI public repositories. See also Datasets S1-S7. Multiple web portals will provide access to 3'UTRome data, including UTRome.org, AceView.org, modENCODE.org, and WormBase.org.

2.6: Reference

1. de Moor CH, Meijer H, Lissenden S: **Mechanisms of translational control by the 3' UTR in development and differentiation**. *Semin Cell Dev Biol* 2005, **16**(1):49-58.
2. Wickens M, Bernstein DS, Kimble J, Parker R: **A PUF family portrait: 3'UTR regulation as a way of life**. *Trends Genet* 2002, **18**(3):150-157.
3. Bartel DP: **MicroRNAs: target recognition and regulatory functions**. *Cell* 2009, **136**(2):215-233.
4. He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, Goodson S, Powers S, Cordon-Cardo C, Lowe SW, Hannon GJ *et al*: **A microRNA polycistron as a potential human oncogene**. *Nature* 2005, **435**(7043):828-833.
5. Chatterjee S, Pal JK: **Role of 5'- and 3'-untranslated regions of mRNAs in human diseases**. *Biol Cell* 2009, **101**(5):251-262.
6. Stein L, Mangone M, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J: **WormBase: network access to the genome and biology of Caenorhabditis elegans**. *Nucleic Acids Res* 2001, **29**(1):82-86.
7. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, De La Cruz N, Davis P, Duesbury M, Fang R *et al*: **WormBase: a comprehensive resource for nematode research**. *Nucleic Acids Res* 2010, **38**(Database issue):D463-467.
8. Materials SO.
9. Thierry-Mieg D, Thierry-Mieg J: **AceView: a comprehensive cDNA-supported gene and transcripts annotation**. *Genome Biol* 2006, **7 Suppl 1**:S12 11-14.
10. Dupuy D, Li QR, Deplancke B, Boxem M, Hao T, Lamesch P, Sequerra R, Bosak S, Doucette-Stamm L, Hope IA *et al*: **A first version of the Caenorhabditis elegans Promoterome**. *Genome Res* 2004, **14**(10B):2169-2175.
11. Reboul J, Vaglio P, Rual JF, Lamesch P, Martinez M, Armstrong CM, Li S, Jacotot L, Bertin N, Janky R *et al*: **C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression**. *Nat Genet* 2003, **34**(1):35-41.
12. Mangone M, Macmenamin P, Zegar C, Piano F, Gunsalus KC: **UTRome.org: a platform for 3'UTR biology in C. elegans**. *Nucleic Acids Res* 2008, **36**(Database issue):D57-62.
13. Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH: **Massively parallel sequencing of the polyadenylated transcriptome of C. elegans**. *Genome Res* 2009, **19**(4):657-666.
14. Shin H, Hirst M, Bainbridge MN, Magrini V, Mardis E, Moerman DG, Marra MA, Baillie DL, Jones SJ: **Transcriptome analysis for Caenorhabditis elegans based on novel expressed sequence tags**. *BMC Biol* 2008, **6**:30.
15. Zisoulis DG, Lovci MT, Wilbert ML, Hutt KR, Liang TY, Pasquinelli AE, Yeo GW: **Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans**. *Nat Struct Mol Biol* 2010, **17**(2):173-179.
16. Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu WL, Duke K, Kiraly M *et al*: **A global analysis of Caenorhabditis elegans operons**. *Nature* 2002, **417**(6891):851-854.
17. Liu Y, Huang T, MacMorris M, Blumenthal T: **Interplay between AAUAAA and the trans-splice site in processing of a Caenorhabditis elegans operon pre-mRNA**. *RNA* 2001, **7**(2):176-181.

18. Wang ZF, Whitfield ML, Ingledue TC, 3rd, Dominski Z, Marzluff WF: **The protein that binds the 3' end of histone mRNA: a novel RNA-binding protein required for histone pre-mRNA processing**. *Genes Dev* 1996, **10**(23):3028-3040.
19. Marzluff WF, Wagner EJ, Duronio RJ: **Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail**. *Nat Rev Genet* 2008, **9**(11):843-854.
20. Keall R, Whitelaw S, Pettitt J, Muller B: **Histone gene expression and histone mRNA 3' end structure in Caenorhabditis elegans**. *BMC Mol Biol* 2007, **8**:51.
21. Lall S, Grun D, Krek A, Chen K, Wang YL, Dewey CN, Sood P, Colombo T, Bray N, Macmenamin P *et al*: **A genome-wide map of conserved microRNA targets in C. elegans**. *Curr Biol* 2006, **16**(5):460-471.
22. Okamura K, Balla S, Martin R, Liu N, Lai EC: **Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila melanogaster**. *Nat Struct Mol Biol* 2008, **15**(6):581-590.
23. Lund E, Liu M, Hartley RS, Sheets MD, Dahlberg JE: **Deadenylation of maternal mRNAs mediated by miR-427 in Xenopus laevis embryos**. *RNA* 2009, **15**(12):2351-2363.
24. Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, Enright AJ, Schier AF: **Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs**. *Science* 2006, **312**(5770):75-79.

Acknowledgements: Supported in part by grants from NIH (U01-HG004276) to F.P., K.C.G., J.K.K., N.R.; KAKENHI from the Ministry of Education, Culture, Sports, Science and Technology of Japan to Y.K., S.S., Y.S.; M.D.A and the Pew Charitable Trusts to J.K.K.; and by the Intramural Research Program of NIH, National Library of Medicine to J.T-M., and D.T-M. We thank J.V. Moran, T. Blumenthal, A. Billi, D. Mecenas, and B. Bargmann for discussions; T. Shin'I, and Exelixis Inc. for *C. elegans* cDNA traces; for technical assistance, T. Nawy and B. Brown (statistical analysis), R. Sachidanandam, R. Lyons and S. Genik (deep sequencing), P. MacMenamin and D. Schaub (3'UTRome database), M. Morris (data submission), L. Huang (stage analysis). 3'UTRome datasets are available from NCBI Trace Archive, dbEST, SRA, GEO, and from modENCODE [8]. See Materials and Methods for details. Annotations are displayed at NCBI AceView (http://www.aceview.org) [9] and http://www.UTRome.org [12].

2.7: Reference for Supplementary Online Materials

1. Stiernagle T: **Maintenance of C. elegans**. *WormBook* 2006:1-11.
2. Hodgkin J, Horvitz HR, Brenner S: **Nondisjunction Mutants of the Nematode CAENORHABDITIS ELEGANS**. *Genetics* 1979, **91**(1):67-94.
3. Dupuy D, Li QR, Deplancke B, Boxem M, Hao T, Lamesch P, Sequerra R, Bosak S, Doucette-Stamm L, Hope IA *et al*: **A first version of the Caenorhabditis elegans Promoterome**. *Genome Res* 2004, **14**(10B):2169-2175.
4. Vaglio P, Lamesch P, Reboul J, Rual JF, Martinez M, Hill D, Vidal M: **WorfDB: the Caenorhabditis elegans ORFeome Database**. *Nucleic Acids Res* 2003, **31**(1):237-240.
5. Reboul J, Vaglio P, Rual JF, Lamesch P, Martinez M, Armstrong CM, Li S, Jacotot L, Bertin N, Janky R *et al*: **C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression**. *Nat Genet* 2003, **34**(1):35-41.
6. Suzuki Y, Yoshitomo-Nakagawa K, Maruyama K, Suyama A, Sugano S: **Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library**. *Gene* 1997, **200**(1-2):149-156.
7. Thierry-Mieg D, Thierry-Mieg J: **AceView: a comprehensive cDNA-supported gene and transcripts annotation**. *Genome Biol* 2006, **7 Suppl 1**:S12 11-14.
8. Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH: **Massively parallel sequencing of the polyadenylated transcriptome of C. elegans**. *Genome Res* 2009, **19**(4):657-666.
9. Shin H, Hirst M, Bainbridge MN, Magrini V, Mardis E, Moerman DG, Marra MA, Baillie DL, Jones SJ: **Transcriptome analysis for Caenorhabditis elegans based on novel expressed sequence tags**. *BMC Biol* 2008, **6**:30.
10. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656-664.
11. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
12. Lopez R, Silventoinen V, Robinson S, Kibria A, Gish W: **WU-Blast2 server at the European Bioinformatics Institute**. *Nucleic Acids Res* 2003, **31**(13):3795-3798.
13. Murthy KG, Manley JL: **Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus**. *J Biol Chem* 1992, **267**(21):14804-14811.
14. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J: **Galaxy: a web-based genome analysis tool for experimentalists**. *Curr Protoc Mol Biol* 2010, **Chapter 19**:Unit 19 10 11-21.
15. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool**. *Nucleic Acids Res* 2004, **32**(Database issue):D493-496.
16. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature**. *Nucleic Acids Res* 2006, **34**(Database issue):D140-144.

17. Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N: **Discovering microRNAs from deep sequencing data using miRDeep**. *Nat Biotechnol* 2008, **26**(4):407-415.

18. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M *et al*: **Combinatorial microRNA target predictions**. *Nat Genet* 2005, **37**(5):495-500.

19. Lall S, Grun D, Krek A, Chen K, Wang YL, Dewey CN, Sood P, Colombo T, Bray N, Macmenamin P *et al*: **A genome-wide map of conserved microRNA targets in C. elegans**. *Curr Biol* 2006, **16**(5):460-471.

20. Zisoulis DG, Lovci MT, Wilbert ML, Hutt KR, Liang TY, Pasquinelli AE, Yeo GW: **Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans**. *Nat Struct Mol Biol* 2010, **17**(2):173-179.

2.8: Figures

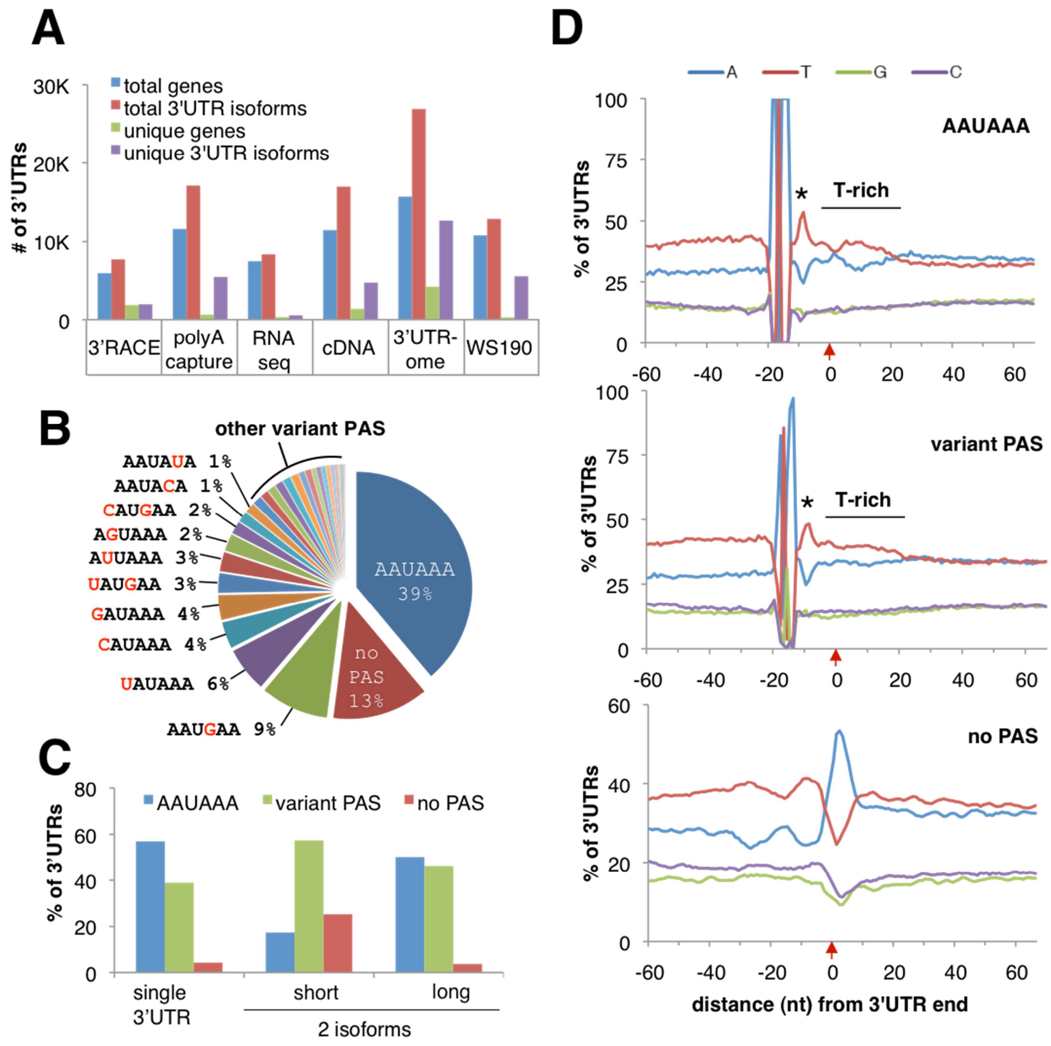


Figure 2.1: The 3'UTRome and 3'UTR polyadenylation signals.
A) The number of genes and isoforms detected in, or specific to, each dataset, and cumulative totals in WS190 and 3'UTRome annotations. B) PAS motif frequencies: AAUAAA (39%), variant PAS (1-9%), no PAS (13%). C) PAS usage in genes with one or two (short and long) 3'UTR isoforms. D) Nucleotide distribution spanning ±60 nt around the polyA addition site, in 3'UTRs with: AAUAAA (top), ten most common variant PAS (middle), no PAS (bottom). Alignments, centered at -19nt, show T-spike at 5 nt downstream of PAS (asterisk), polyA addition site (red arrow), and T-rich region downstream of cleavage site. The A-rich peak downstream of "no PAS" is not enriched for AAAAAA, suggesting an A-rich motif at that location rather than artifactual A-rich ends.

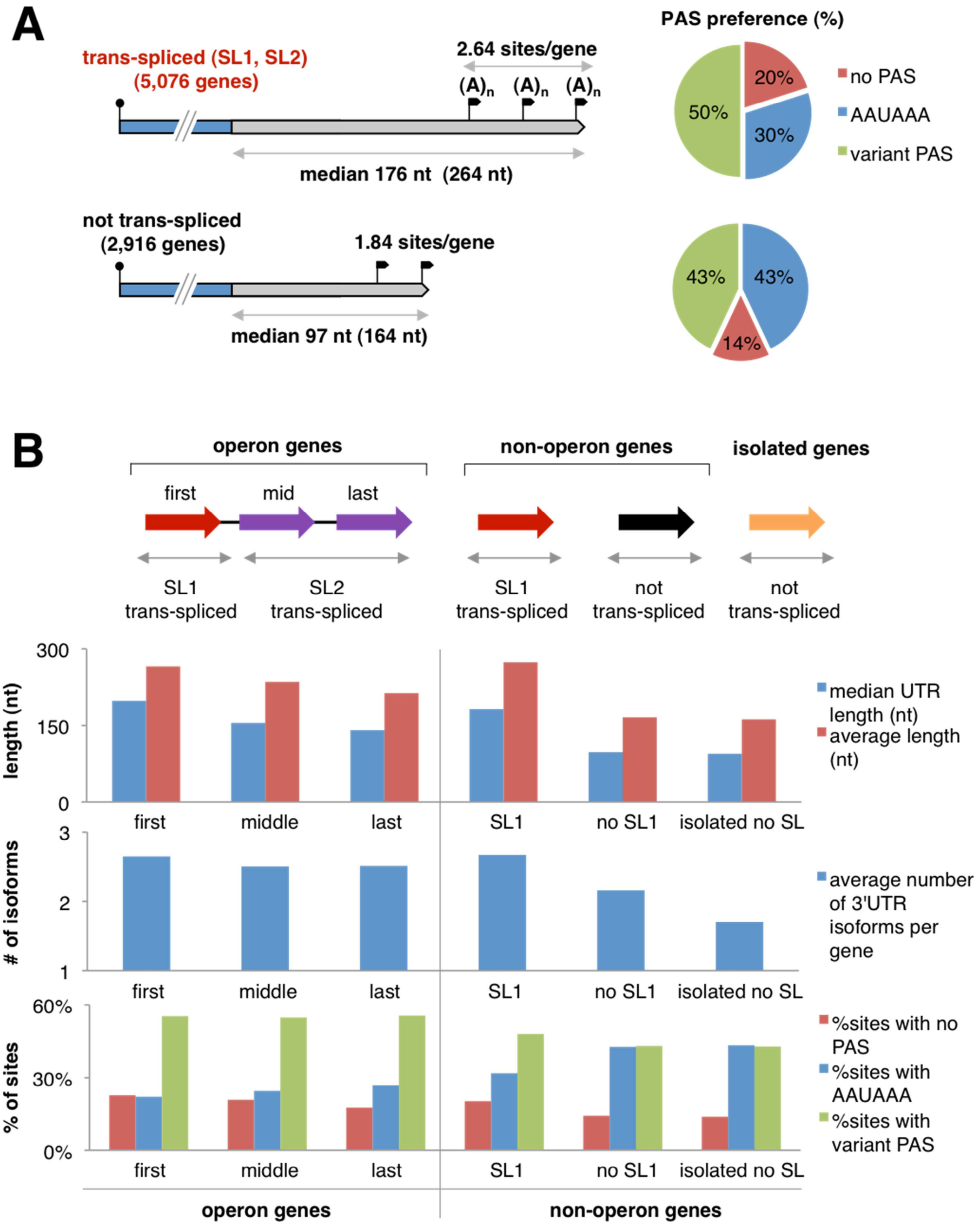


Figure 2.2: 3'UTRs in operons and trans-spliced vs. non-trans-spliced mRNAs. A) Trans-spliced (top) and non-trans-spliced (bottom) mRNAs: 3'UTR median (and average) lengths, number of 3'UTR isoforms per gene (polyA sites, black flags), and PAS preference (pie charts: % 3'UTRs with AAUAAA, variant PAS, and no PAS). B) Top panel: Schematic of operon (left, n=574 operons), non-operon (center, n=4,348 genes),

and isolated (right, n=2,098) genes. Initial operon genes (red) are SL1-trans-spliced; downstream genes (purple) usually acquire one of SL2-SL12. Non-operon genes are either SL1-trans-spliced (red, n=3,530) or not trans-spliced (black, n=818). Isolated genes (having no neighbors within 2 kb) are not trans-spliced (orange, n=2,098). Lower panels: 3'UTR lengths, number of isoforms, and PAS sites for operon and non-operon genes.

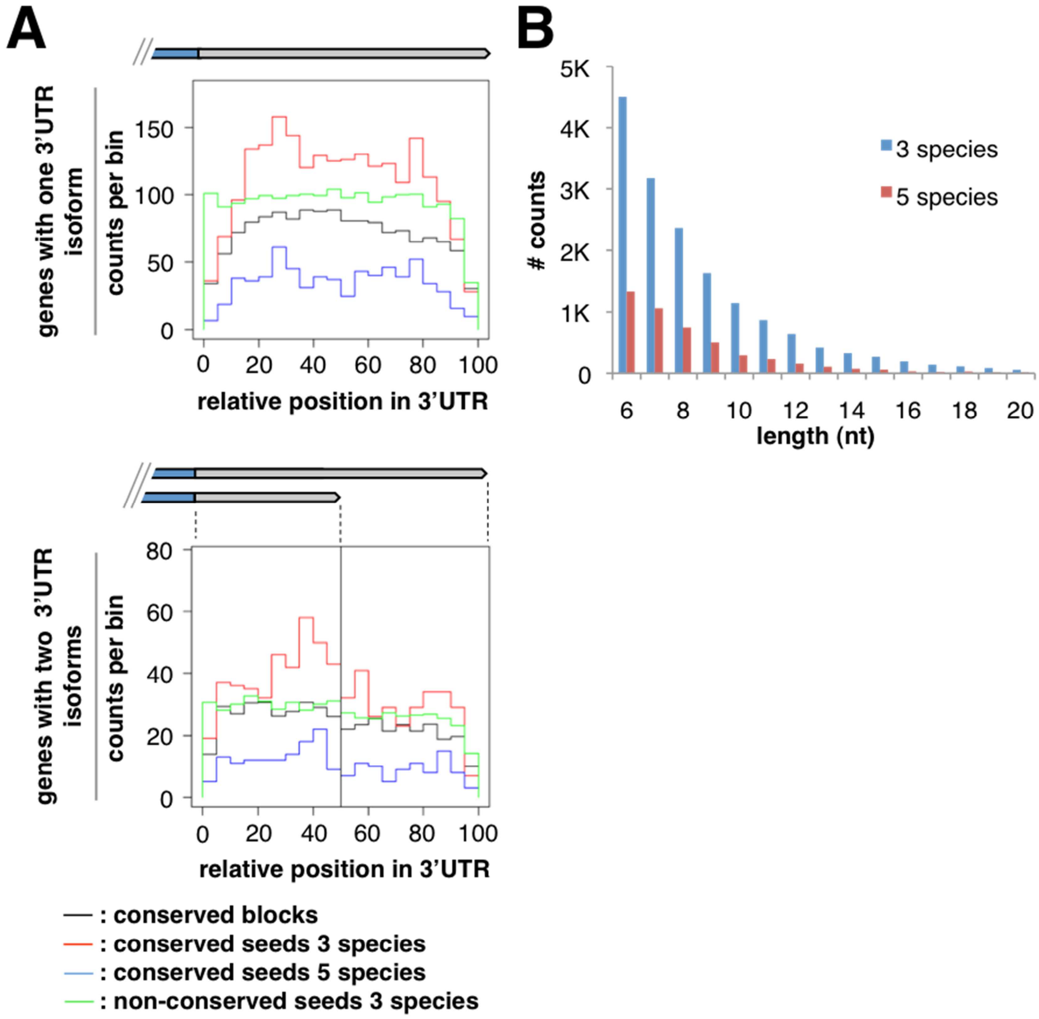


Figure 2.3: Conserved sequence elements in 3'UTRs.

A) Histogram distributions of conserved sequence blocks (black, counts shown at 1/5th scale), conserved miRNA seeds in three (red; *C. elegans, C. remanei*, *C. briggsae*) and five (blue; *C. elegans, C. remanei*, *C. briggsae*, *C. brenneri*, *C. japonica*) species, and non-conserved miRNA seeds (green, 1/25th scale) along the normalized length of 3'UTRs, in genes with one isoform (top) or exactly 2 isoforms (bottom). For genes with one isoform, length scale is 100%; for two isoforms, 0-50% represents short isoform span, 51-100% the span exclusive to long isoform. Counts were binned by fraction of total length, and thus varied in absolute length. B) Length distribution (up to 20 nt) of conserved sequence blocks in 3'UTRs (excluding miRNA target and PAS sites), in three (blue; n=16,204) and five species (red; n=4,758). See also Table S7.

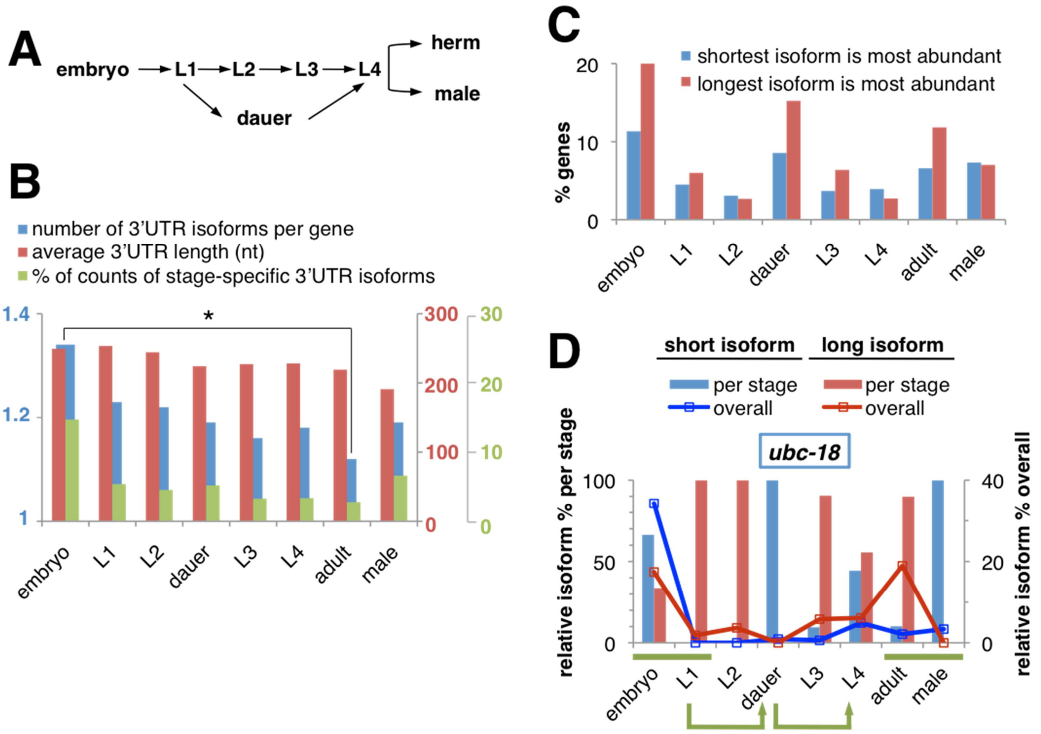
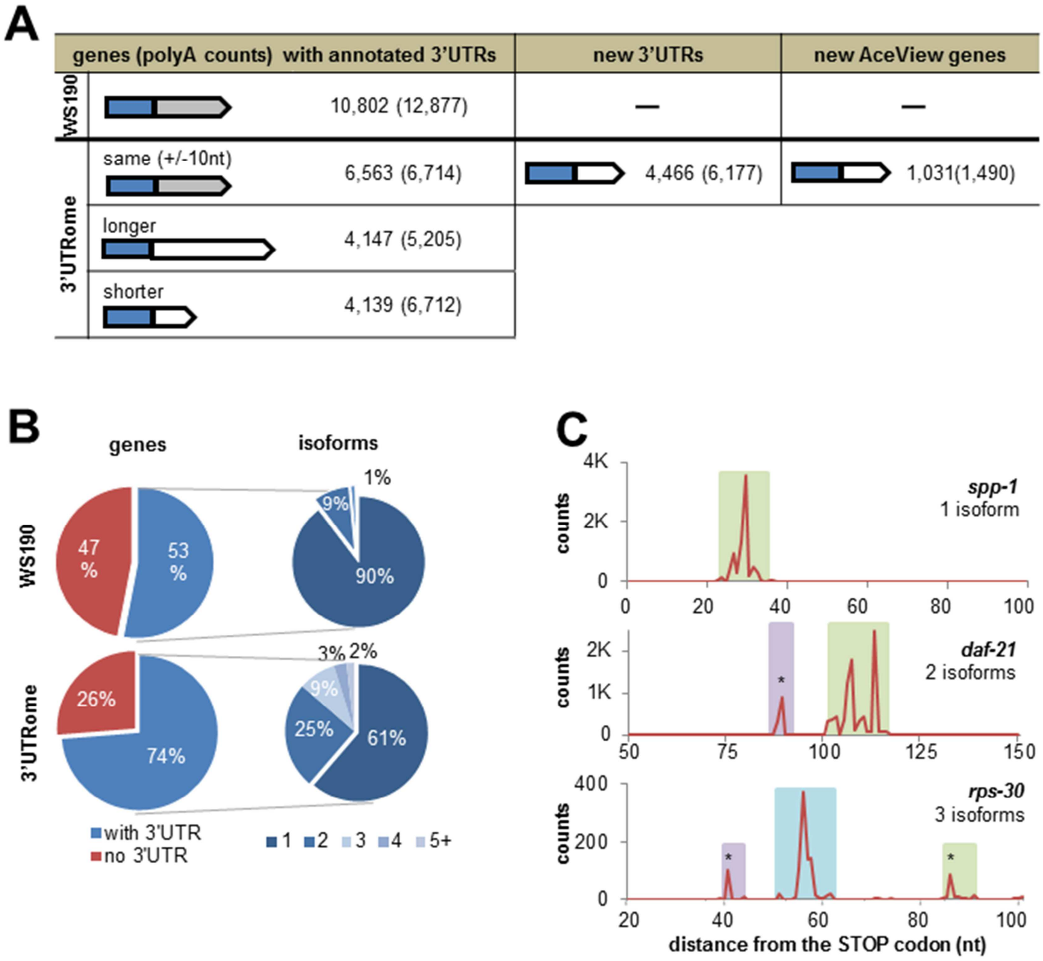


Figure 2.4: 3'UTRs during development.
A) *C. elegans* developmental transitions: embryogenesis, four larval stages, and adults. In unfavorable environments, L1 larvae arrest in ‘dauer’ stage, and can re-enter the lifecycle as L4 larvae. B) Blue: The number of 3’UTR isoforms per gene decreases significantly during development (p~0.004, permutation test). Red: The average length of 3’UTRs decreases during development. Adult males have shorter average 3’UTRs than hermaphrodites. Green: Embryos show more stage-specific 3’UTR isoforms for genes expressed during multiple developmental stages (see Table S8). C) Proportion of genes showing stage-specific expression of alternative 3’UTR isoforms (see Table S9). Embryos and dauers favor longer 3’UTR isoforms. D) Differential 3’UTR isoform expression during development (*ubc-18* shown; see Datasets S5, S6 for details). Bar chart: relative abundance of short vs. long 3’UTR isoforms for *ubc-18* in each stage (sum per stage=100%, left y-axis). Line graph: relative abundance across all stages (sum per gene across all stages=100%, right y-axis). Green bars highlight differences in 3’UTR isoform usage in embryo-to-L1 and between adult hermaphrodite and male stages. Green arrows: dauer entry and exit transitions.



		genes (polyA counts) with annotated 3'UTRs	new 3'UTRs	new AceView genes
WS190		10,802 (12,877)	—	—
3'UTRome	same (+/-10nt)	6,563 (6,714)	4,466 (6,177)	1,031(1,490)
	longer	4,147 (5,205)		
	shorter	4,139 (6,712)		

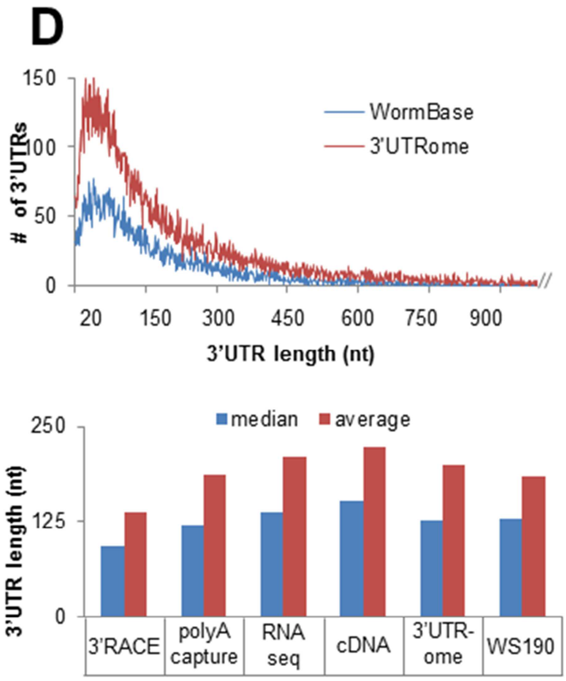


Figure 2.S1: Overview of the 3’UTRome

A,B,D) Comparison with WormBase (WS190) gene models. A) The 3'UTRome contains 3'UTRs of similar, longer, and shorter length for WS190 genes with annotated 3'UTRs (left column); 3'UTRs for WS190 genes with no annotated 3'UTRs (middle column); and 3'UTRs for transcriptional units not annotated in WS190 (AceView genes) (right column). B) WormBase WS190 contains 3'UTR annotations for 10,802 protein coding genes (53% of total); of these, only 10% are annotated with two or more 3’UTR isoforms. Our 3'UTRome covers 14,918 WS190 coding genes (74%), 39% of which possess two or more isoforms. C) Observed counts of polyA sites from independent sequence reads cluster together, defining one or more 3'UTR isoforms. Variability within polyA clusters (colored boxes) spans up to ~20 nt. Asterisks denote newly identified 3’UTR isoforms. D) Top panel: The length distribution of 3'UTRs in WS190 and 3’UTRome datasets are homothetic. Bottom panel: median (blue bar) and average (red bar) length of 3’UTRs detected in each dataset.

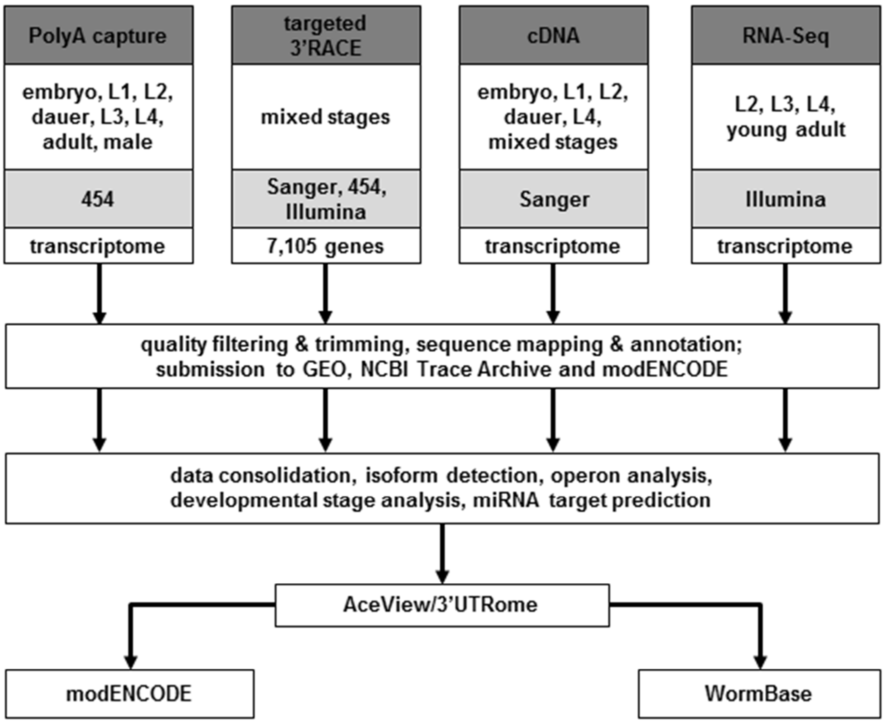

PolyA capture
embryo, L1, L2, dauer, L3, L4, adult, male
454
transcriptome
targeted 3'RACE
mixed stages
Sanger, 454, Illumina
7,105 genes
cDNA
embryo, L1, L2, dauer, L4, mixed stages
Sanger
transcriptome
RNA-Seq
L2, L3, L4, young adult
Illumina
transcriptome
quality filtering & trimming, sequence mapping & annotation; submission to GEO, NCBI Trace Archive and modENCODE
data consolidation, isoform detection, operon analysis, developmental stage analysis, miRNA target prediction
AceView/3'UTRome
modENCODE
WormBase

Figure 2.S2: Overview of 3'UTRome computational pipeline

The 3'UTRome project is composed of four datasets. PolyA capture and targeted 3'RACE were generated in this study, while publicly available cDNA and RNA-Seq data were reanalyzed and curated as part of this effort. Barcoded polyA capture tags contain the 3' end portions of 3'UTRs from staged samples; 3'RACE products directed at 7,105 coding genes were cloned from mixed-stage samples. The cDNA dataset represents AceView-curated cDNA and EST sequences using, where possible, the original traces from cDNA libraries produced by the Kohara laboratory, Exelixis, and others obtained from the NCBI trace repository, as well as cDNA sequences from NCBI sequence repositories (GenBank, dbEST). The RNA-Seq dataset consists of published data for staged mRNA samples from the modENCODE *C. elegans* transcriptome project *(8)* and previously reported L1-stage data *(9).* Datasets were sequenced as indicated (gray shading). Sequences were processed (to remove vector, linker, barcode, and polyA sequences), filtered for read quality, and aligned to the *C. elegans* WS190/CE6 genome. The consolidated datasets were used to define a compendium of 3'UTR isoforms, which was used for downstream analyses of 3'UTR structure and function. Raw data and annotations for the compendium are available in public repositories, including NCBI GEO and Trace Archive, the 3'UTR-centric 3'UTRome database (http://www.utrome.org), AceView (http://www.aceview.org), modENCODE (http://www.modencode.org), and WormBase (http://ww.wormbase.org). Supplementary Materials and Methods provide additional details on data production and analysis.

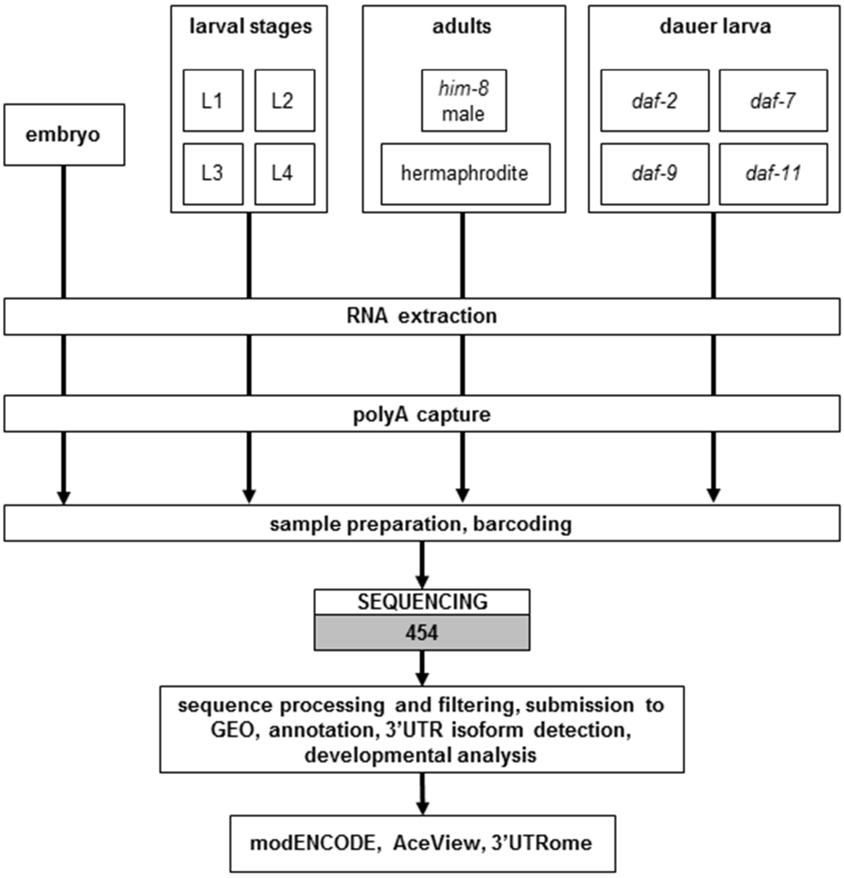

larval stages
adults
dauer larva
embryo
L1
L2
L3
L4
him-8 male
hermaphrodite
daf-2
daf-7
daf-9
daf-11
RNA extraction
polyA capture
sample preparation, barcoding
SEQUENCING
454
sequence processing and filtering, submission to GEO, annotation, 3'UTR isoform detection, developmental analysis
modENCODE, AceView, 3'UTRome

Figure 2.S3: Workflow for polyA capture assay

Barcoded polyA capture libraries were prepared using total RNA from staged animals and sequenced by Roche/454. Reads were filtered for quality, processed to remove adaptor and barcode sequences, and aligned to the WS190/CE6 genome build. Raw and processed sequence files were submitted to GEO. Alignments were consolidated with the other 3'UTR datasets and annotated with respect to WS190 and AceView gene models. Data and annotations are available in AceView, 3'UTRome, and modENCODE databases (see Supplementary Materials and Methods for details).

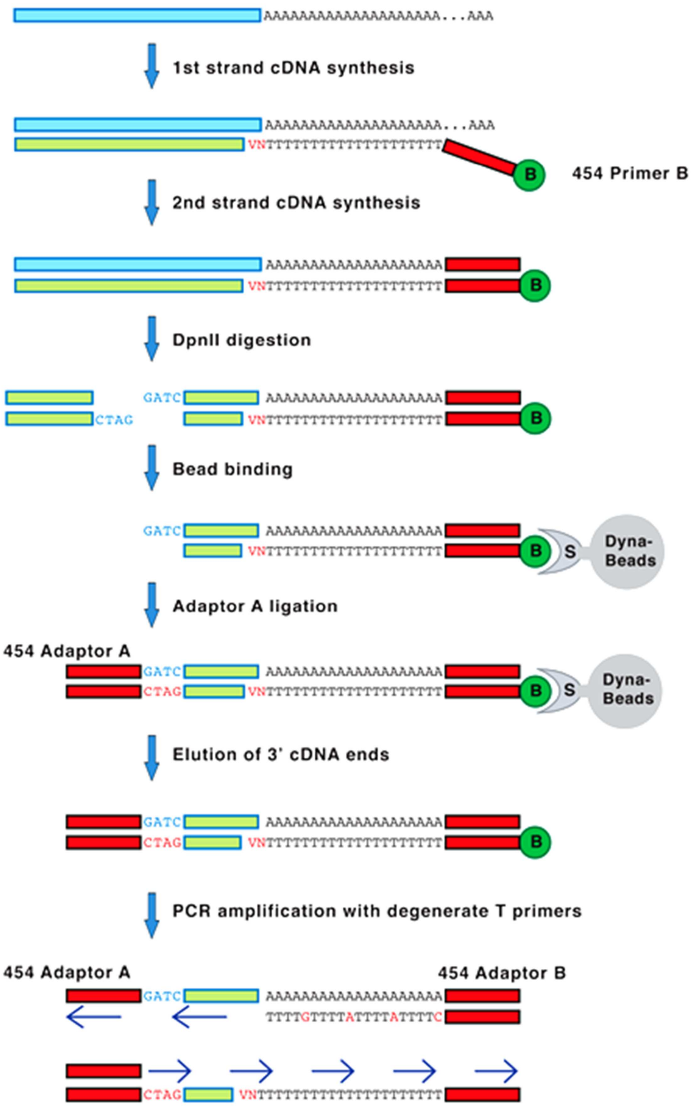


Figure 2.S4: PolyA capture protocol

Total RNA from staged samples (Figure S3) served as template for a first-strand reverse transcriptase (RT) reaction with an anchored, biotinylated poly-dT primer. Second-strand synthesis with T4 DNA polymerase produced dsDNA products that were digested with *DpnII*. Three-prime terminal fragments were recovered using streptavidin beads, ligated with barcoded 454 sequencing primers, PCR amplified, and subjected to pyrosequencing (see Supplementary Materials and Methods for details).

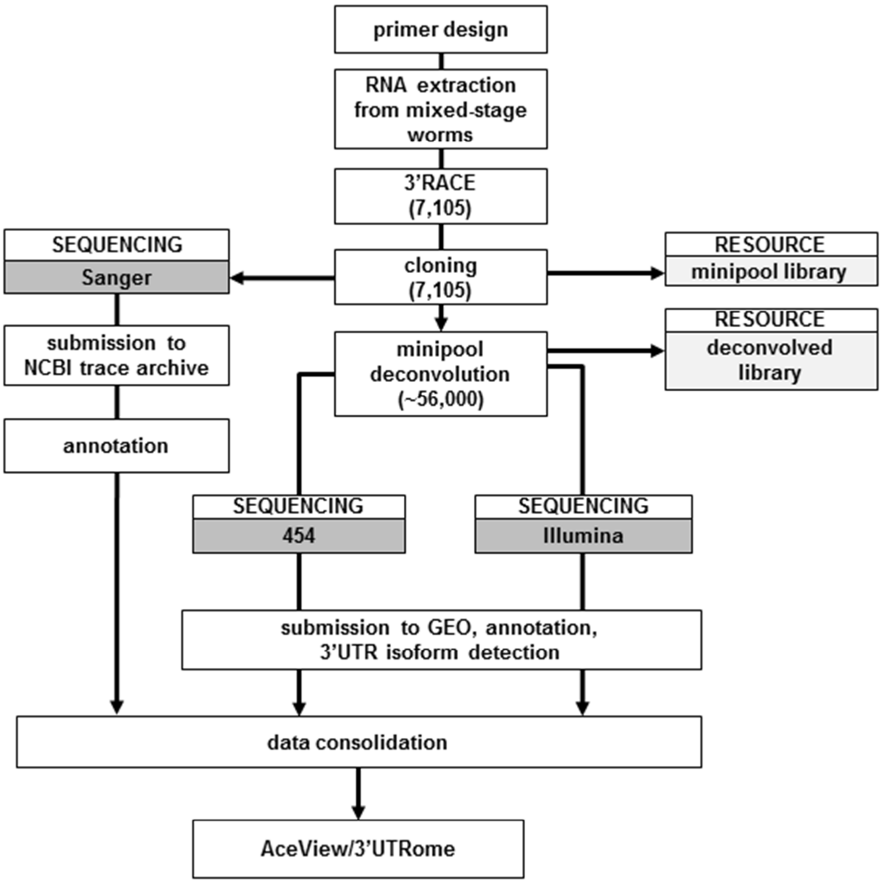


Figure 2.S5: Workflow for 3'RACE

A 3'RACE cloning pipeline was designed to target 3'UTRs of 7,105 CDSs for 6,741 genes previously included in the Promoterome (*3*) and ORFeome (*4-5*) collections. 3'RACE products were generated from total RNA isolated from mixed developmental stages, cloned into Gateway™ vectors, and collected as minipools of products for each target. Minipools were sequenced using the Sanger method. Eight individual colonies per minipool were isolated and re-pooled into eight bar-coded libraries containing one individual clone per targeted gene. Barcoded libraries were sequenced using Illumina and Roche/454 platforms. Minipool and deconvolved single-clone sequences were trimmed for vector and barcode sequences, filtered for quality, and aligned to the WS190/CE6 genome sequence. Alignments that extended beyond the CDS-specific primer were annotated and consolidated with other 3'UTRome datasets in AceView and 3'UTRome databases (see Supplementary Materials and Methods for details).

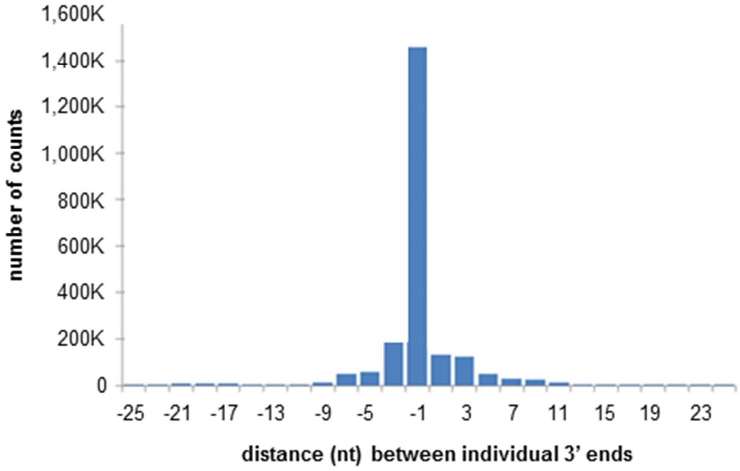


Figure 2.S6: Distance between individual 3'ends and the representative polyA addition site for a cluster

Frequency distribution of distance (in nucleotides) between the representative polyA site in a cluster and all other polyA sequence tags in the same cluster. Data are cumulative for all polyA clusters in the 3'UTRome. 86% of individual polyA tags fall within 4nt of the representative polyA site.

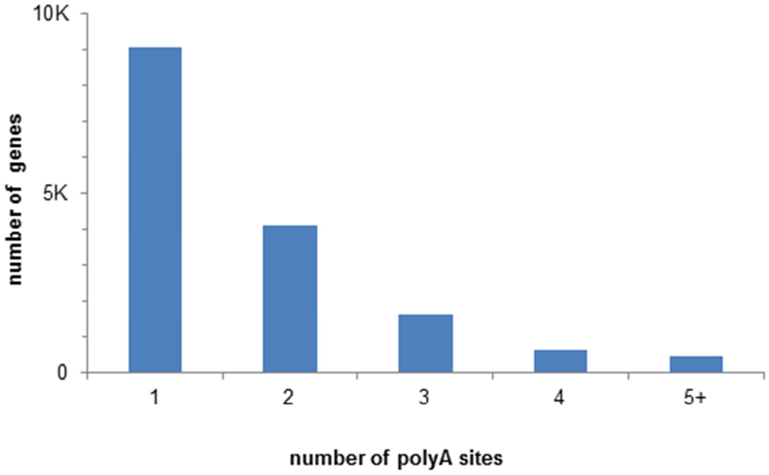


Figure 2.S7: Distribution of the number of polyA sites per gene

The frequency distribution of distinct representative polyA sites per gene in the 3'UTRome. Around 40% of all genes with an annotated 3'UTR contain more than one alternative polyA site. Among genes with a large number of alternative 3'UTR isoforms are those encoding the small GTPase RAB-11.1 (6 isoforms), the LIN-61 paralog MBTR-1 (7 isoforms), and the RNA helicase VBH-1 (8 isoforms).

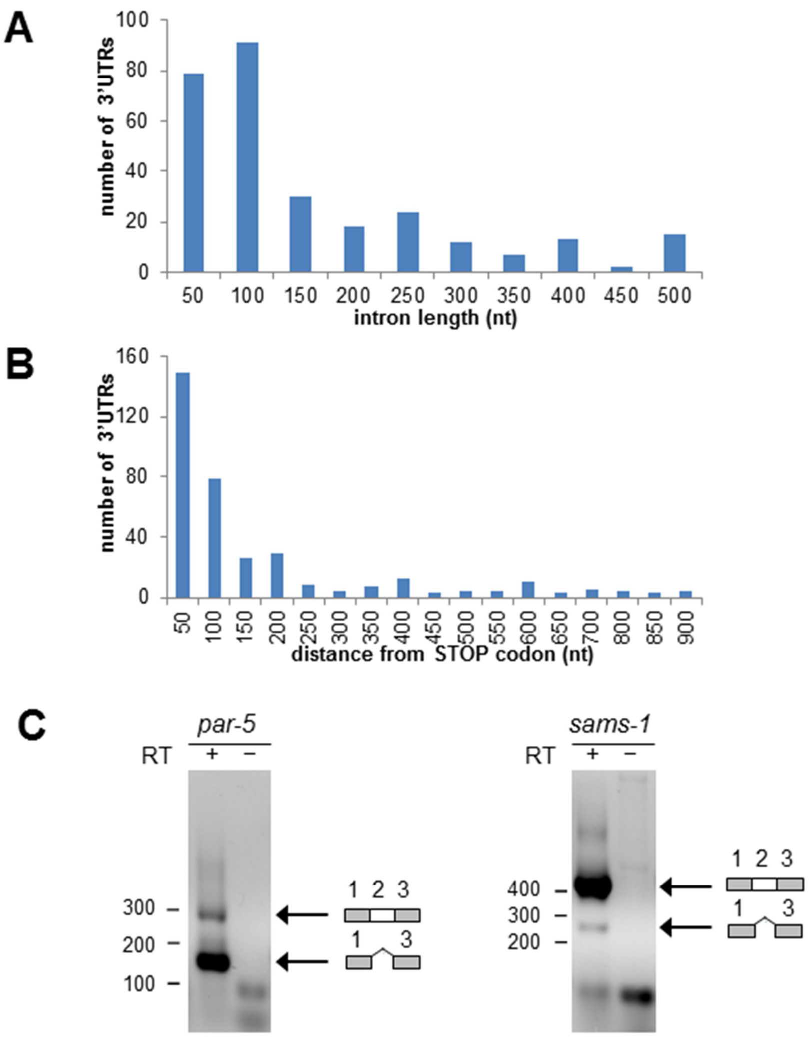


Figure 2.S8: Introns in 3'UTR regions

363 intron-containing 3'UTRs for 192 unique genes were used in this analysis. A) Length distribution (in nucleotides) of introns in 3'UTRs. B) Length distribution of the distance from the STOP codon to the intron start position. In both A and B, intron length is shown in 50 nt bins for simplification. C) Examples of facultative introns. Shown are 3'RACE products from *par-5* and *sams-1* 3'UTRs using mixed-stage total RNA and gene-specific primer pairs flanking the intron (regions 1 and 3), with (+) or without (-) inclusion of reverse transcriptase (RT) in the reaction. Agarose gel electrophoresis lanes with RT each produce two products consistent in size with the retention (top band) or excision (lower band) of region 2. Small bands below 100 nt represent unamplified primers and primer dimers (see Supplementary Online Materials and Methods for details). We observe that in some of these 3'UTRs, putative binding sites for miRNAs or ALG-1 (*20*) are contained within an intronic sequence.

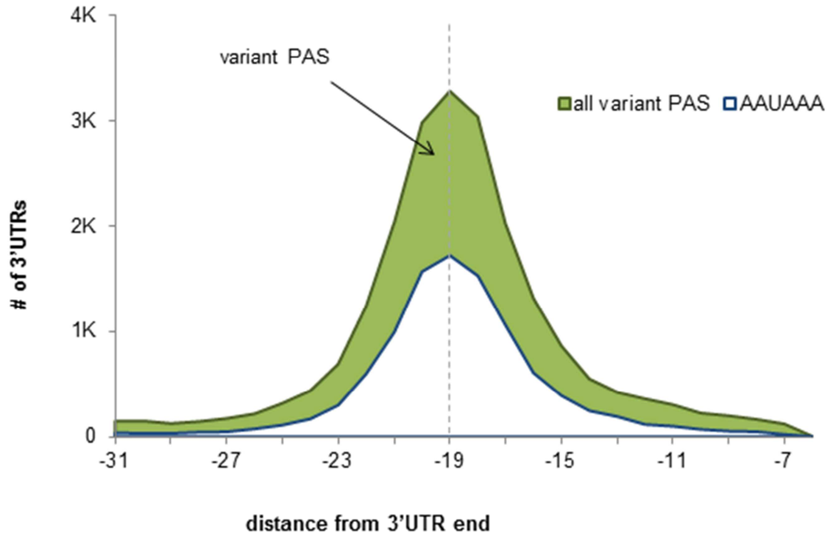


Figure 2.S9: Distribution of the canonical AAUAAA and variant PAS elements relative to the cleavage and polyA addition site

Start position for all PAS motifs (green line), AAUAAA (blue line), and variant PAS (green shading) peak at 19 nt upstream of the polyA addition site. See Supplementary Materials and Methods for details on the identification of PAS motifs and assignment of the most likely PAS motif for each 3'UTR.

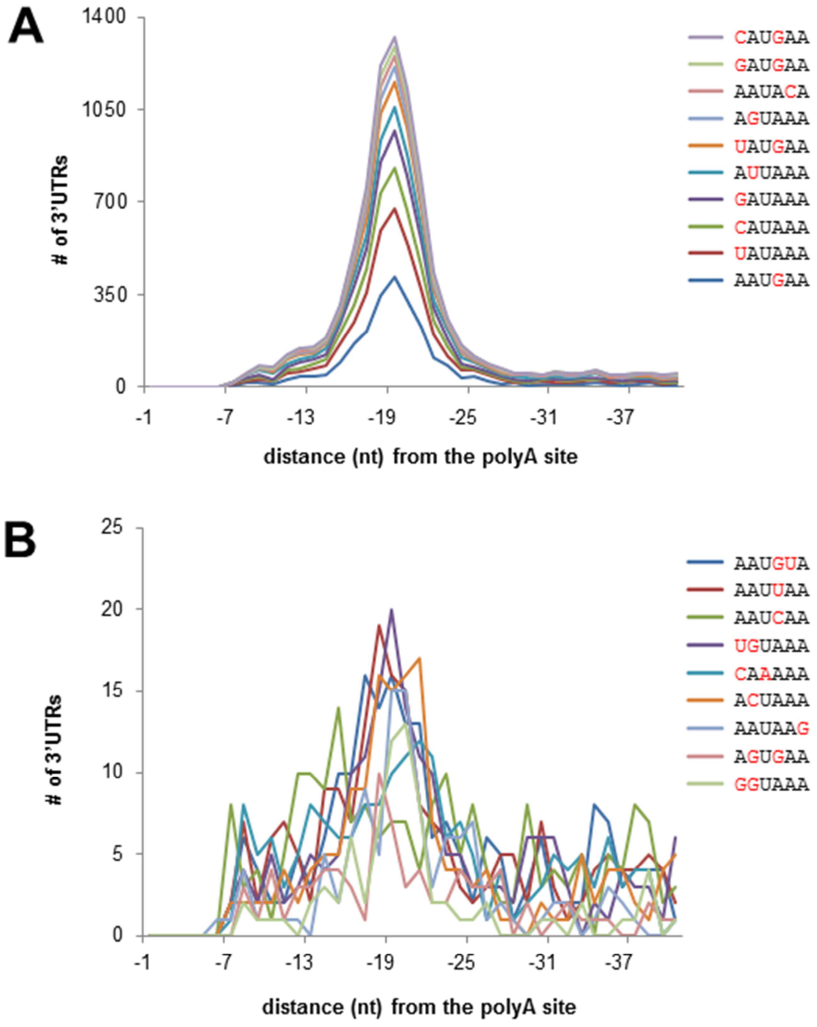


Figure 2.S10: Distribution of variant PAS elements relative to the cleavage and polyA addition site

In an unbiased search of all possible hexamers in the regions upstream of polyA sites in the 3'UTRome, the most common variant PAS hexamers show an enrichment that peaks at 19-20 nucleotides upstream of the polyA site. Using this as a guide, the most likely PAS motif for each polyA site was assigned using an ordered list of motifs according the the frequency of each motif in this region (see Supplementary Materials and Methods for details). The distribution of the most common motif, the canonical AAUAAA, which peaks at position -19, is not shown in this figure.

A) Ten of the most common variant PAS motifs (each assigned to ≥1% of all polyA sites). The most common PAS variants contain a U in the third position and an A in the sixth position. B) Nine of the least common variant PAS motifs (each assigned to ≤1% of all polyA sites). Total counts for each motif are given in Table S5.

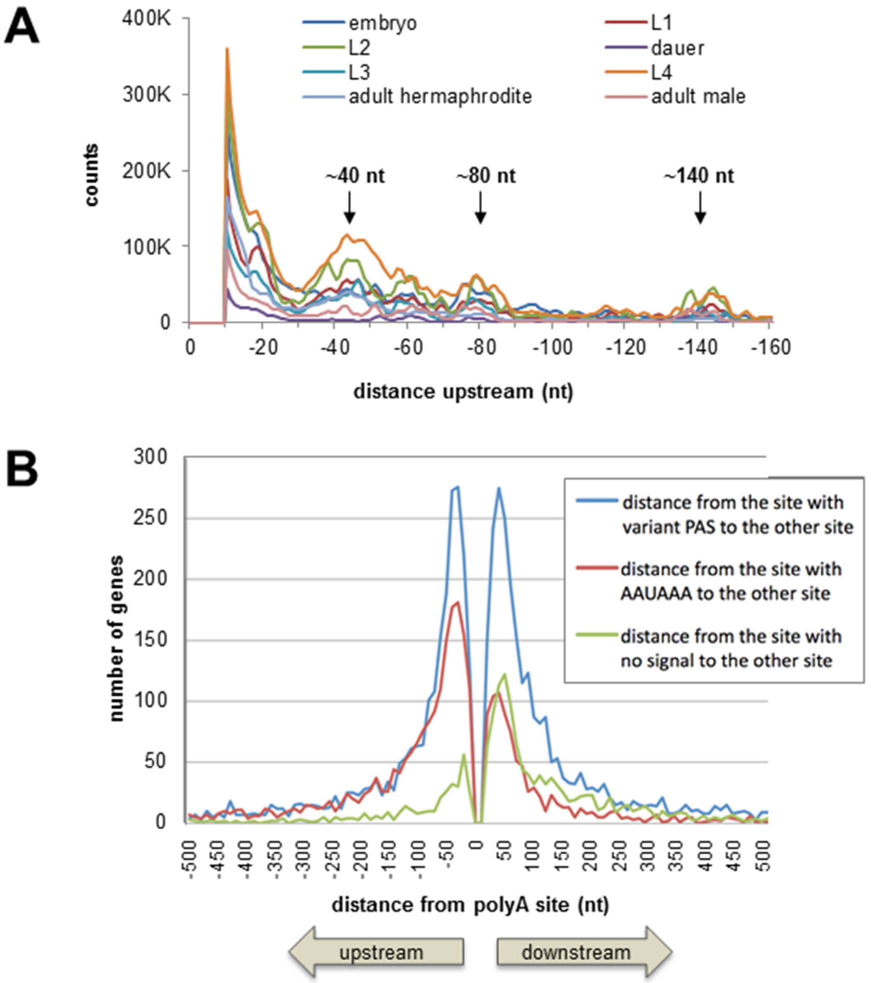


Figure 2.S11: Relationships between alternative polyA addition sites for the same transcript

A) The autocorrelation of polyA addition sites, pooled by stage, showing the average support count at each position relative to the most highly supported polyA site (aligned at 0 nt). The data show a main peak (arrow) ~40-45 bases upstream of the dominant polyA site. B) The distance between adjacent polyA sites peaks at ±45nt. PolyA addition sites with the canonical AAUAAA PAS motif (red) show a propensity to have a neighboring polyA site upstream; conversely, sites with no detectable PAS (green) tend to have a neighboring site downstream. Sites with a variant PAS (blue) are equally likely to have a neighboring site upstream or downstream.

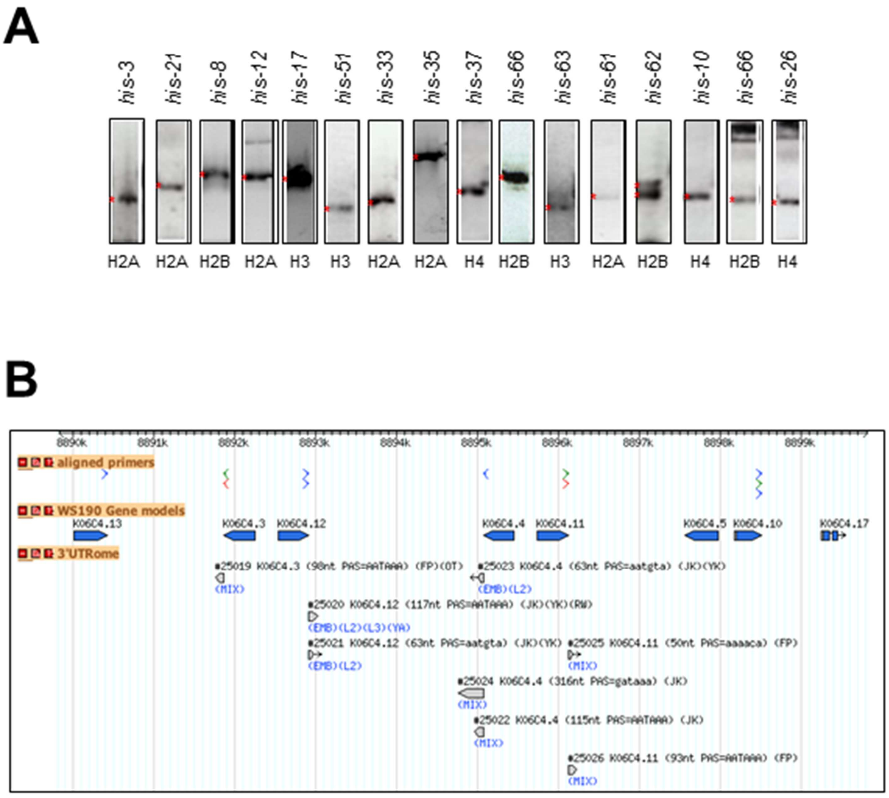


Figure 2.S12: Polyadenylated 3'UTRs for histone genes

A) The electrophoretic analysis on 1% agarose E-Gels of selected 3’RACE clones corresponding to 3’UTRs of histone genes obtained with the 3'RACE pipeline. PCR amplicons (red asterisks) correspond to unique or multiple 3’UTR isoforms. B) Histone gene cluster on chromosome V. Several histone genes with corresponding 3’UTRs detected in multiple developmental stages are shown. See Table S6 for the comprehensive list of histone 3’UTRs and PAS usage.

Combined with the observation that depletion of the SLBP homolog CDL-1 by RNAi severely depletes histone protein but not mRNA levels (*21*), our data lend support to the hypothesis that replication-dependent histone transcripts in *C. elegans* are first cleaved and polyadenylated using a PAS-directed mechanism, and are later post-processed to their final stem-loop form and regulated at the translational level by factors including CDL-1.

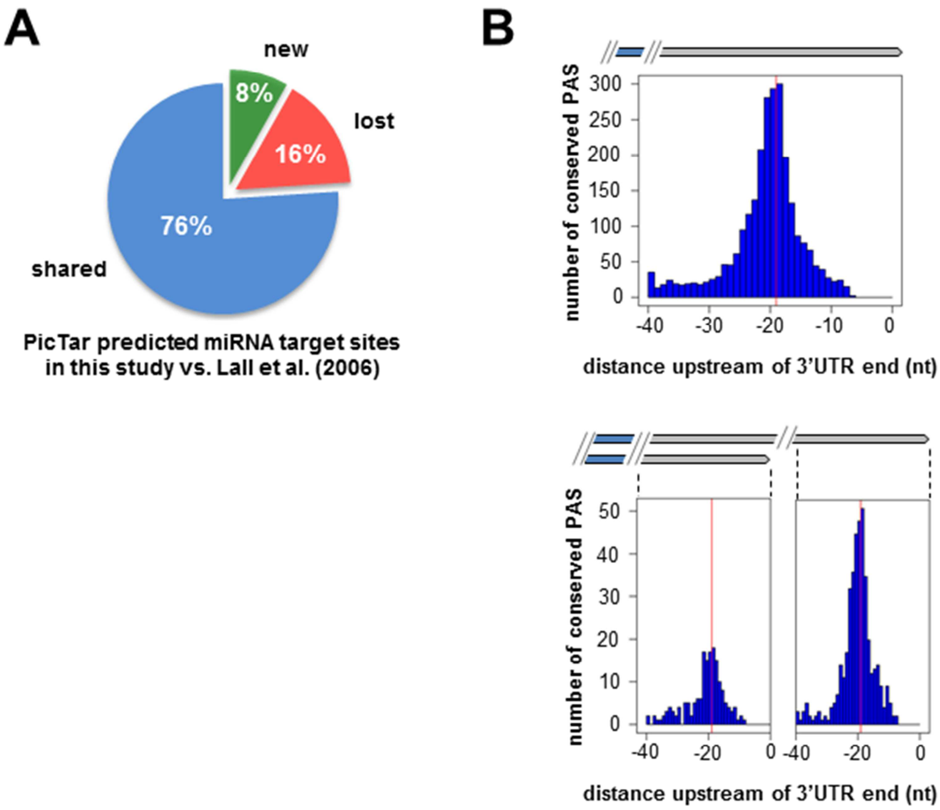


Figure 2.S13: PicTar miRNA target predictions and PAS conservation

A) Differences in PicTar predicted miRNA target sites within sequences spanned by the 3'UTRome, from this study in comparison with our previous predictions for *C. elegans* (*19*), as a percentage of the total number of predictions from both studies. See also Table S7. B) Distribution of conserved PAS motifs within 40 nt upstream of 3'UTR ends in three-way alignments between *C. elegans*, *C. briggsae*, and *C. remanei*, for (top) genes with one isoform (n=2,573 3'UTRs) or (bottom) exactly two isoforms (short, n=173; long, n=419). Red lines indicate the peak at -19 nt from the 3'UTR polyA addition site. See Supplementary Materials and Methods for additional details.

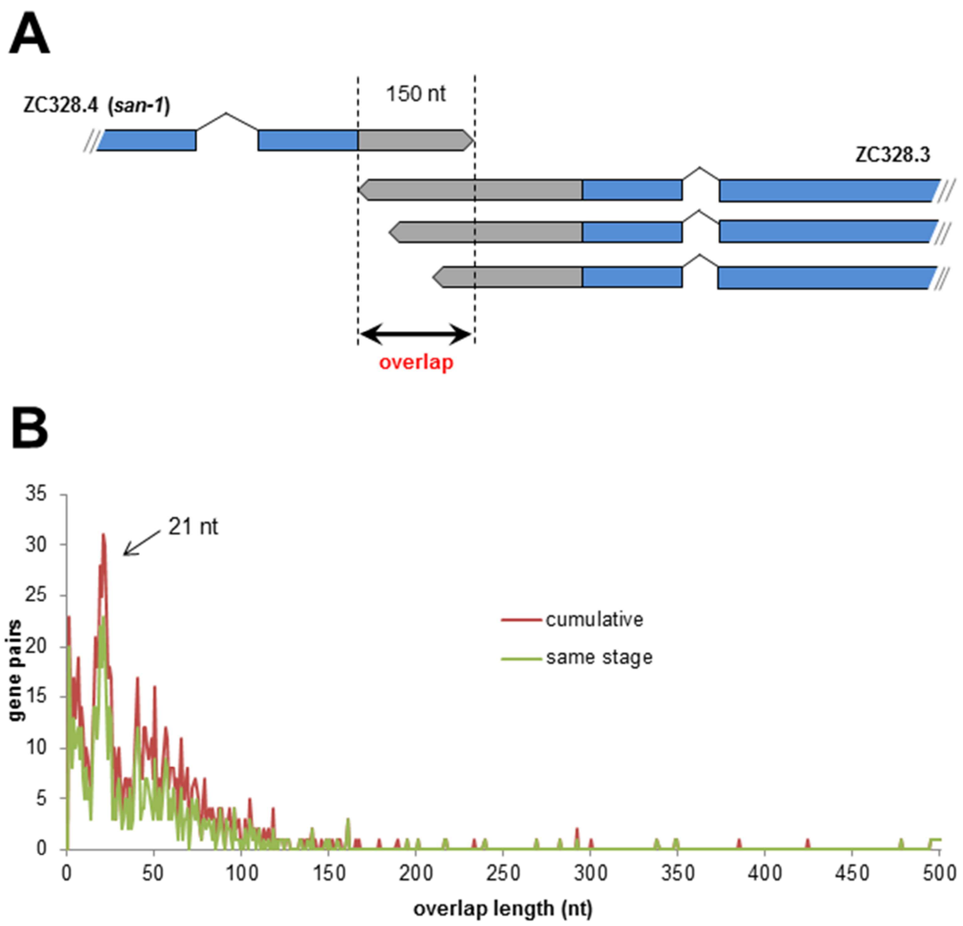
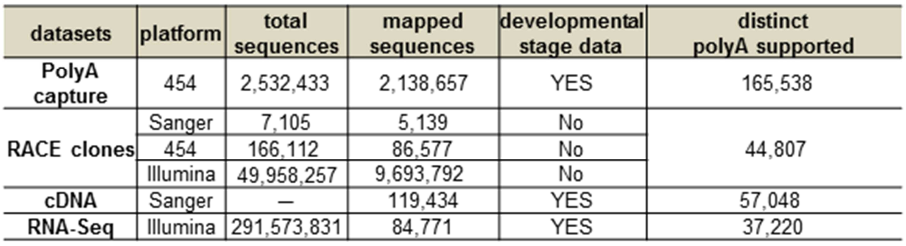


Figure 2.S14: 3'UTRs on opposite strands sometimes overlap

The 3'UTRome contains 1,876 convergently transcribed neighboring genes with overlapping regions that extend from the distal end of each putative transcript into the 3'UTR or CDS of the neighboring gene (see also Supplementary Dataset S4). For 1,240 of these genes, overlapping 3'UTR isoforms are co-expressed during at least one developmental stage. If both genes are transcribed simultaneously in the same cell, their 3'UTRs could potentially pair as dsRNA and trigger the production of endogenous siRNAs (endo-siRNAs) *(22),* which could down-regulate their mRNA levels.

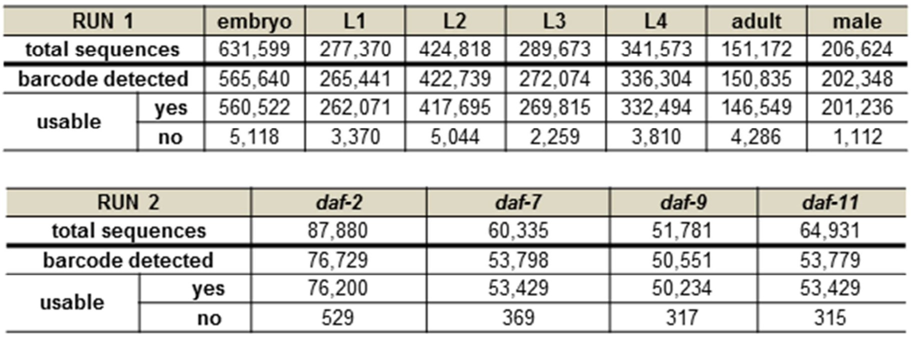
A) Example of a 3'UTR overlap between the gene encoding mitotic spindle checkpoint protein ZC328.4 (*san-1*) and the uncharacterized gene ZC328.3. B) Length distribution (nt) of overlapping 3' end annotations for gene pairs on opposite strands, for cumulative overlapping pairs (red, n=938 pairs) or pairs detected in the same developmental stage (green, n=620 pairs). Overlapping pairs involve ~10% of genes in the 3'UTRome. Overlaps range from 1 to 495 nt, with an average overlap length of ~44 nt and median overlap length of ~28 nt. The peak in the overlap distribution at ~21 nt suggests that longer overlaps generally may be disfavored to limit recruitment of cellular machinery that could lead to endo-siRNA production *(22).*



datasets	platform	total sequences	mapped sequences	developmental stage data	distinct polyA supported
PolyA capture	454	2,532,433	2,138,657	YES	165,538
RACE clones	Sanger	7,105	5,139	No	44,807
	454	166,112	86,577	No	
	Illumina	49,958,257	9,693,792	No	
cDNA	Sanger	—	119,434	YES	57,048
RNA-Seq	Illumina	291,573,831	84,771	YES	37,220

Table 2.S1: Sequence data in the 3'UTRome

Total number of raw and mapped sequences and the number of distinct polyA clusters supported for each data stream. Three of the datasets, polyA capture, cDNA and RNA-seq, provide developmental stage information allowing us to link distinct 3'UTR isoforms to specific developmental stages. See Figures S2-S5 for details on the different pipelines.

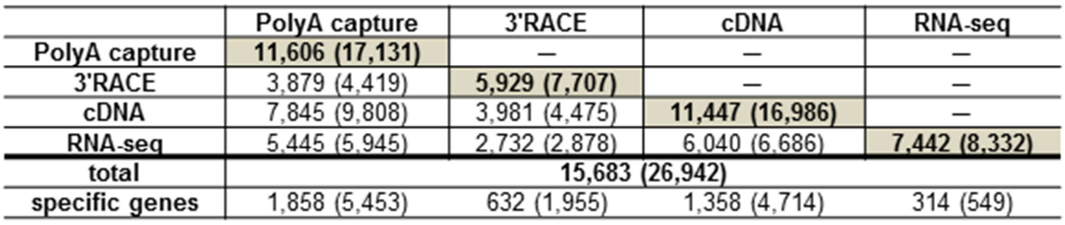


RUN 1		embryo	L1	L2	L3	L4	adult	male
total sequences		631,599	277,370	424,818	289,673	341,573	151,172	206,624
barcode detected		565,640	265,441	422,739	272,074	336,304	150,835	202,348
usable	yes	560,522	262,071	417,695	269,815	332,494	146,549	201,236
	no	5,118	3,370	5,044	2,259	3,810	4,286	1,112

RUN 2		*daf-2*	*daf-7*	*daf-9*	*daf-11*
total sequences		87,880	60,335	51,781	64,931
barcode detected		76,729	53,798	50,551	53,779
usable	yes	76,200	53,429	50,234	53,429
	no	529	369	317	315

Table 2.S2: Summary of the polyA capture 454 sequencing runs

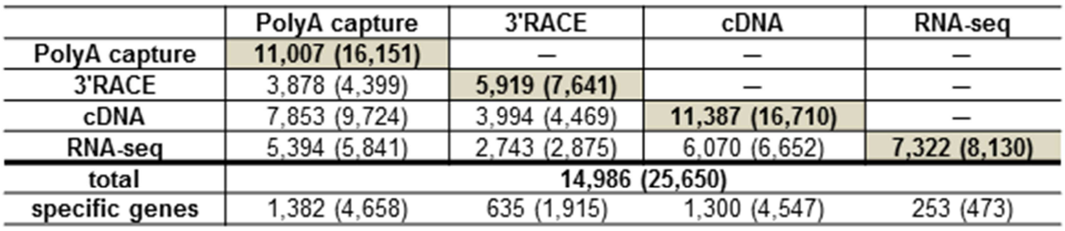
Roche/454 reads produced by the polyA capture in individual developmental stages, males, and dauer mutants. The sequences obtained (total sequences) were scanned for the detection of a barcode (barcode detected). Reads containing a sequence contiguous with a polyA site were classified as 'usable'.



	PolyA capture	3'RACE	cDNA	RNA-seq
PolyA capture	**11,606 (17,131)**	–	–	–
3'RACE	3,879 (4,419)	**5,929 (7,707)**	–	–
cDNA	7,845 (9,808)	3,981 (4,475)	**11,447 (16,986)**	–
RNA-seq	5,445 (5,945)	2,732 (2,878)	6,040 (6,686)	**7,442 (8,332)**
total	**15,683 (26,942)**			
specific genes	1,858 (5,453)	632 (1,955)	1,358 (4,714)	314 (549)

Table 2.S3: Gene and 3’UTR isoform coverage for individual datasets and overlaps among datasets in the 3'UTRome using AceView gene models

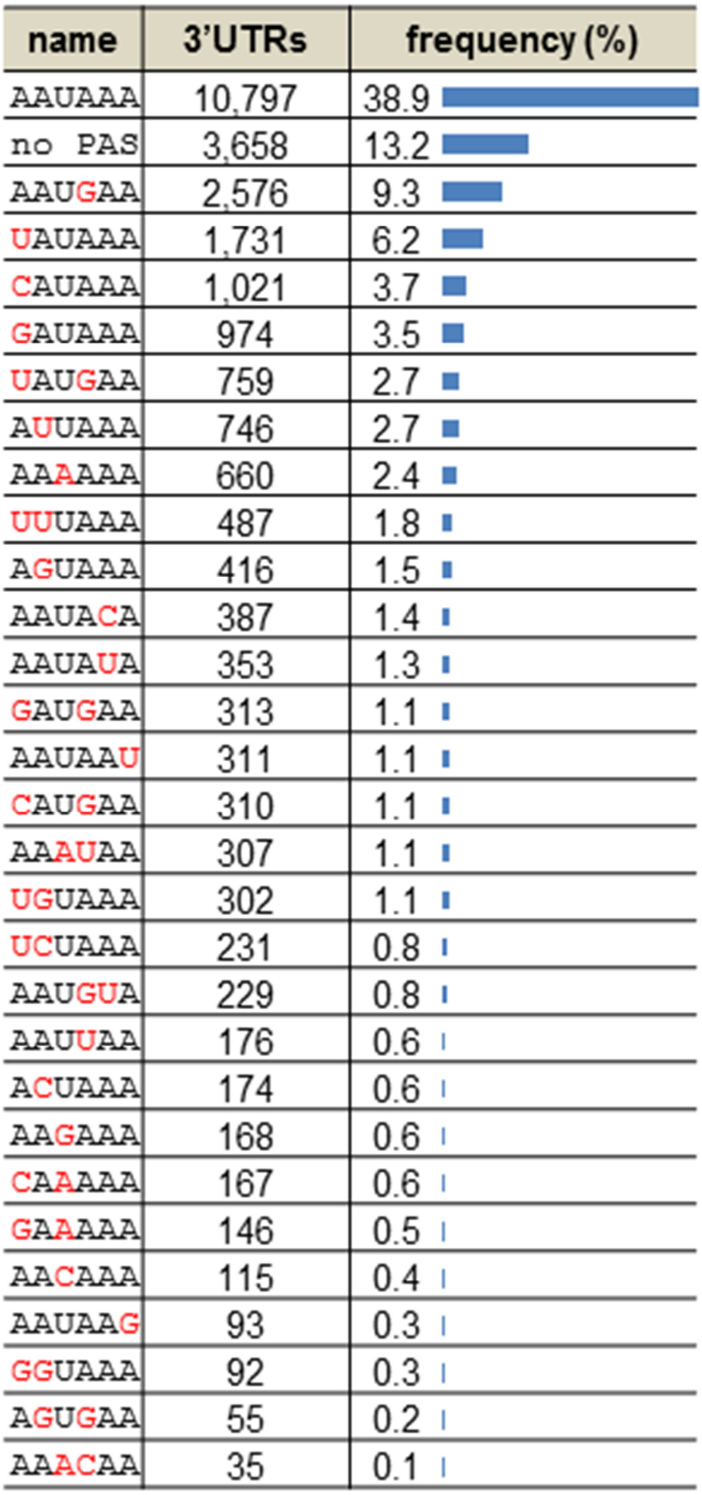
Diagonal cells show the total number of coding genes and distinct polyA ends (in parentheses) for each of the four independent datasets; off-diagonal cells show intersections between each pair of datasets. The last row shows the total number of coding genes and distinct polyA ends that are specific to each individual dataset.



	PolyA capture	3'RACE	cDNA	RNA-seq
PolyA capture	**11,007 (16,151)**	–	–	–
3'RACE	3,878 (4,399)	**5,919 (7,641)**	–	–
cDNA	7,853 (9,724)	3,994 (4,469)	**11,387 (16,710)**	–
RNA-seq	5,394 (5,841)	2,743 (2,875)	6,070 (6,652)	**7,322 (8,130)**
total	**14,986 (25,650)**			
specific genes	1,382 (4,658)	635 (1,915)	1,300 (4,547)	253 (473)

Table 2.S4: Subset of 3’UTRome matching WS190 gene models

The subset of data from Table S3 that are compatible with WormBase WS190 gene models. See Table S3 legend for additional details.

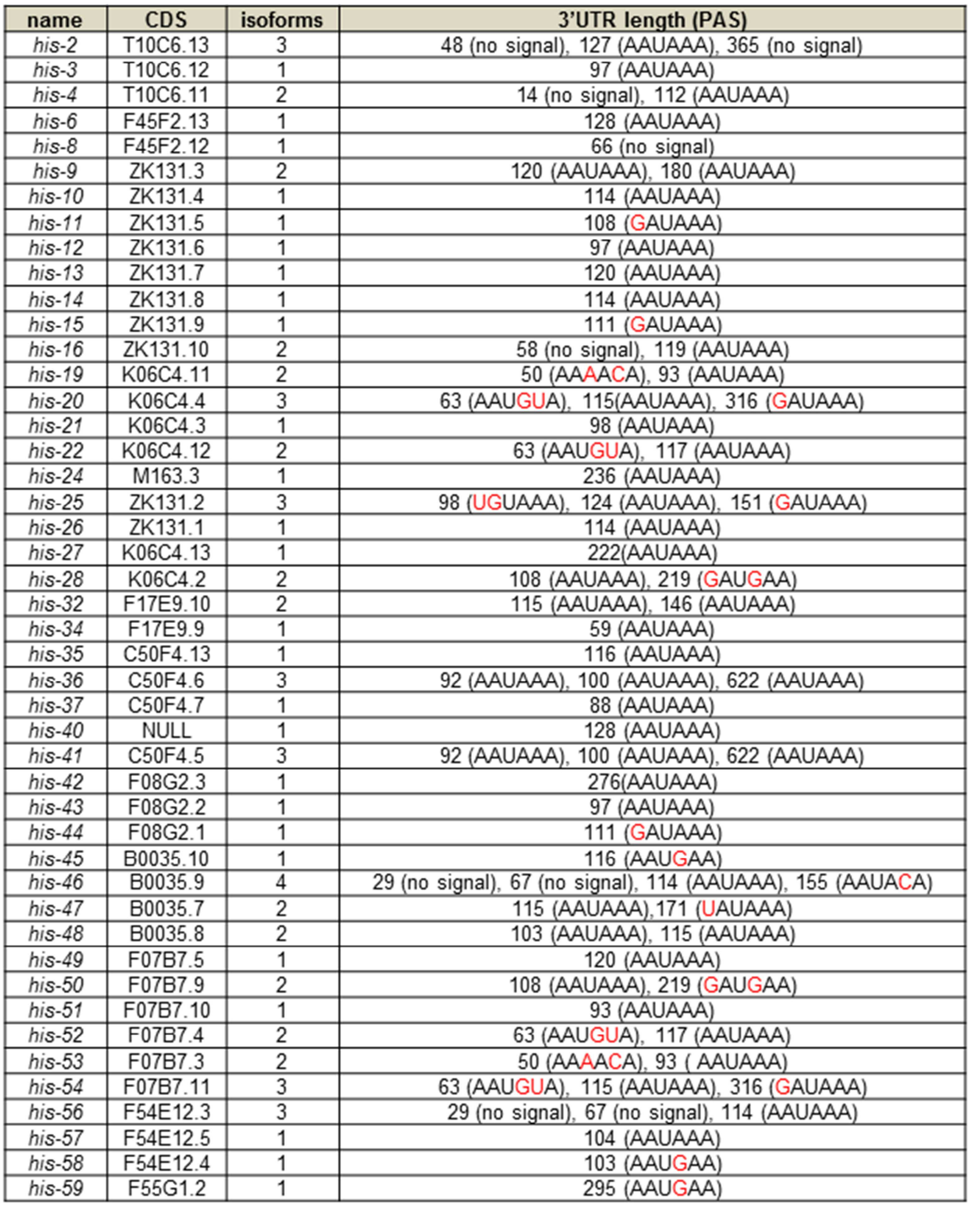


name	3'UTRs	frequency (%)
AAUAAA	10,797	38.9
no PAS	3,658	13.2
AAUGAA	2,576	9.3
UAUAAA	1,731	6.2
CAUAAA	1,021	3.7
GAUAAA	974	3.5
UAUGAA	759	2.7
AUUAAA	746	2.7
AAAAAA	660	2.4
UUUAAA	487	1.8
AGUAAA	416	1.5
AAUACA	387	1.4
AAUAUA	353	1.3
GAUGAA	313	1.1
AAUAAU	311	1.1
CAUGAA	310	1.1
AAAUAA	307	1.1
UGUAAA	302	1.1
UCUAAA	231	0.8
AAUGUA	229	0.8
AAUUAA	176	0.6
ACUAAA	174	0.6
AAGAAA	168	0.6
CAAAAA	167	0.6
GAAAAA	146	0.5
AACAAA	115	0.4
AAUAAG	93	0.3
GGUAAA	92	0.3
AGUGAA	55	0.2
AAACAA	35	0.1

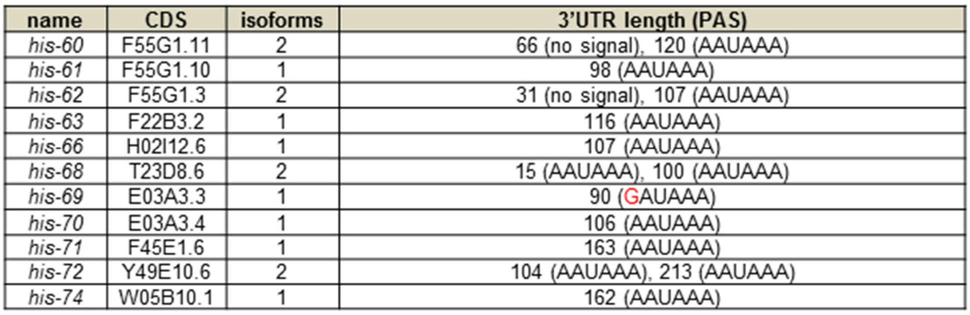
Table 2.S5: Identification of putative PAS elements

An unbiased search for over-represented hexamers in the last 50 nt of 3'UTRs in the 3'UTRome identified a handful of sequences whose start positions all peaked at around

19 nt upstream of the polyA cleavage site. Using these results as a guide, we searched all 3’UTRs recursively for the most likely PAS site utilized by each 3’UTR (see Supplementary Materials and Methods for details). The most common motif, the “canonical” PAS element AAUAAA, is observed in 39% of 3’UTRs; the other elements consist of variations of this motif differing by one or two nucleotides. This apparent diversity of PAS motifs suggests that the recognition of PAS sites in worms is more flexible than higher eukaryotes, where mutation in any position of the canonical AAUAAA element disrupts the 3’ end processing of mRNAs (*23*), and may perhaps be more akin to the 3’ end processing mechanism of yeast, where presence of an AU rich region is sufficient to allow docking of the processing machinery (*24*).



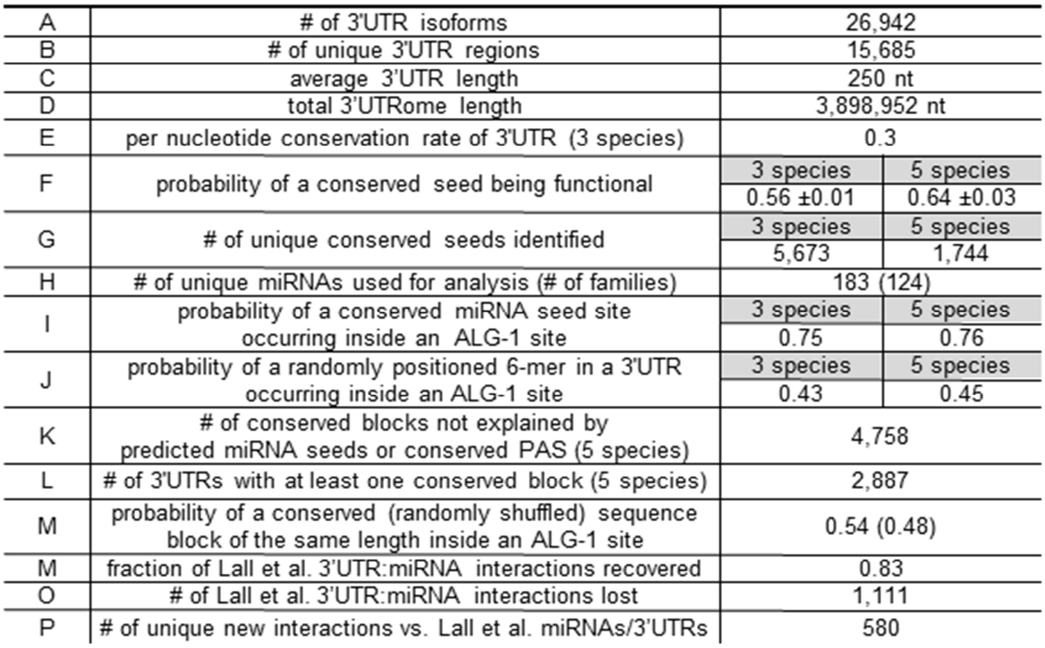
name	CDS	isoforms	3'UTR length (PAS)
his-2	T10C6.13	3	48 (no signal), 127 (AAUAAA), 365 (no signal)
his-3	T10C6.12	1	97 (AAUAAA)
his-4	T10C6.11	2	14 (no signal), 112 (AAUAAA)
his-6	F45F2.13	1	128 (AAUAAA)
his-8	F45F2.12	1	66 (no signal)
his-9	ZK131.3	2	120 (AAUAAA), 180 (AAUAAA)
his-10	ZK131.4	1	114 (AAUAAA)
his-11	ZK131.5	1	108 (GAUAAA)
his-12	ZK131.6	1	97 (AAUAAA)
his-13	ZK131.7	1	120 (AAUAAA)
his-14	ZK131.8	1	114 (AAUAAA)
his-15	ZK131.9	1	111 (GAUAAA)
his-16	ZK131.10	2	58 (no signal), 119 (AAUAAA)
his-19	K06C4.11	2	50 (AAAACA), 93 (AAUAAA)
his-20	K06C4.4	3	63 (AAUGUA), 115(AAUAAA), 316 (GAUAAA)
his-21	K06C4.3	1	98 (AAUAAA)
his-22	K06C4.12	2	63 (AAUGUA), 117 (AAUAAA)
his-24	M163.3	1	236 (AAUAAA)
his-25	ZK131.2	3	98 (UGUAAA), 124 (AAUAAA), 151 (GAUAAA)
his-26	ZK131.1	1	114 (AAUAAA)
his-27	K06C4.13	1	222(AAUAAA)
his-28	K06C4.2	2	108 (AAUAAA), 219 (GAUGAA)
his-32	F17E9.10	2	115 (AAUAAA), 146 (AAUAAA)
his-34	F17E9.9	1	59 (AAUAAA)
his-35	C50F4.13	1	116 (AAUAAA)
his-36	C50F4.6	3	92 (AAUAAA), 100 (AAUAAA), 622 (AAUAAA)
his-37	C50F4.7	1	88 (AAUAAA)
his-40	NULL	1	128 (AAUAAA)
his-41	C50F4.5	3	92 (AAUAAA), 100 (AAUAAA), 622 (AAUAAA)
his-42	F08G2.3	1	276(AAUAAA)
his-43	F08G2.2	1	97 (AAUAAA)
his-44	F08G2.1	1	111 (GAUAAA)
his-45	B0035.10	1	116 (AAUGAA)
his-46	B0035.9	4	29 (no signal), 67 (no signal), 114 (AAUAAA), 155 (AAUACA)
his-47	B0035.7	2	115 (AAUAAA),171 (UAUAAA)
his-48	B0035.8	2	103 (AAUAAA), 115 (AAUAAA)
his-49	F07B7.5	1	120 (AAUAAA)
his-50	F07B7.9	2	108 (AAUAAA), 219 (GAUGAA)
his-51	F07B7.10	1	93 (AAUAAA)
his-52	F07B7.4	2	63 (AAUGUA), 117 (AAUAAA)
his-53	F07B7.3	2	50 (AAAACA), 93 (AAUAAA)
his-54	F07B7.11	3	63 (AAUGUA), 115 (AAUAAA), 316 (GAUAAA)
his-56	F54E12.3	3	29 (no signal), 67 (no signal), 114 (AAUAAA)
his-57	F54E12.5	1	104 (AAUAAA)
his-58	F54E12.4	1	103 (AAUGAA)
his-59	F55G1.2	1	295 (AAUGAA)



name	CDS	isoforms	3'UTR length (PAS)
his-60	F55G1.11	2	66 (no signal), 120 (AAUAAA)
his-61	F55G1.10	1	98 (AAUAAA)
his-62	F55G1.3	2	31 (no signal), 107 (AAUAAA)
his-63	F22B3.2	1	116 (AAUAAA)
his-66	H02I12.6	1	107 (AAUAAA)
his-68	T23D8.6	2	15 (AAUAAA), 100 (AAUAAA)
his-69	E03A3.3	1	90 (GAUAAA)
his-70	E03A3.4	1	106 (AAUAAA)
his-71	F45E1.6	1	163 (AAUAAA)
his-72	Y49E10.6	2	104 (AAUAAA), 213 (AAUAAA)
his-74	W05B10.1	1	162 (AAUAAA)

Table 2.S6: Cumulative list of polyadenylated 3'UTRs detected in histone genes

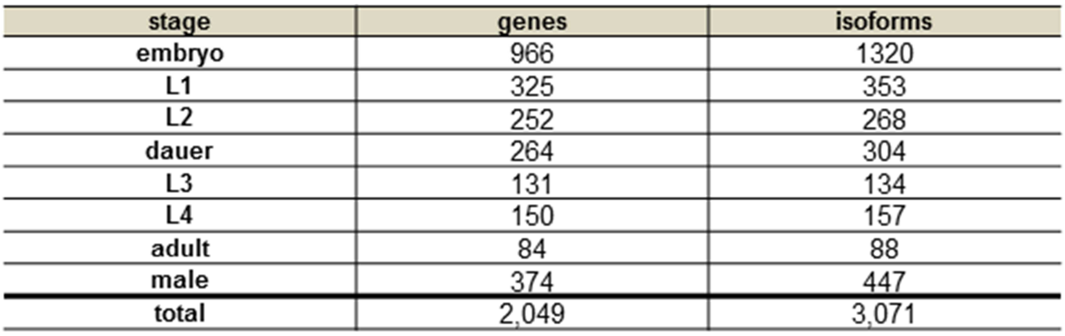
Summary of 3'UTR isoforms detected in histone genes, showing the putative PAS element for each representative 3’UTR. Nucleotides that deviate from the canonical PAS motif are highlighted in red.



A	# of 3'UTR isoforms	26,942	
B	# of unique 3'UTR regions	15,685	
C	average 3’UTR length	250 nt	
D	total 3'UTRome length	3,898,952 nt	
E	per nucleotide conservation rate of 3'UTR (3 species)	0.3	
F	probability of a conserved seed being functional	3 species	5 species
		0.56 ±0.01	0.64 ±0.03
G	# of unique conserved seeds identified	3 species	5 species
		5,673	1,744
H	# of unique miRNAs used for analysis (# of families)	183 (124)	
I	probability of a conserved miRNA seed site occurring inside an ALG-1 site	3 species	5 species
		0.75	0.76
J	probability of a randomly positioned 6-mer in a 3'UTR occurring inside an ALG-1 site	3 species	5 species
		0.43	0.45
K	# of conserved blocks not explained by predicted miRNA seeds or conserved PAS (5 species)	4,758	
L	# of 3'UTRs with at least one conserved block (5 species)	2,887	
M	probability of a conserved (randomly shuffled) sequence block of the same length inside an ALG-1 site	0.54 (0.48)	
M	fraction of Lall et al. 3’UTR:miRNA interactions recovered	0.83	
O	# of Lall et al. 3’UTR:miRNA interactions lost	1,111	
P	# of unique new interactions vs. Lall et al. miRNAs/3’UTRs	580	

Table 2.S7: Summary statistics for PicTar miRNA target predictions and other conserved sequence blocks in genomic regions spanned by the 3'UTRome compendium

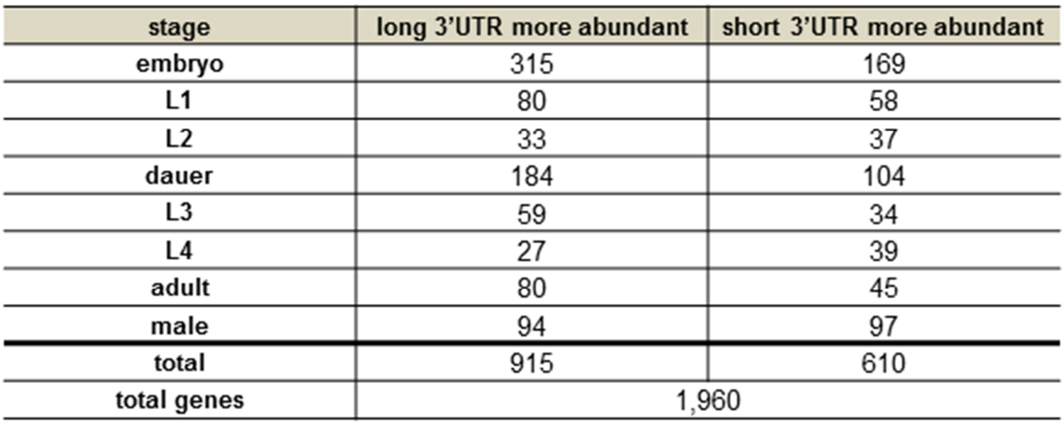
A) Total number of 3'UTRs used for miRNA target predictions. **B)** Number of unique 3'UTR regions, obtained by merging 3'UTRs with overlapping genomic coordinates. **C)** Average length of all unique 3’UTRs. **D)** The unique 3'UTRome comprises ~4M nucleotides. **E)** 30% percent of nucleotides in *C. elegans* 3’UTR are conserved in *C. remanei* and *C. briggsae*. Nucleotides in CDS, 5’UTR or intergenic regions were not considered in this analysis. **F)** Probability of a conserved miRNA seed being functional based on alignments of three or five species, obtained by creating artificial miRNAs resembling the original miRNAs (*18*) and comparing the number of target sites for the artificial miRNAs with the “real” target sites. **G)** Number of unique conserved miRNA seeds in the genome of three or five species. **H)** In total, 183 miRNAs were used. They comprise 174 miRBase (database release 14) miRNAs and 9 novel miRNAs determined by miRDeep2 (*17*), grouped in 124 miRNA families. **I)** The probability of a conserved miRNA seed within an ALG-1 binding site (*20*) in three or five species, calculated as the ratio of all miRNA target sites located in an ALG-1 binding site when considering only 3’UTRs that have an ALG-1 site and at least one miRNA target site. **J)** Probability of a shuffled seed sites (randomly positioned with the same 3’UTR) occurring within an ALG-1 binding site for three or five species. The probability is 30% less for shuffled sites than for the original miRNA seed position, signifying that miRNA seeds located in ALG-1 sites are indeed accurate signals. **K)** Number of conserved blocks, defined as at least 6 nt long and present in five species, that cannot be explained by a conserved predicted miRNA target seed site or a conserved PAS. **L)** Number of 3'UTR regions that contain at least one of such conserved blocks. **M)** Probability of a conserved block occurring within an ALG-1 binding site vs. randomly positioned blocks of the same length distribution within 3'UTRs is not significantly different. For analyses in K-M, regions overlapping a CDS in an alternative transcript were excluded. **N,O,P)** For the same miRNAs and 3'UTR regions, 83% of previously predicted miRNA target sites from Lall et al. (*19*) are identical with predictions using the empirically defined 3'UTRs in the 3'UTRome; 1,111 miRNA target sites are exclusively found in Lall et al., and 580 sites are newly predicted. Three species alignments always included *C. elegans, C. remanei*, and *C. briggsae*. Five species alignments also included *C. brenneri* and *C. japonica.* See Supplementary Materials and Methods for additional details.



stage	genes	isoforms
embryo	966	1320
L1	325	353
L2	252	268
dauer	264	304
L3	131	134
L4	150	157
adult	84	88
male	374	447
total	2,049	3,071

Table 2.S8: Number of genes present in multiple developmental stages but with stage-specific 3'UTR isoforms

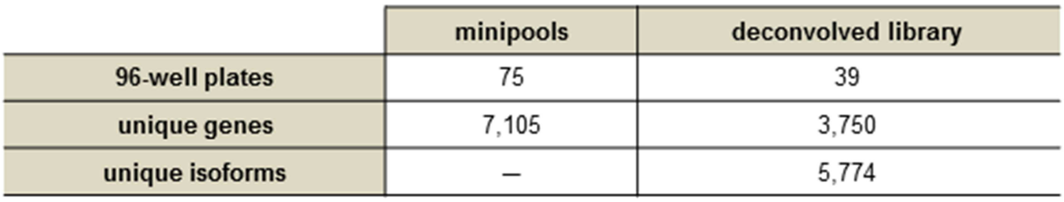
We have scanned the 3'UTRome for genes expressed in 1) at least two developmental stages, 2) with at least two 3'UTR isoforms, and 3) where one of these isoforms was stage-specific. The results shown here were used for the analysis described in Figure 4B.



stage	long 3'UTR more abundant	short 3'UTR more abundant
embryo	315	169
L1	80	58
L2	33	37
dauer	184	104
L3	59	34
L4	27	39
adult	80	45
male	94	97
total	915	610
total genes	1,960	

Table 2.S9: Number of genes with two 3'UTR isoforms detected in the staged polyA capture dataset

A subset of annotated genes from the polyA capture dataset with two 3'UTR isoforms used for the analyses in Figure 4. A 3'UTR isoform is defined as abundant if: 1) the total number of counts across all stages is larger than 5, and 2) if it is supported by at least twice the number of counts than the other 3'UTR isoform (see Supplementary Materials and Methods for details).



	minipools	deconvolved library
96-well plates	75	39
unique genes	7,105	3,750
unique isoforms	–	5,774

Table 2.S10: 3’UTR clones available in the 3’UTRome library

The 3’RACE approach produced sequence-validated 3’UTR clones that are available to the community to study 3’UTR biology. The UTR library collection will be updated on an ongoing basis and will expand to contain minipools and unique 3’UTR isoforms for all *C. elegans* 3’UTRs for protein-coding transcripts. See Supplementary Dataset S7 and the 3’UTR data repository http://www.utrome.org for clone availability.

Supplementary datasets from the publication:

These datasets can be obtained from the journal site at:
http://www.sciencemag.org/content/329/5990/432/suppl/DC1

Supplementary Dataset 2.S1. AceView genes in the 3'UTRome.

Comprehensive list of AceView genes with annotated 3'UTRs in the 3'UTRome. All gene names are linked to the current AceView annotation at NCBI (http://www.aceview.org). The file can be downloaded in HTML format.

Supplementary Dataset 2.S2. The complete 3'UTRome dataset.

A key is enclosed with Dataset S2 that describes all of the individual components.

Supplementary Dataset 2.S3. 3'UTR coordinates attached to AceView genes.

We used AceView gene annotations (http://www.aceview.org) *(7)* to map 1,490 unique, fully supported 3'UTR isoforms in genomic regions with either no annotated gene models or no compatible CDS ends in WormBase WS190. This table contains genome coordinates of 3'UTRs for these new genes. The file can be downloaded in Microsoft© Excel format.

Supplementary Dataset 2.S4. Convergently transcribed genes with overlapping 3'UTRs.

A list of genes in the 3'UTRome whose transcripts overlap (1 nt to 495 nt), indicating gene names, overlap length (nt), genome coordinates, and whether the two overlapping 3'UTRs are co-expressed in the same developmental stage. These data were used for the analysis described in Figure S14. The file can be downloaded in Microsoft© Excel format.

Supplementary Dataset 2.S5. List of genes displaying changes in 3'UTR length between developmental stages.

A comprehensive list of genes with two 3'UTR isoforms showing a change in the expression of long vs. short 3'UTR isoforms between developmental stages. All data are derived from the polyA capture dataset and are based on the number of Roche/454 read counts identified per 3'UTR end. The file contains two worksheets: The worksheet labeled "All genes–counts" lists the raw tag counts and counts normalized to the total counts in the embryo dataset. The worksheet labeled "All genes–relative abundance" shows the number of reads normalized within and across all developmental stages. Genes that exhibit 3'UTR isoform switching across developmental stages (shown individually in Supplementary Dataset S6) are indicated in the last two columns, labeled "potential isoform switch" and " 'high-confidence' isoform switch" (defined as a difference of ≥ 2-fold). See Supplementary Materials and Methods for details. The file can be downloaded in Microsoft© Excel format.

Supplementary Dataset 2.S6. Individual graphs of genes displaying 3’UTR isoform switching during development.

Individual graphs for 612 genes with two 3’UTR isoforms that exhibit a detectable switch in the expression of the long vs. short isoform across developmental stages, and with at least 20 total Roche/454 polyA tag counts per gene. All data are derived from the polyA capture dataset and are based on the number of Roche/454 read counts identified per 3'UTR end. For each graph the gene name, chromosome location, strand (in parentheses), genomic coordinate of the 3’UTR start, and lengths of the two 3’UTR isoforms are indicated. Green boxes highlight genes for which the relative abundance of 3’UTR isoform ‘*a*’ vs. ‘*b*’ is ≥ 2-fold in at least one particular stage and then ”switches” so that the ratio of ‘*b*’ vs. ‘*a*’ is ≥ 2-fold in another stage; in addition, the difference in expression between isoform ‘*a*’ and ‘*b*’ was required to be ≥ 5 counts. The cumulative list is given in Supplementary Dataset S5. See Supplementary Materials and Methods for details of the analysis. This file can be downloaded in Adobe© PDF format.

Supplementary Dataset 2.S7. The 3’UTRome clone library.

List of 3’UTR clones released. The clones are available to the community in the form of bacterial minipools and isolated 3’UTR isoforms. The library is cloned into the Gateway™ entry vector P2R-P3 and is compatible with the Promoterome (*3*) and ORFeome (*4,5*) libraries. The file can be downloaded in Microsoft© Excel format.

Chapter 3: Comparison of results of two parallel studies of developmental stage 3'UTRomes in *C. elegans*

3.1: Introduction

The previous chapter describes our efforts in generating a high quality genome-scale 3'UTRome of *C. elegans*. Our work was done as part of the modEncode project and incorporated data from various methods: PolyA capture, 3'RACE, full cDNA cloning and RNAseq transcriptome data from publicly available sources. [1] While this effort was in progress, a parallel group also proceeded to sequence the 3'UTRs of *C. elegans* in different developmental stages [2]. Both methods utilized high-throughput sequencing technology and identified 1000s of 3'UTRs, vastly increasing the number of known 3'UTRs in *C. elegans*. Since both groups worked independently, the common conclusions arrived at by both groups had immediate direct validation, and the differences could be solved by cross comparison. Since the results of both the methods are not complete subsets of each other, integrating the results of both will help the research community by providing a high quality validated dataset. Hence in this chapter I have attempted the comparison of results between both datasets and have performed a cross validation of my analysis with their results. I have also tried to answer criticisms in their report and arrived at quality control measures, which will reconcile the disparities between the methods. These new filters will be useful in analyzing further polyA capture sequence data without going through troublesome reviews.

3.2: Protocol differences between polyA-capture and 3P-seq methods

The first difference between the two methods was the sequencing method used. While our effort utilized pyrosequencing methods from Roche 454 life sciences, Jan et.al [2] used Illumina Solexa sequencing. The choice of method depends on preference of sequence quality, length and throughput. Pyrosequencing provides longer sequences (~250 nt) with higher quality, but at lower throughput (~300,000 sequences), while Illumina sequencing provides shorter sequences of the order of ~36 nt with lower quality at the 3'end. However it provides higher throughput (~4-5 million reads). This is the first difference noticeable comparing the raw data of the two methods (Fig 3.1A). Jan et. al has almost ten times the throughput as ours. However there is a tradeoff for choosing high throughput. Our sequences are long enough to faithfully map back to the genome, so we can confidently define the 3'end of the 3'UTR and in most cases the 5'end if it reaches the CDS end. With smaller Illumina sequences, the 5'end had to be deduced computationally.

The next major difference comes from the actual protocol of the 3'UTR capture. 3P-seq (PolyA Position Profiling by sequencing) protocol begins with a splint ligation that binds to the ends of the polyA tail and appends a biotinylated primer to the polyA tail. Partial digestion with RNAse T1 (which cuts after Gs) leaves the 3' end portion of the 3'UTR and the polyA tail attached to the primer. Reverse transcription with dTTP of the sequence antisense to the polyA tail is followed by partial digestion with RNAse H which cuts double strands. The remaining sequence will have the 3'end of the 3'UTR plus a few residues of adenine at the tail. This fragment is then amplified and sequenced. Our polyA capture protocol involves binding of an dinucleotide (NV)-anchored oligo(dT) biotinylated primer to the 3'end of the 3'UTR. Reverse strand cDNA synthesis followed by DpnII restriction enzyme digestion results in fragments (max length ~250 nt) with a

5'Dpn II GATC site and part of the 3'UTR and a 20nt polyA sequence at the 3' end which can be barcoded. This is then amplified and sequenced. The usage of oligo (dT) primer in polyA capture has been criticized in Jan et.al since it could facilitate binding to A-rich genomic regions and result in sequencing of non-3'UTR regions called "false priming". Another difference in the protocols is that in polyA capture, we selectively filter abundant ribosomal genes which occupy ~40-50% in the library. However, this filtering is not done in 3P-seq and hence just a few ribosomal genes dominate 18% of their data.

3.3: Sequence processing

There are differences in the way the sequences are processed between the two methods (Fig 3.1B). The sequences obtained in both methods are mapped to the WS190 genome. Whereas we used both unique and multiple loci matching sequences, Jan et. al used only those sequences that mapped to a unique location. Both approaches used a similar clustering algorithm to account for heterogeneity at 3'ends and abundance filters to remove low abundant tags. However, there is a difference in the evidence filter used. Since our data has more than one data source, we considered 3'UTRs that are either present in more than one source, or if it is from only one source then it should come from more than one stage (Fig 3.1C). To remove artifacts due to false priming we enforce evidence in more than one source. 3P-seq only has one source, and so they pick 3'UTRs that are present in more than one library and if it is only present in one library then it should have evidence in 2 different locations in the same cluster. Another difference is in the annotation of a 3'end to a gene. In most cases our reads reach the upstream CDS end and if not we annotate it to the nearest upstream end. In 3P-seq, an end is annotated to a gene only if it has RNAseq transcriptome evidence for a transcript there. Jan et. al also used an abundance constraint that the tag abundance should be >=5% of the average mRNA abundance and should be >=1% of the total abundance of all 3'UTRs for that gene. Also if a tag has another tag within 40nt then it is ignored. These differential filters could change the way a 3'UTR is called between the two methods, suggesting computational processing-based differences, but not necessarily differences in biological conclusions. Based on these filters we can see that 56% of the raw reads in 3P-seq have been filtered (Fig 3.1D) which could have contained valid data.

3.4: Comparison of 3'UTR overlap between 3P-seq and 3'UTRome

To find the overlap between the two datasets, I downloaded the raw and processed data provided by Jan et al. First, comparing the processed annotations of Jan et al with our annotations, we see overlap in ~13,000 genes (Fig 3.2A). There were 1,980 genes uniquely seen in our dataset, while ~3,056 genes were seen uniquely in 3P-seq. However when we look at isoforms, only ~15,825 isoforms out of 27,971 total annotated isoforms were seen to overlap between the datasets. ~12,100 3'UTRs were seen uniquely in our data while ~9,800 3'UTRs were unique to 3P-seq. These non-overlapping 3'UTRs had to be accounted for and Jan et al raised criticisms saying ~3,500 of these 3'UTRs may be enriched in false primed 3'UTRs. To examine whether the 3'UTRs unique in our dataset were false primed 3'UTRs or were filtered in the 3P-seq data by the computational filters (making them likely true positives), I looked for the evidence of our 3'UTRs in the 3P-seq raw data before processing with overlap within 20nt of the two ends. Surprisingly, I found evidence for >7,300 out of 12,146 3'UTR isoforms of our unique 3'UTRs within 20nt of their raw 3'UTR ends (Fig 3.2B). A majority of these 3'UTRs seem to have been filtered by their >2 downstream A filter. Based on this, I had two subsets of our 3'UTRs, a) present in the 3P-seq processed data (termed 3P-seq 3'UTR from now on) and b) our 3'UTRs present in 3P-seq raw data. Comparing these subsets with our data, I derived two non-overlapping datasets: ~12,000 3'UTRs non-overlapping with their annotated 3'UTRs("set A") and ~4,800 3'UTRs non-overlapping with their raw 3'UTR ends (set B)

3.4.1: Comparison of source distribution of 3'UTRs between 3P-seq and 3'UTRome

One of our major evidence filters was that the 3'UTR should be seen in more than one data source (i.e., polyA capture, 3'RACE, full length cDNA and RNAseq). Hence for the ~12,000 3'UTRs non-overlapping with their annotated 3'UTRs("set A") and ~4,800 3'UTRs non-overlapping with their raw 3'UTR ends (set B) the first thing we wanted to see was the source from which they were derived. To that end, we looked at the source distribution of both the set A (Fig 3.3A, left panel) and set B (Fig 3.3A right). We see that 72% of the 12,000 3'UTRs and 66% of the 4,800 3'UTRs have evidence in at least one other source. Even if we assume polyA capture to contain a high incidence of false priming, the other sources don't have this artifact. Based on this, at least ~65-70% (8,699 out of 12,146 3'UTRs of set A and 3,160 out of 4,815 3'UTRs in set B) of the non-overlapping 3'UTRs are likely valid because they were observed in an alternate source.

3.4.2: Comparison of PAS distribution of 3'UTRs between 3P-seq and 3'UTRome

Another criticism in Jan et al was that the 3' UTRs that do not contain a canonical or variant PAS motif (i.e. "no PAS" 3'UTRs) in the non-overlapping set could be due to false priming. To address this claim, we examined the PAS distribution of the overlapping and non-overlapping 3'UTRs in our dataset (Fig 3.3B). Out of the 12,146 non-overlapping 3'UTRs in set A, 32% seem to be enriched in the no PAS category and out of the ~4,819 non-overlapping 3'UTRs in set B, 40% seem to be enriched in the no PAS category. This high percentage is indeed enriched compared to the ~6% and ~13%

no PAS 3'UTRs in the overlapping 3'UTRs (Fig 3.3B). This indicates that their criticism might be valid and stringent filtering criteria should be employed in this class of 3'UTRs. Looking at the source distribution we see that out of the 3,886 no PAS 3'UTRs from the 12,146 non overlapping 3'UTRs in set A, 2,907 3'UTRs are arising from a single source namely, 1,119 from cDNA, 1,081 from polyA capture and 679 from 3'RACE. Similarly out of the 1,931 no PAS 3'UTRs from the 4,819 non overlapping 3'UTRs in set B, 1,656 3'UTRs are arising from a single source namely, 583 from cDNA, 555 from polyA capture and 507 from 3'RACE. Out of these, the cDNA libraries have no false priming artifacts. PolyA capture and 3'RACE are likely to contribute to the false priming and these constitute (1,081+ 679 =1,760) 14.5% of 12,146 non overlapping 3'UTRs in set A and (555 + 507= 1,062) or 22% of 4,819 non overlapping 3'UTRs in set B.

3.4.3: Comparison of adenine composition downstream of 3'UTR ends between 3P-seq and 3'UTRome

To identify the reasons for potential false priming, the first step was to examine the downstream composition of the 3'UTR ends. Based on the number of A's seen immediately downstream we see that while the overlapping 3'UTRs show a decreasing number of As, the non-overlapping 3'UTRs show an increasing number of As especially after 8 As in the case of both the 12,146 and the 4,819 non-overlapping 3'UTRs (Fig 3.4A left and right). The increased downstream number of As could indeed be a likely cause of false priming. While only 250 out of 15,825 3'UTRs overlapping with 3P-seq dataset in set A (948 out of 23,156 3'UTR3'UTRs in set B) derived from single source (polyA capture or 3'RACE), 1,760 out of 12,146 3'UTRs were derived from single source in the non-overlapping 3'UTRs in set A (1,062 out of 4,819 3'UTRs in set B). Looking at the percentage distribution of the 3'UTRs between the overlapping and non-overlapping in set A, we see that 75% (11,885 out of 15,825) of the 3'UTRs in the overlapping set have less than or equal to 8 downstream As and 90%(14,423 out of 15,825) of the 3'UTRs in the overlapping set have less than or equal to 10 downstream As. However only 60% (7,425 out of 12,146) of 3'UTRs in the non-overlapping set has less than or equal to 8 downstream As. Furthermore, if we look at the 1,760 potential false primed 3'UTRs in set A arising from the noPAS 3'UTRs derived from a single source (polyA capture + 3' RACE), only 55% (972) of the 3'UTRs have less than or equal to 8 downstream As. Similarly only 50% (552 out of 1,062) of the potentially false primed noPAS 3'UTRs in set B arose from a single source. A method to estimate the sufficient number of A's downstream to false prime is provided later in this chapter.

3.4.4: Comparison of 3'UTR length distribution between 3P-seq and 3'UTRome

The criticism in Jan et al postulated that the proximal UTRs in our dataset, i.e. the shorter 3'UTRs, may be enriched with false primed 3'UTRs. Therefore, we examined the length distribution of the 3'UTRs in both the overlapping and non-overlapping datasets. Looking at overlap with both their processed data and their raw data shows similar 3'UTR length profile between the overlapping and the non-overlapping 3'UTRs.(Fig 3.4B). Therefore, we conclude that the false priming may not be biased on the 3'UTR length.

3.4.5: Derivation of false priming filter criteria

The results of the previous sections suggest that while there is some evidence of false priming in the non-overlapping dataset, the extent of the false priming may not be as extensive as Jan et. al suggested. To obtain a better understanding, I used the modENCODE consolidation data from Gerstein et al [3] that was used for comparison in the Jan et al. study. The next few analyses were performed using the Gerstein et al data with their annotations so that I could directly compare results of my analysis with the Jan et al analyses. There are some differences in using this dataset. First, this dataset only incorporates ~8,500 3'UTRs from our dataset and a similar number from 3P-seq (Fig 3.5A). The source distribution of the 3'UTRs shows similar patterns between our data and 3P-seq. ~3,800 3'UTRs are unique to polyA capture and 3P-seq each (Fig 3.5B).

I performed overlap analysis on this dataset using the same method as described in section 3.4. The data was split into the following categories. Based on overlap between the polyA capture and 3p-seq datasets, 3'UTRs were designated as unique or not unique. Based on the number of isoforms per gene, genes with one 3'UTR isoform were designated as single 3'UTRs. For genes with multiple 3'UTRs isoforms, the longest was designated as the distal 3'UTR and the shortest was designated as the proximal 3'UTR. Altogether, 2,398 3'UTRs were derived from single isoform genes, 2,248 3'UTRs were derived from two isoform genes, 1,323 3'UTRs were derived from three isoform genes and 1,996 3'UTRs were derived from genes with more than three isoforms.

I first compared the PAS distribution between single 3'UTRs unique and non-unique to each study. The overlapping "single not unique" 3'UTRs (959 3'UTRs) show similar PAS profile between polyA-capture and 3P-seq (Fig 3.6A). However, the non-overlapping "single unique" 3'UTRs (1,427 3'UTRs) show twice the number of no PAS incidences in our data (374 out of 821 3'UTRs in polyA capture compared to 142 out of 606 3'UTRs in 3P-seq) (Fig 3.6B). This suggests an enrichment of noPAS 3'UTRs in polyA capture which could arise from false priming artifacts. Next, I compared the PAS distribution of overlapping distal and proximal not unique 3'UTRs and non-overlapping distal and proximal unique 3'UTRs. Similar to the single 3'UTRs, the overlapping 3'UTRs show similar PAS profiles between our data and 3P-seq (Fig 3.7A,B). However, the non-overlapping dataset shows increased no PAS occurances in distal (200 out of 459

3’UTRs in polyA capture compared to 131 out of 552 3’UTRs in 3P-seq) and proximal (1,202 out of 2,567 3’UTRs in polyA capture compared to 855 out of 2,412 3’UTRs in 3P-seq) unique 3’UTRs in the polyA-capture dataset (Fig 3.7C,D). This data suggests differences between our data and 3P-seq occur mainly at the no PAS sites and the discrepancy is not dependent on 3’UTR length as postulated by Jan et al. [2]

Understanding one’s own data and knowing the limitations in it helps to obtain maximum information output. Thus far, we conclude that the issue in the data results from the non-overlapping noPAS 3’UTRs that are derived from single sources, namely polyA capture and 3’RACE**.** The next step was to derive the filters that would remove this anomaly. For this, we need to know where our data deviates from 3P-seq. The first step is to look at the downstream A distribution of the 3’UTR end. I first looked at the distribution of the number of consecutive A’s in the 20nt downstream region of the 3’ end. To look at deviation, I calculated the percentage difference between polyA capture and 3P-seq for different numbers of downstream consecutive As. The difference was calculated for the single, proximal and distal unique 3’UTRs which seemed to show false primed 3’UTRs. The plot shows peaking at 4 consecutive As for single, proximal and distal cases. (Fig 3.8A) This suggests that a filter set at 4 consecutive A’s in the 20 nt immediately downstream of the 3’UTR end would limit false primed 3’UTRs and bring the distribution closer to 3P-seq data.

Next we looked at the total number of A’s in the 20 nucleotides downstream that may trigger false priming. Here I calculated the deviation between the two datasets as the distribution of difference in percentage of the number of 3’UTRs between polyA capture and 3P-seq for single, proximal and distal unique 3’UTRs as before for the total number of A’s seen in a 20nt window immediately downstream of the 3’end of the 3’UTR. The difference between polyA capture and 3P-seq becomes positive for all cases after 8 total A’s. (Fig 3.8B), suggesting that a filter for total number of A’s in the 20 nucleotides downstream at 8 would uniformly remove a large portion of false primed candidates across all three cases.

Next we looked at the total number of A’s in a moving 8 nt window. For this we calculated the deviation between the two datasets as the difference in percentage distribution of the number of 3’UTRs between polyA capture and 3P-seq for single, proximal and distal unique 3’UTRs as before for the total number of As seen in an 8nt moving window which slides along the 20nt immediate downstream region of the 3’UTR end. The difference between polyA capture and 3P-seq becomes positive for all cases after 5As in a 8nt moving window(Fig 3.8C). This suggests that a filter set at limiting the total number of A’s to 5 in a 8 nt window would be effective at removing false primed artifacts across single and multiple isoform cases.

Based on the above results, the following filtering rules were derived:

- No more than 4 consecutive A’s in the 20 nt downstream of the 3’end
- No more than 8 total A’s in the 20 nucleotides downstream of the 3’end
- No more than 5 total A’s in an 8nt window in the 20 nt downstream of the 3’end

The next step was to identify how the application of these filters will affect the output. For this I tried a two way approach. First, I applied the filters for all the 3’UTRs in our dataset and the 3P-seq dataset. This was termed “full filter” in (Fig 3.8D). Based on this we have two datasets – our 3’UTRs passing the full filter and 3P-seq 3’UTRs passing the full filter. On a second approach I only applied the filters on the no PAS 3’UTRs which were enriched for the false primed 3’UTRs. This was termed “selective filter” (Fig 3.8D). From this, we also have two datasets: our 3’UTRs passing the selective filter and 3P-seq 3'UTRs passing the selective filter. Comparing the results we see that applying the filters on the whole dataset results in the loss of 2,753 (32% loss) 3'UTRs from polyA capture and 2,232 (26% loss) 3'UTRs from 3P-seq. This not only removes 903 noPAS 3'UTRs from polyA capture and 445 noPAS 3'UTRs from 3P-seq but also removes 315 valid AATAAA 3'UTRs and 1,537 3'UTRs with alternative PAS from polyA capture and 347 AATAAA 3'UTRs and 1,440 3'UTRs with alternative PAS from 3P-seq(Fig 3.8D). Hence, this full filtering removes valid 3'UTRs in addition to false primed 3'UTRs, suggesting that the rules may be better selectively applied. Selective filtering brings our dataset closer to the 3P-seq distribution by removing 903 noPAS 3'UTRs from polyA capture and 445 noPAS 3'UTRs from 3P-seq. The similar percentage distribution between selectively filtered polyA capture and 3P-seq clearly shows the effectiveness of the filters in removing spurious 3'UTRs without loss of useful information . Hence I suggest use of the FP filters on the no PAS 3'UTRs on our future datasets.

3.5: Conclusion

In this chapter I consolidated and compared the results of two parallel studies in developmental stage specific 3'UTRomes. The comparison showed similarities and differences between the two protocols and their computational processing. Looking at the data in a non-biased way shows the flaws in each method and how they could be improved. Both methods identified thousands of 3'UTRs that were previously not annotated before and were greatly helpful in generating a comprehensive 3'UTRome of *C. elegans*. First, the consolidation shows that while there are ~15,825 3'UTRs common to polyA capture and 3P-seq there were 12,146 3'UTRs which were unique to polyA capture and had no evidence in the annotated 3'UTR list provided with 3P-seq. Further analysis shows evidence for 7,331 3'UTRs in the raw 3P-seq data which were filtered out by stringent computational filtering. Second, answering criticism that the polyA capture unique 3'UTRs are enriched for false primed 3'UTRs and are mostly proximal noPAS 3'UTRs, we do see enrichment for false primed 3'UTRs in the noPAS 3'UTRs unique to polyA capture. However the scale of enrichment is not as high as reported by Jan et.al and it is not biased on the length of the 3'UTR. We see similar enrichment in single, proximal and distal 3'UTRs. Third, analyzing the consolidation provided by modEncode we were able to derive rules which will effectively remove the false primed 3'UTRs from polyA capture datasets. Finally, selectively applying these rules to noPAS 3'UTRs, we were able to show a similar distribution between polyA capture and 3P-seq. With this we were able to answer the criticism about our data and efficiency of the polyA capture protocol, and show that with just a few computational filters we could achieve similar results as 3P-seq.

3.6: Reference

1. Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V *et al*: **The Landscape of C. elegans 3'UTRs**. *Science* 2010, **329**(5990):432-435.
2. Jan CH, Friedman RC, Ruby JG, Bartel DP: **Formation, regulation and evolution of Caenorhabditis elegans 3′UTRs**. *Nature* 2010, **469**(7328):97-101.
3. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K *et al*: **Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project**. *Science* 2010, **330**(6012):1775-1787.

3.7: Figures

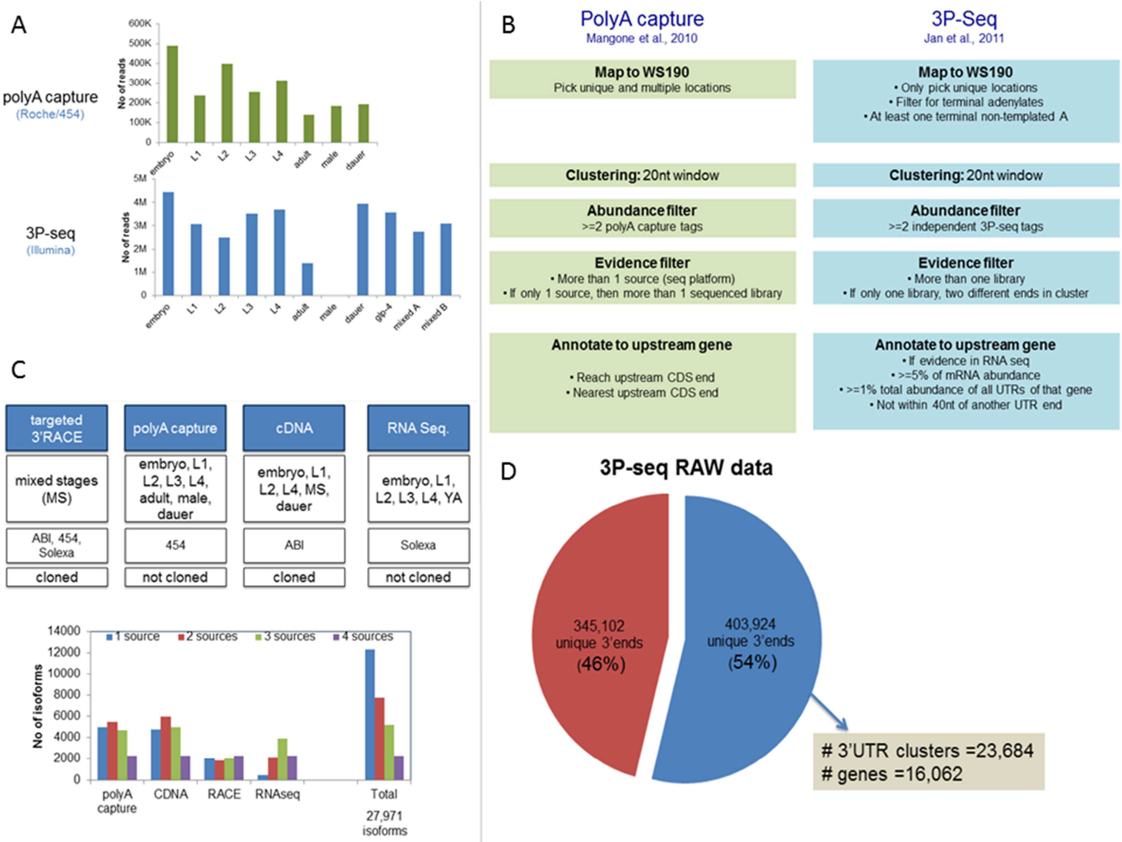


Figure 3.1: Raw data comparison between polyA capture and 3P-seq
A: Raw sequence abundance between polyA capture and 3P-seq methods.

B: The schematic processing pipeline of polyA capture and 3P-seq sequences.

C: The top panel: Data sources that comprise the polyA capture dataset. The bottom panel: Overlap of 3'UTRs between different datasources

D: Numbers of filtered and non-filtered sequence reads in 3P-seq.

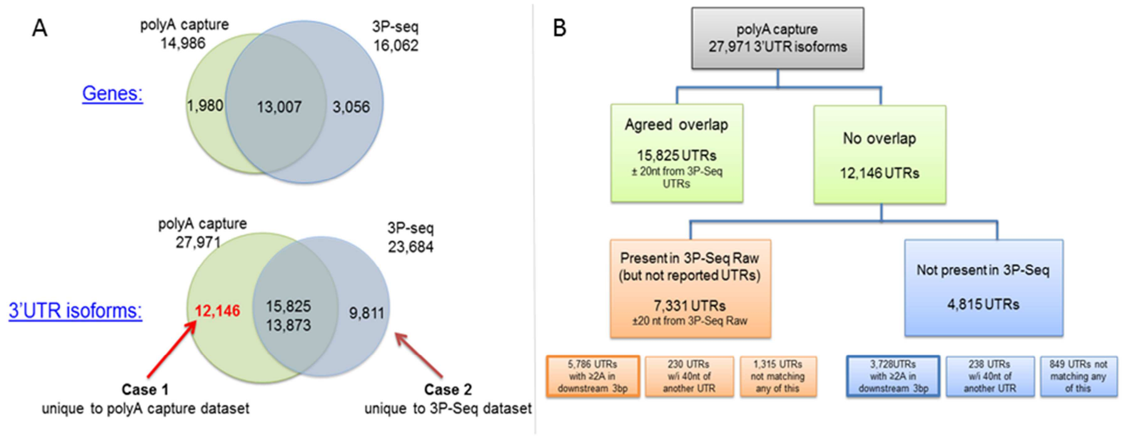


Figure 3.2: Overlap of 3'UTRs between polyA capture and 3P-seq
A: Overlap of genes between polyA capture and 3P-seq (top panel). Isoform overlap between polyA capture and 3P-seq (bottom panel).

B: A hierarchical representation of the overlap between the polyA capture data and 3P-seq data.

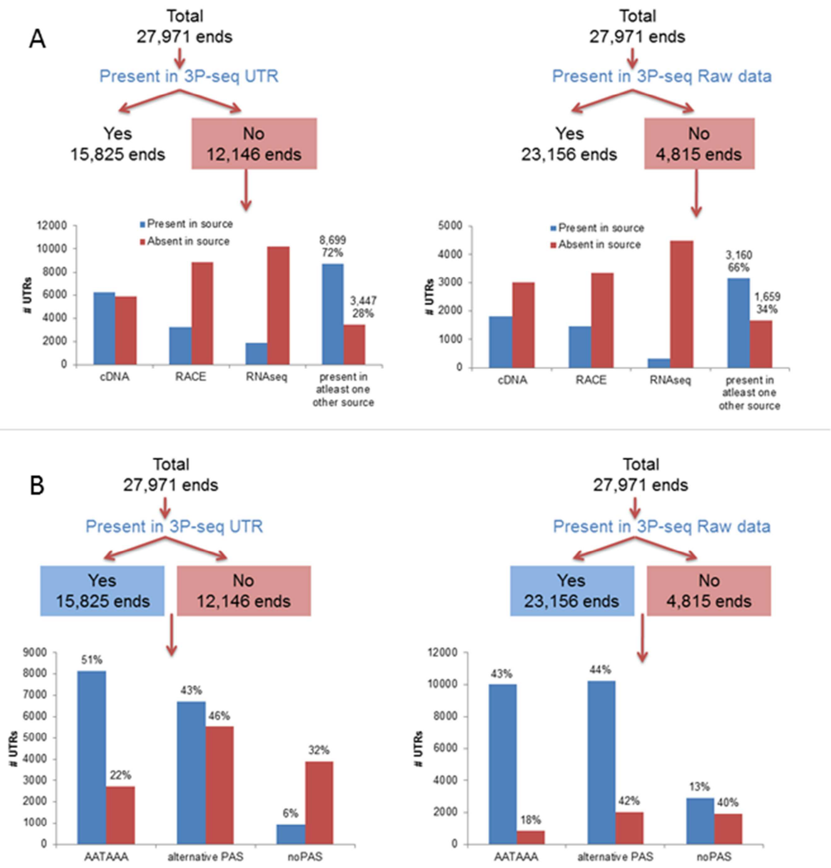


Figure 3.3: Source and PAS distribution of polyA capture 3'UTRs with respective to processed and raw data of 3P-seq

A: Source distribution of the 3'UTRs non-overlapping with 3P-seq processed data (left panel) and 3P-seq raw data (right panel).

B: PAS distribution of the 3'UTRs overlapping and non-overlapping with 3P-seq processed data (left panel) and 3P-seq raw data (right panel).

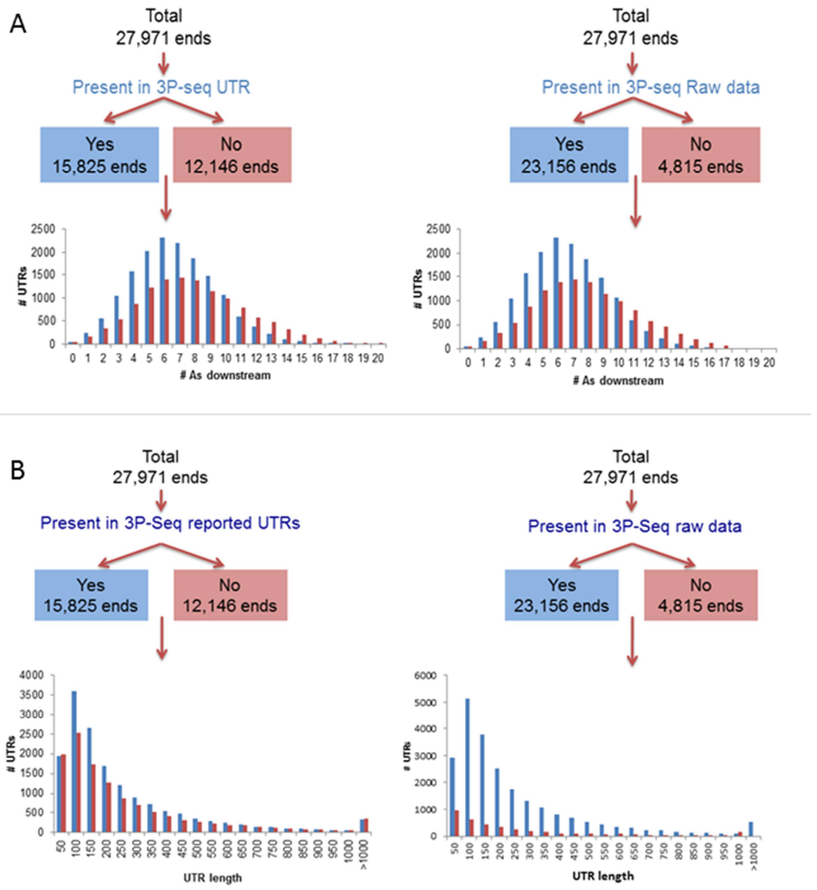
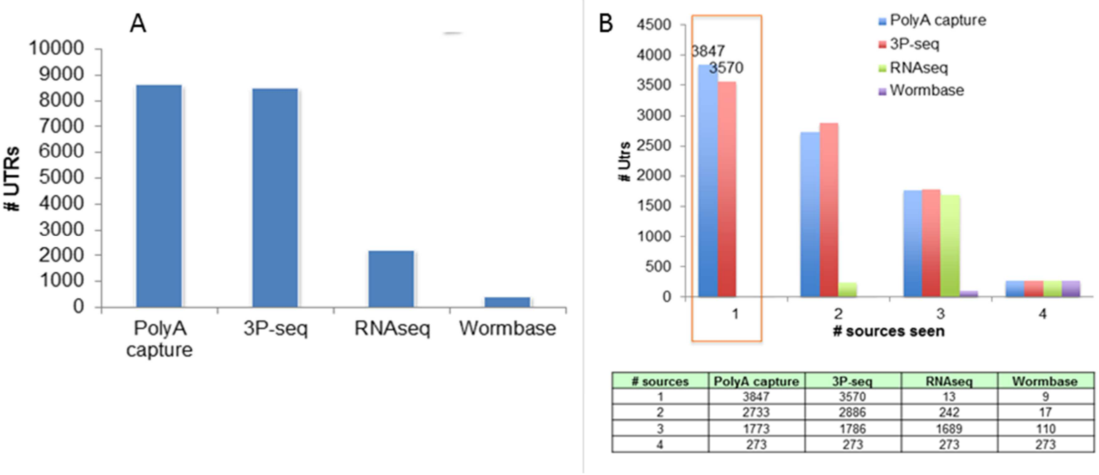


Figure 3.4: PAS site position and UTR length distribution of polyA capture 3'UTRs with respective to processed and raw data of 3P-seq

A: PAS position distribution of the 3'UTRs overlapping and non-overlapping with 3P-seq processed data (left panel) and 3P-seq raw data (right panel).

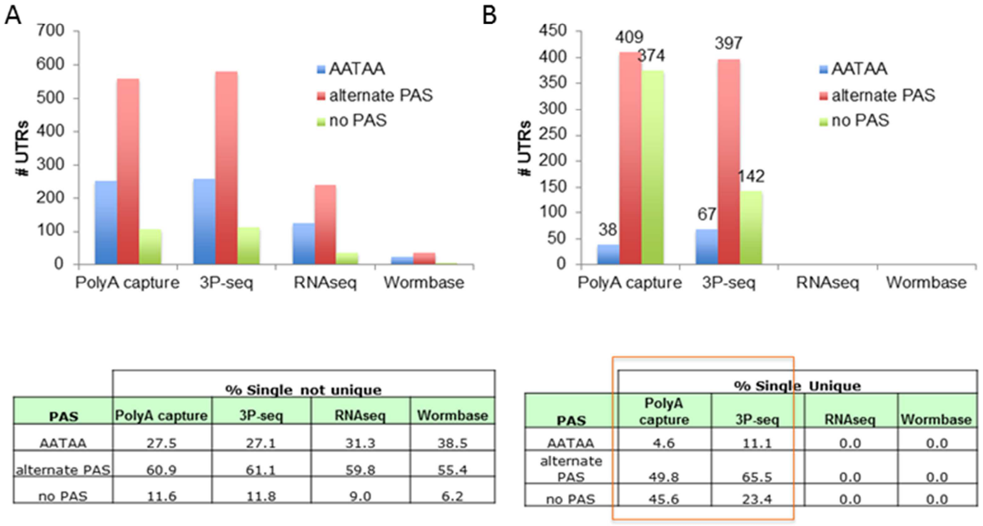
B: Length distribution of the 3'UTRs overlapping and non-overlapping with 3P-seq processed data (left panel) and 3P-seq raw data (right panel).



# sources	PolyA capture	3P-seq	RNAseq	Wormbase
1	3847	3570	13	9
2	2733	2886	242	17
3	1773	1786	1689	110
4	273	273	273	273

Figure 3.5: Source distribution of 3’UTRs obtained from Gerstein et.al[3]
A: Number of 3’UTRs represented by each source as consolidated by Gerstein et al

B: This plot provides frequency distribution of the 3’UTRs from each source overlapping with the UTRs identified in the other sources



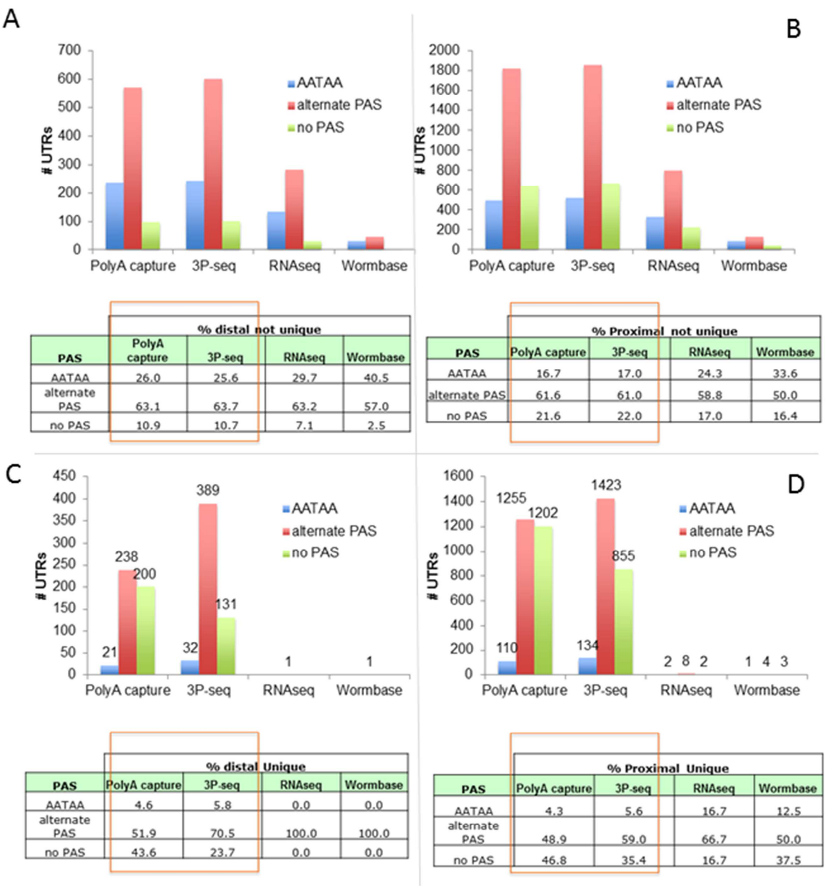
	% Single not unique			
PAS	PolyA capture	3P-seq	RNAseq	Wormbase
AATAA	27.5	27.1	31.3	38.5
alternate PAS	60.9	61.1	59.8	55.4
no PAS	11.6	11.8	9.0	6.2

	% Single Unique			
PAS	PolyA capture	3P-seq	RNAseq	Wormbase
AATAA	4.6	11.1	0.0	0.0
alternate PAS	49.8	65.5	0.0	0.0
no PAS	45.6	23.4	0.0	0.0

Figure 3.6: PAS distribution of UTRs from single isoform genes obtained from Gerstein et.al[3]

A: PAS distribution for the single isoforms that are not unique to a dataset. The bottom panel gives the % distribution for the plot in the top panel.

B: PAS distribution for the single isoforms that are unique to a dataset. The bottom panel gives the % distribution for the plot in the top panel.



	% distal not unique			
PAS	PolyA capture	3P-seq	RNAseq	Wormbase
AATAA	26.0	25.6	29.7	40.5
alternate PAS	63.1	63.7	63.2	57.0
no PAS	10.9	10.7	7.1	2.5

	% Proximal not unique			
PAS	PolyA capture	3P-seq	RNAseq	Wormbase
AATAA	16.7	17.0	24.3	33.6
alternate PAS	61.6	61.0	58.8	50.0
no PAS	21.6	22.0	17.0	16.4

	% distal Unique			
PAS	PolyA capture	3P-seq	RNAseq	Wormbase
AATAA	4.6	5.8	0.0	0.0
alternate PAS	51.9	70.5	100.0	100.0
no PAS	43.6	23.7	0.0	0.0

	% Proximal Unique			
PAS	PolyA capture	3P-seq	RNAseq	Wormbase
AATAA	4.3	5.6	16.7	12.5
alternate PAS	48.9	59.0	66.7	50.0
no PAS	46.8	35.4	16.7	37.5

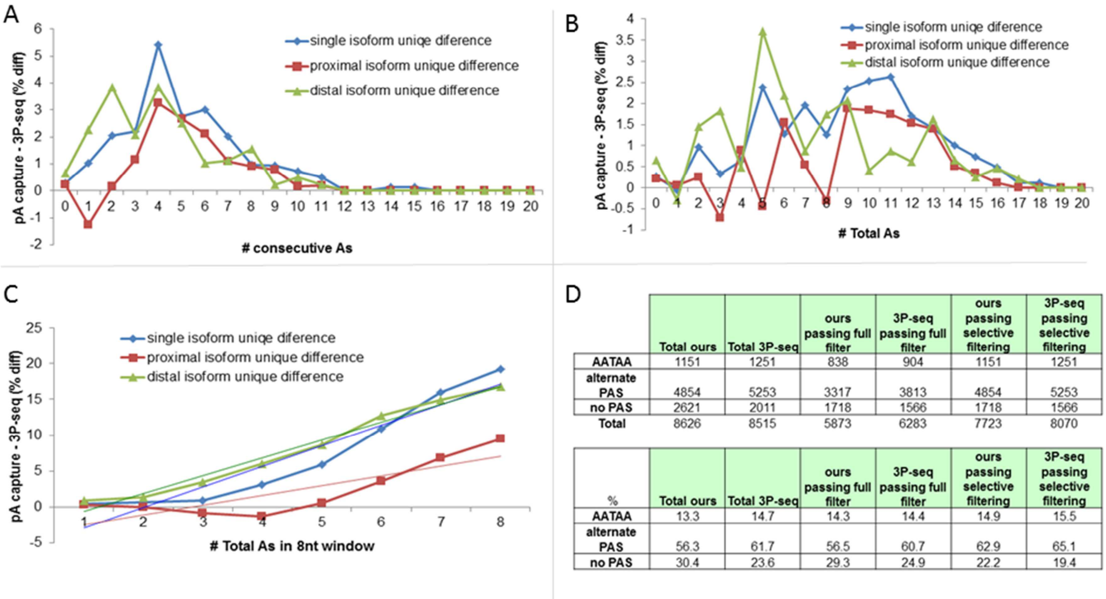
Figure 3.7: PAS distribution of UTRs from multiple isoform genes obtained from Gerstein et.al[3]

A: PAS distribution for the distal isoforms that are not unique to a dataset. The bottom panel gives the % distribution for the plot in the top panel.

B: PAS distribution for the proximal isoforms that are not unique to a dataset. The bottom panel gives the % distribution for the plot in the top panel.

C: PAS distribution for the distal isoforms that are unique to a dataset. The bottom panel gives the % distribution for the plot in the top panel.

D: PAS distribution for the proximal isoforms that are unique to a dataset. The bottom panel gives the % distribution for the plot in the top panel.



	Total ours	Total 3P-seq	ours passing full filter	3P-seq passing full filter	ours passing selective filtering	3P-seq passing selective filtering
AATAA	1151	1251	838	904	1151	1251
alternate PAS	4854	5253	3317	3813	4854	5253
no PAS	2621	2011	1718	1566	1718	1566
Total	8626	8515	5873	6283	7723	8070

%	Total ours	Total 3P-seq	ours passing full filter	3P-seq passing full filter	ours passing selective filtering	3P-seq passing selective filtering
AATAA	13.3	14.7	14.3	14.4	14.9	15.5
alternate PAS	56.3	61.7	56.5	60.7	62.9	65.1
no PAS	30.4	23.6	29.3	24.9	22.2	19.4

Figure 3.8: comparison between polyA capture and 3P-seq to derive false priming filter

A: Percentage deviation between polyA capture and 3P-seq for different numbers of consecutive As in the 20nt downstream of the 3'end.

B: Percentage deviation between polyA capture and 3P-seq for different numbers of total As in the 20nt downstream of the 3'end.

C: Percentage deviation between polyA capture and 3P-seq for different numbers of total As in an 8nt window in the 20nt downstream of the 3'end.

D: The number of 3'UTRs from polyA capture and 3P-seq that pass the false priming filters. When the full dataset is filtered it is called "Full filter" and when only the no PAS is filtered it is called "selective filter".

Chapter 4: Generation of pathway specific 3’UTRomes in axonal and synapse development

4.1: Introduction

Traditional studies in neuron development have focused on transcriptional control. However post-transcriptional regulations, including alternative splicing, alternative polyadenylation, and regulation by small RNAs, have emerged as important mechanisms to generate additional layers of diversity in neuronal plasticity models. A number of studies demonstrate alternative 3’UTR selection to be a critical determinant during neuronal development. For example, brain derived neurotrophic factor (BDNF) exhibits two different 3’UTRs and during neuron rest the long 3’UTR is repressed and the short 3’UTR is translated. When a neuron is activated during seizure, the long 3’UTR is translated while the short 3’UTR is repressed [1]. Similarly, in rat hippocampus, GluR2 mRNAs exhibit two isoforms. While the short isoform is translated during neuron rest, stimulus resulting in neuron activation changes the translation to the longer 3’UTR [2]. This phenomenon is precise enough to regulate a single neuron in *C. elegans* where *die-1* 3’UTR is differentially repressed in just the ASER chemosensory receptor neuron while being active in the ASEL neuron [3]. These events highlight the precision of alternative polyadenylation-mediated regulation and its widespread nature across different organisms in development and in functioning of neurons.

Synapse development in *C. elegans* has been shown to be impaired by mutagenesis of *rpm-1[4], syd-1[5]* and *syd-2[6]* genes. But this does not affect the locomotion of the animal. However, when *rpm-1* and *syd-2* are both mutated, those animals have defective synapses and uncoordinated movement [7, 8]. This shows the existence of parallel pathways in synapse development. Recent work by our collaborators in the Yishi Jin lab at UCSD identified *sydn-1,* which is vital for synapse and axon development, and functions in an alternate pathway from *rpm-1 and syd-2* [9]. In addition, mutation of *pfs-2*, which encodes a factor in the 3'end processing machinery, suppresses the *sydn-1* mutant phenotypes in axon and synapse development. *pfs-2* is a member of the conserved WD Repeat protein, and in yeast it is a part of the CF II/PF I complex, interacting directly with the subunits of the CF II/PF I and CF IA complexes to play a bridging role in the assembly of the polyA complex [10]. Loss of *pfs-2* in fission yeast also resulted in chromosomal segregation defects and lethality in addition to defects in mRNA 3'end processing [11]. The genetic interaction of *pfs-2* and *sydn-1* places a critical role for 3'end processing machinery during neuronal development.

In this chapter, I aimed to computationally study the role of *rpm-1* and *sydn-1* in 3' end processing in the context of synapse development. We sequenced the 3'UTRs from wild type, *rpm-1*, *sydn-*1, and *rpm-1;sydn-1* mutants using our polyA capture method [12]. Comparing wild type with these genetic mutants will provide insights into how differential 3'UTR formation and alternative 3'UTR isoform expression can affect axon and synapse development.

4.2: Materials and Methods

Strains. The Bristol N2 was used as the reference wild-type strain. Mutant alleles used in this study include: *rpm-1(CZ1252), sydn-1(CZ4741), rpm-1;sydn-1(CZ4738).* All four strains were synchronized and raised at 20℃. The s amples were collected at L1 stage

(~8hr post hatching). RNA preparation and polyA capture protocol was performed as previously described [12]. Deep sequencing was performed on the Genome Sequencer FLX system (Roche/454 Life Sciences, Branford, CT).

Sequence processing. The raw sequence reads were processed with a custom Perl script to remove the 5’ and 3’ sequencing linkers. Reads where the linker could not be identified and those with length <15nt after linker removal were discarded. The number of sequences passing this filter is given in (Fig 4.1A). Sequences ≥15 nt in length were aligned to the WS190 genome using BLAT [13], with a maximum intron size of 1000, minimum window size of 5, and maximum gap of 6. Best matches were selected, and multiple alignments reported if present in more than one genomic location. Sequence reads whose alignment did not map within 5 nt of the 3’end of the sequence were removed. The number of sequences passing this filter is given in (Fig 4.1A). The abundance of reads that mapped to multiple loci were normalized to the total number of genomic loci to which they map.

Clustering of 3’ ends. The 3’ ends of the alignments were clustered with a custom algorithm for iterative clustering to handle 3’end heterogeneity. The alignment clusters the ends within a 20 nt window and the most abundant 3’end is designated as the representative of that cluster. The sum of the reads in that cluster is defined as the abundance of that cluster. This process is iterated many times until there are no representative ends within 20 nt. From 408,377 distinct 3’ends, 24,109 clusters were defined. The number of clusters per library is given in Fig 4.1B.

False priming filter. To handle artifacts that occur due to false priming, we used the optimal filtering criteria derived in the previous chapter. Clusters whose ends had 5 or more consecutive A’s in the 20 nt downstream region were filtered. Clusters whose 20 nt downstream region had more than 8 A’s were filtered. Clusters whose 20 nt downstream region had more than 5 A’s in an 8 nt sliding window were filtered; clusters that failed this filter but had a valid PAS site upstream of the 3’end were still considered. The number of clusters passing this filter per library is given in Fig 4.1B.

Abundance filter. In order to remove low-abundant isoforms, any cluster with abundance of <2 reads in all the libraries were filtered. The number of clusters passing this filter is given in Fig 4.1B. This filter shows that a significant number of the clusters had fewer than 2 reads.

PAS motif analysis. The 50 nt regions immediately upstream of all polyA sites were scanned in an unbiased way for all possible 5 to 10-mer sequences to identify any statistically over-represented motifs. The only motifs returned from this exercise were the canonical PAS sequence (AAUAAA) and several closely related variants. The distribution of all over-represented hexamers peaked at a start position of -19 nt from the polyA site, which was taken as the most likely position of the PAS site. All of the 3'UTR isoforms in the compendium were then scanned for the canonical PAS sequence and any hexamer with an edit distance of 1 or 2 nt. Because it is not possible to definitively identify the "real" PAS site, we scanned for hexamers in a preferred order based on their observed frequency of occurrence in bona fide 3'UTRs between 10 and 30 nt upstream of the polyA site. Those occurring at a frequency of ≥1% as putative PAS motifs were considered. We used the first occurrence of a putative motif in the ordered list as the most likely functional PAS sequence. 3’UTRs that did not contain one of the resulting 26 putative PAS motifs within this interval were termed “no PAS”. The results of the consolidated PAS analysis are given in Fig 4.1C.

4.3: Results

4.3.1: Defects in synapse biogenesis do not affect overall trends of 3’end formation.

Since it was previously known that sydn-1 was a regulator of pfs-2, a member of the 3’end processing machinery, we examined if the loss of *sydn-1* results in global defects in 3’end processing of transcripts. Loss of *sydn-1* did not result in a change in utilization of PAS (Fig 4.2A). The relative percentages of canonical AATAAA and alternative PAS variants in the *sydn-1* mutants were similar to that of N2 and *rpm-1*. The 3’UTRs in the *sydn-1* mutant also exhibit similar percentages of individual PAS elements as in N2 and *rpm-1* (Fig 4.2B). Since the relative distribution of canonical and variant PAS sites did not show any difference, we asked if there was any abnormality in the positional distribution of the PAS sites. However, the positional distribution for all the samples peak at 19nt upstream of the cleavage site (Fig 4.2C). We then annotated the 3’ends to the gene models in WS190 along with our updated 3’UTRome annotations. We next asked if there were any differences in the functional regions of the genome where the isoforms map. Although the abundance of libraries varied between the mutants and N2 (Fig 4.2D, top panel), the relative distribution of the mapping classes was very similar (Fig 4.2D, bottom panel). There was a slight increase in the percentage of unannotated 3’UTRs that did not have any previous annotations. However, analyzing this without better gene annotations is not currently possible.

4.3.2: Isoform differences between the 3’UTR libraries.

Would disruption of 3’end processing possibly affect the isoform expression? First we examined isoform frequency distribution in each library. We see that the majority of the genes had 1, 2 or 3 isoforms (Fig 4.3A). Hence we restricted further analysis to these three classes. Since the mutant libraries are defective in synapse and axon biogenesis, we asked if there was a major difference in the number of neuronal 3’UTRs expressed between the N2 and the mutants. Although there was an abundance difference between N2 and the mutants, the number of neuronal genes expressed was very similar (Fig 4.3B). The *rpm-1;sydn-1* double mutant shows a decrease in the number of neuronal genes but this could also be due to the low coverage in that library. All four libraries expressed isoforms that were unique to that library and that came from both already annotated isoforms as well new isoforms unique to the dataset (Fig 4.3C). Performing the pairwise comparison between the N2 vs mutants and between the different mutants we can see a significant uniqueness in isoform expression between the N2 and mutants (Fig 4.3D). The next logical question was to separate them based on the number of isofoms per gene and look for differences in each class.

4.3.3: Comparison of single isoform genes

Since global analysis of the isoforms didn’t show major differences in PAS utilization and PAS position or neuronal gene expression between the wild type and mutants, we looked at single isoform genes. These are genes that only express one isoform across all libraries. Based on the expression in individual libraries, they can be further classified into three subclasses: present in N2 and present in the mutant, present in N2 and absent in the mutant, absent in N2 and present in the mutant. We see that the distribution of the three subclasses is similar for the global and neuronal genes (Fig 4.4A), indicating that the global and neuronal trends are similar. Interestingly, when we look at the PAS distribution of the individual classes significant differences arise, the dominant class belongs to the isoforms common to N2 and mutants (1,927 out of 2,951 (65%) single isoforms in *rpm-1*, 1,833 out of 3,138 (58%) single isoforms in *sydn-1* and 1,423 out of 1,881 (76%) single isoforms in *rpm-1;sydn-1* are common between N2 and the mutants). This class shows enrichment for the AATAAA motif (1,253 out of 2,341 (54%) isoforms). However, the isoforms that are expressed uniquely in the N2 or mutants seem to be enriched in the alternative PAS and noPAS (1,327 out of 2,328 (57%) isoforms unique to mutants and 451 out of 721 (63%) isoforms unique to N2) (Fig 4.4B). This enrichment for alternative PAS usage in the single isoforms is a variation from our previous work [12] in N2 where we showed that single isoforms were enriched in canonical AATAAA. The common class of 3'UTRs are more dominant (~58-76%) and would have masked the effect of the alternative PAS usage in the unique 3'UTRs and this explains why we didn't notice this subtle variation in the global trend. Similar results were noticed for the subset of neuronal genes alone (821 single isoform genes) (Fig 4.4B right panel). Here again the common 3'UTRs expressed were enriched with the AATAAA motif (196 out of 353 (56%) isoforms) while the unique 3'UTRs expressed enriched alternative and no PAS motifs (216 out of 347(62%) isoforms unique to the mutants and 85 out of 121 (70%) isoforms unique to N2).

4.3.4: Comparison of two isoform genes.

We next examined genes expressing two isoforms to determine if the enrichment in alternative PAS usage in the mutants for single-isoform genes was also present in genes expressing two 3'UTR isoforms (1,221 genes). First we divided the 3'UTR isoforms into three subclasses, common to N2 and mutant, unique to mutant and unique to N2. Then we examined the PAS distribution across each class for enrichment in alternative PAS usage. However this did not show any enrichment for all genes or for the subset of neuronal genes (Fig 4.5A). Next, we separated the 3'UTRs into short and long isoforms for each gene and examined the PAS usage. In the isoforms common to N2 and the mutants, the longest isoform had increased AATAAA motif usage (269 out of 564 (47%) isoforms) compared to the unique isoforms (153 out 494 (31%) 3'UTRs unique to mutants and 53 out 163 (33%) 3'UTRs unique to N2) (Fig 4.5B left). The subset of neuronal expressed genes also exhibited similar trend (Fig 4.5B right). Next, we looked for correlation between the length and abundance of the isoforms in N2 and the mutants. Out of 1,015 isoforms expressed in N2 for two isoform genes, 497 (49%) had abundant short isoforms. Out of 1,005 isoforms expressed in *rpm-1* for two isoform genes 474 (47%) had abundant short isoforms. Out of 991 isoforms expressed in *sydn-1* for two isoform genes, 511 (52%) had abundant short isoforms. Out of 775 isoforms expressed in *rpm-1;sydn-1* for two isoform genes, 423 (55%) had abundant short isoforms. Similar abundances between long and the short isoforms for the N2 and the mutants suggest no bias on the length of 3'UTRs affecting abundance on a global scale. One possible explanation could be that the effect of length on abundance is on a per-gene basis or it could be at the protein level and not at the transcript level where we measure. However, we see a slight increase in the percentage for *sydn-1* and *rpm-1; sydn-1* and slight decrease in rpm-1 in the abundance of short isoforms.

4.3.5: Alternate isoform utilization in two isoform genes

In our previous study of developmental stage specific 3'UTRs, we identified cases where different 3'UTR lengths were utilized across different developmental stages for the same gene. Furthermore, the utilization of the 3'UTRs changed from short to long or long to short during developmental transitions. We called these cases "isoform switching". With our current data since all the worms were staged for L1 developmental stage we can't currently look for developmental stage switching. However, we could examine genes with two isoforms to identify cases of alternative 3'UTR utilization between N2 and mutants and between different mutants. We searched for cases where the most abundant isoform switched from the long to short or vice versa between the mutants and N2 or between mutants. Out of the 1221 genes with two isoforms, our search resulted in 317 (26%) genes which exhibit alternative 3'UTR utilization between N2 and one of the mutants and 277 (23%) genes which exhibit alternative 3'UTR utilization across the mutants (Fig 4.6A). Of these, 62 out of the 317 (20%) genes and 59 out of the 277 (21%) genes have neuronal expression. 176 genes overlap between these two switching lists and out of those 39 are neuronal expressed. The complete list of switching genes is provided at the end of this chapter. Two examples have been given to highlight this phenomenon (Fig 4.6B). In the first example we see that while the N2, *rpm-1* and the double mutant use the longest isoform, the *sydn-1* mutant preferentially expresses the shorter isoform. In the second example, we see that while N2 and *rpm-1* express both isoforms, *sydn-1* and the double mutants express only either the longest or the shortest isoform. The fact that >20% of the genes expressed with two isoforms express alternate isoform utilization between N2 and the mutants suggests that disruption of the synaptogenesis pathway has effects on the polyA site selection. Of these genes, only 20% were neuronally expressed. This suggests that these non-neuronal expressed genes could also play a role in both synaptogenesis and polyA site selection. Careful annotation of these genes in future work would facilitate further understanding.

4.3.6: Comparison of three isoform genes.

We extended our previous analysis to genes with three isoforms, exhibiting similar results. The overall PAS distribution shows no global variation, similar to the two isoform genes. Hence we have to look at the length distribution of these isoforms (Fig

4.7A). When we look at individual cases, again we see that the longest AATAAA is enriched only in the common isoforms while all the differential isoforms express alternate PAS with increased “no PAS” isoforms in the mutant alone isoforms, similar to the two isoform cases (Fig 4.7A). Next, we separated the 3’UTRs into short, middle and long isoforms for each gene and examined the PAS usage. In the isoforms common to N2 and the mutants, the longest isoform had increased AATAAA motif usage (59 out of 113 (52%) isoforms) compared to the unique isoforms (29 out 125 (23%) 3’UTRs unique to mutants and 13 out 41 (32%) 3’UTRs unique to N2) (Fig 4.7B top). The subset of neuronal expressed genes also exhibited a similar trend (Fig 4.7B bottom). Next, we looked for correlation between the length and abundance of the isoforms in N2 and the mutants. Out of 522 isoforms expressed in N2 for three isoform genes, 92 (18%) had abundant short isoforms, 89 (17%) had abundant middle isoforms and 74 (14%) had abundant long isoforms. Out of the 483 isoforms expressed in *rpm-1* for three isoform genes, 78 (16%) had abundant short isoforms, 93 (19%) had abundant middle isoforms and 83 (17%) had abundant long isoforms. Out of 461 isoforms expressed in *sydn-1* for three isoform genes, 93 (20%) had abundant short isoforms, 85 (28%) had abundant middle isoforms, 62 (13%) had abundant long isoforms. Out of 341 isoforms expressed in *rpm-1;sydn-1* for three isoform genes, 81 (24%) had abundant short isoforms, 78 (23%) had abundant middle isoforms and 44 (13%) had abundant long isoforms. Similar abundances between long, middle and the short isoforms for the N2 and the mutants suggest no bias on the length of 3’UTRs affecting abundance on a global scale.

4.3.7: Alternate isoform utilization in three isoform genes

Similar to our analysis on two isoform genes, we continued our search for alternative isoform utilization on three isoform genes. We looked for cases where the most abundant isoform switches between the long to short/middle, middle to short/long and short to long/middle or vice versa between the mutants and N2 or between mutants. Out of the 279 genes with three isoforms, 167 (60%) genes exhibited alternative 3'UTR utilization between N2 and one of the mutants and 140 (50%) genes which exhibit alternative 3'UTR utilization across the mutants (Fig 4.9A). Of these, 37 out of the 167 (22%) genes and 28 out of the 140(22%) genes have neuronal expression. 103 genes overlap between these two switching lists and out of those 22 are neuronal expressed. The complete list of the switching genes is given at the end of this chapter. Two examples have been provided to demonstrate the switching (Fig 4.9B). The examples are similar to the ones explained in the two isoform cases. Here we have an additional third isoform given in green. In the first example we see differential isoform usage between N2 and mutants and also between mutants. *sydn-1* is expressing the longest isoform while N2 is expressing the shortest isoform and *rpm-1* is expressing the middle length and the double mutant doesn't express the gene. Compared to two isoform genes, we see an increased percent (>50%) of the genes with three isoforms that express alternate isoform utilization. Of these genes, only 20% were neuronal expressed, similar to two isoform genes. The fact that both two and three isoform genes show evidence of alternate 3'UTR utilization between N2 and the mutants further strengthens the claim relating synaptogenesis and 3'end processing.

4.4: Conclusion

In this chapter we show successful application of the polyA capture protocol to study effects of genes that affect neuronal development and 3'UTR formation in *C. elegans*. We have identified 8,589 known 3'UTRs and also 634 new 3'UTRs have been sequenced. This shows that the 3'UTRome is still far from saturation and new 3'UTRs will likely be found based on the tissue, development timing or mutant background of input sample. While global trends do not show any stark irregularities between N2 and the mutants in the utilization of the polyA machinery, we do see many subtle variations. Especially we see many isoforms that are expressed uniquely in N2 or the mutants. These 3'UTRs passed the false priming filters derived in the previous chapter and hence chance of them being false priming artifacts are low. We also see an increased alternative PAS usage in the isoforms that are uniquely expressed in the mutants or N2. Furthermore, we report 484 genes that exhibit alternative isoform utilization between N2 and the mutants or between the mutants. Of these, 20% of the genes are neuronal expressed. These results indicate a link between synaptogenesis and 3'end processing machinery. 80% of the alternative 3'UTR utilizing genes are non-neuronal. This could indicate the role of other genes in neuronal development. Hence this work could be a starting point for future gene-wise analysis.

4.5: Reference

1. Lau AG, Irier HA, Gu J, Tian D, Ku L, Liu G, Xia M, Fritsch B, Zheng JQ, Dingledine R *et al*: **Distinct 3'UTRs differentially regulate activity-dependent translation of brain-derived neurotrophic factor (BDNF)**. *Proceedings of the National Academy of Sciences* 2010, **107**(36):15945-15950.
2. Irier HA, Shaw R, Lau A, Feng Y, Dingledine R: **Translational regulation of GluR2 mRNAs in rat hippocampus by alternative 3′ untranslated regions**. *Journal of Neurochemistry* 2009, **109**(2):584-594.
3. Didiano D, Cochella L, Tursun B, Hobert O: **Neuron-type specific regulation of a 3'UTR through redundant and combinatorially acting cis-regulatory elements**. *Rna* 2010, **16**(2):349-363.
4. Zhen M, Huang X, Bamber B, Jin Y: **Regulation of presynaptic terminal organization by C. elegans RPM-1, a putative guanine nucleotide exchanger with a RING-H2 finger domain**. *Neuron* 2000, **26**(2):331-343.
5. Hallam SJ, Goncharov A, McEwen J, Baran R, Jin Y: **SYD-1, a presynaptic protein with PDZ, C2 and rhoGAP-like domains, specifies axon identity in C. elegans**. *Nat Neurosci* 2002, **5**(11):1137-1146.
6. Zhen M, Jin Y: **The liprin protein SYD-2 regulates the differentiation of presynaptic termini in C. elegans**. *Nature* 1999, **401**(6751):371-375.
7. Nakata K, Abrams B, Grill B, Goncharov A, Huang X, Chisholm AD, Jin Y: **Regulation of a DLK-1 and p38 MAP kinase pathway by the ubiquitin ligase RPM-1 is required for presynaptic development**. *Cell* 2005, **120**(3):407-420.
8. Liao EH, Hung W, Abrams B, Zhen M: **An SCF-like ubiquitin ligase complex that controls presynaptic differentiation**. *Nature* 2004, **430**(6997):345-350.
9. Van Epps H, Dai Y, Qi Y, Goncharov A, Jin Y: **Nuclear pre-mRNA 3'-end processing regulates synapse and axon development in C. elegans**. *Development* 2010, **137**(13):2237-2250.
10. Ohnacker M, Barabino SM, Preker PJ, Keller W: **The WD-repeat protein pfs2p bridges two essential factors within the yeast pre-mRNA 3'-end-processing complex**. *EMBO J* 2000, **19**(1):37-47.
11. Wang SW, Asakawa K, Win TZ, Toda T, Norbury CJ: **Inactivation of the pre-mRNA cleavage and polyadenylation factor Pfs2 in fission yeast causes lethal cell cycle defects**. *Mol Cell Biol* 2005, **25**(6):2288-2296.
12. Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V *et al*: **The Landscape of C. elegans 3'UTRs**. *Science* 2010, **329**(5990):432-435.
13. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656-664.

4.6: Figures

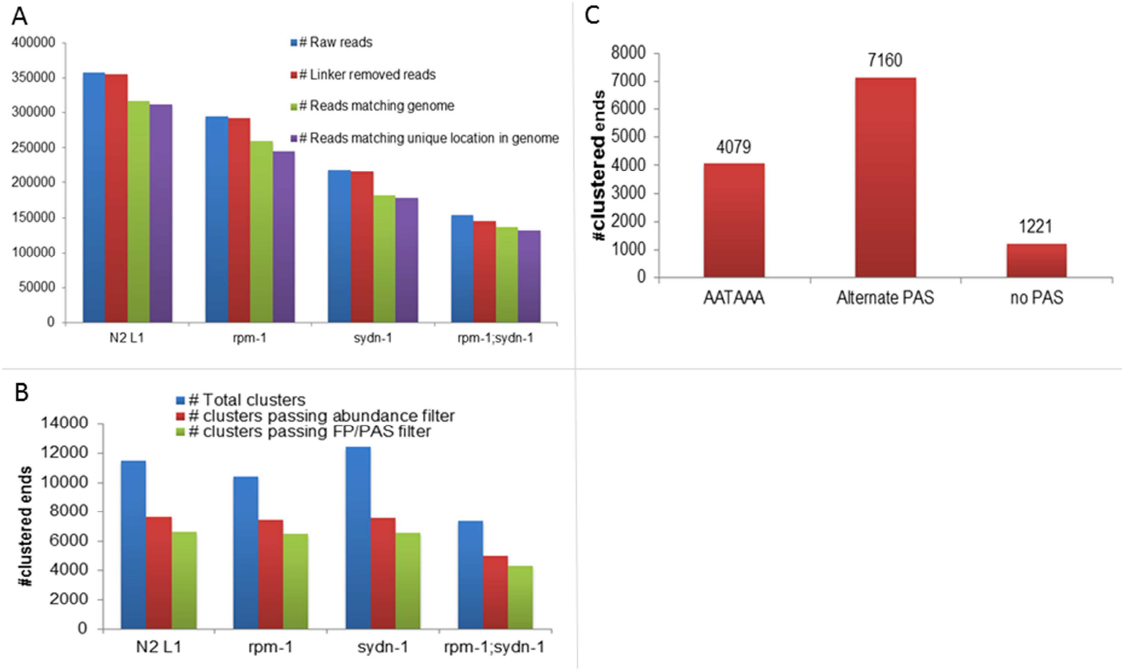
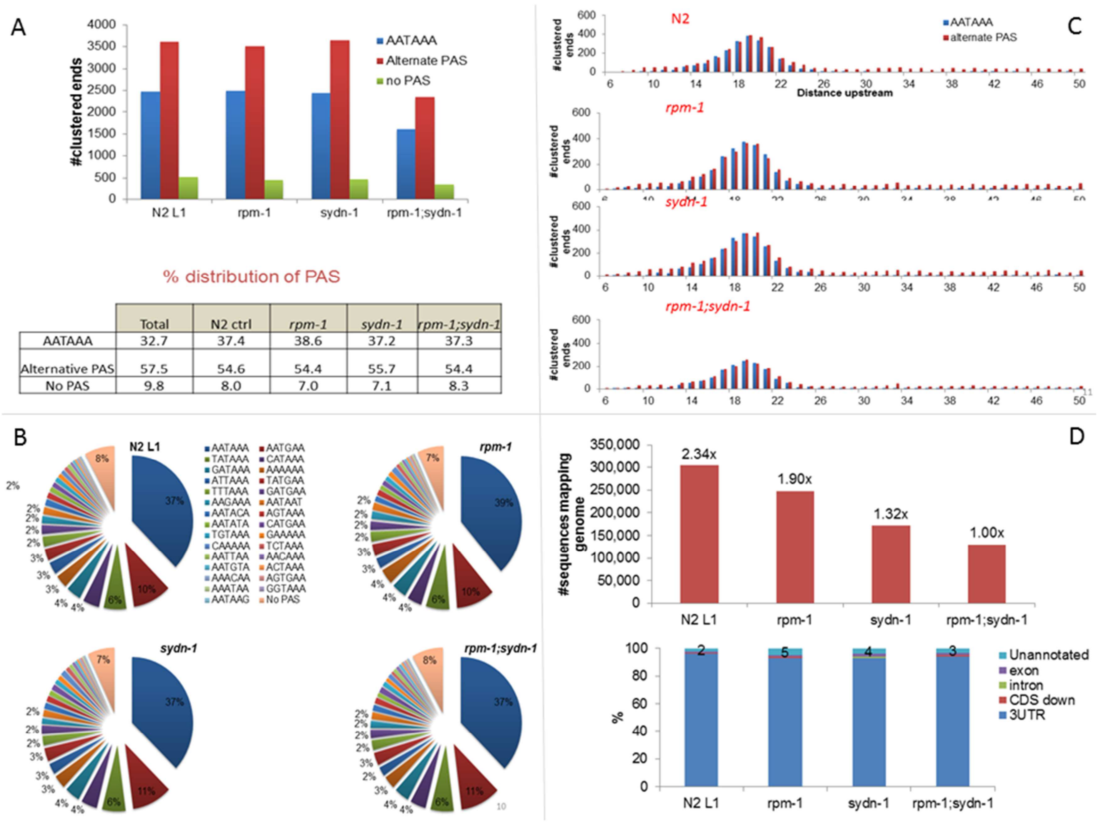


Figure 4.1: Preprocessing of the 3'UTR libraries

A: Sequence statistics for N2, *rpm-1, sydn-1* and *rpm-1;sydn-1* double mutant libraries. The total number of raw reads, the number of linker removed reads, number of reads mapping to the genome and number of reads mapping unique locations in the genome are indicated.

B: Number of 3'UTR isoform clusters for the N2, *rpm-1, sydn-1* and *rpm-1;sydn-1* double mutant libraries. Total number of clusters, number of clusters passing abundance filter, and the number of clusters passing the false priming and PAS site filter are indicated.

C: Global PAS distribution of 3'UTRs across all libraries.



	Total	N2 ctrl	*rpm-1*	*sydn-1*	*rpm-1;sydn-1*
AATAAA	32.7	37.4	38.6	37.2	37.3
Alternative PAS	57.5	54.6	54.4	55.7	54.4
No PAS	9.8	8.0	7.0	7.1	8.3

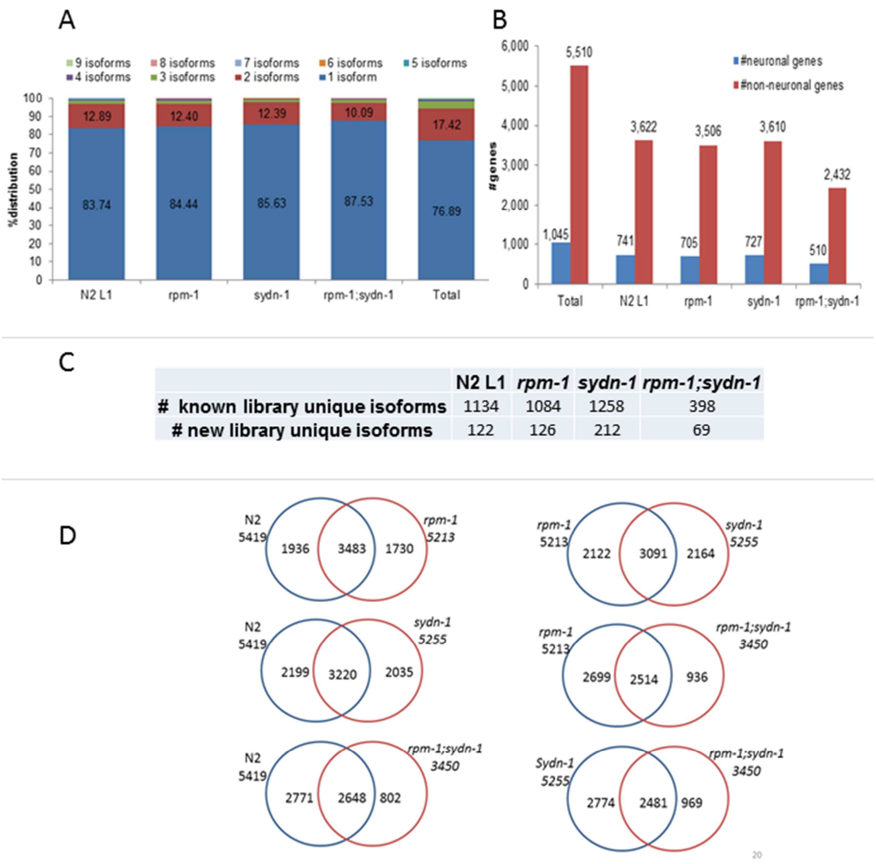
Figure 4.2: PAS site and position distribution

A: PAS distribution per library (top panel). Percentage PAS distribution of each library (bottom panel).

B: Percentage distribution of each individual PAS for each library.

C: Positional distribution of PAS site for each library.

D: Total and relative abundance of each library (top panel). Percentage distribution of genomic regions mapped by polyA end clusters (bottom panel).



	N2 L1	*rpm-1*	*sydn-1*	*rpm-1;sydn-1*
# known library unique isoforms	1134	1084	1258	398
# new library unique isoforms	122	126	212	69

Figure 4.3: comparison of 3'UTRs between individual libraries.
A: Percentage isoform frequency for each library.

B: Number of neuronal and non-neuronal genes expressed in each library.

C: Number of known and new unique isoforms expressed in each library.

D: The Venn diagrams represent the isoform overlap between the different libraries.

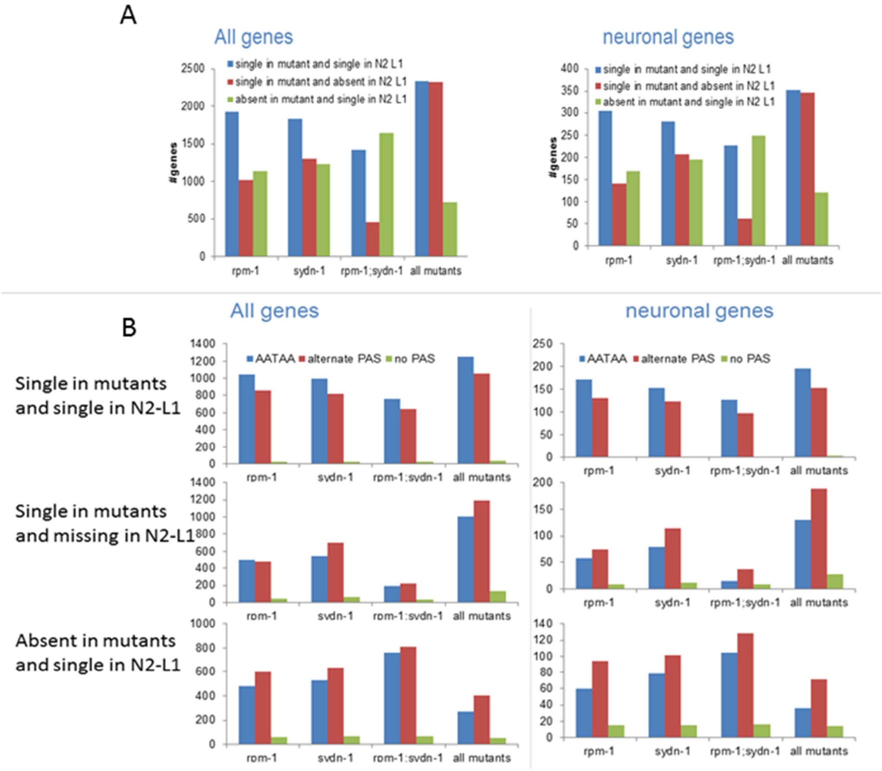


Figure 4.4: 3'UTR isoform analysis in single isoform genes.

A: Number of single expressed isoforms overlapping or unique between N2 and the mutant libraries for all genes (left) and neuronal expressed genes (right)

B: PAS distribution for single expressed isoforms overlapping or unique between N2 and mutant libraries for all genes (left) and neuronal expressed genes (right)

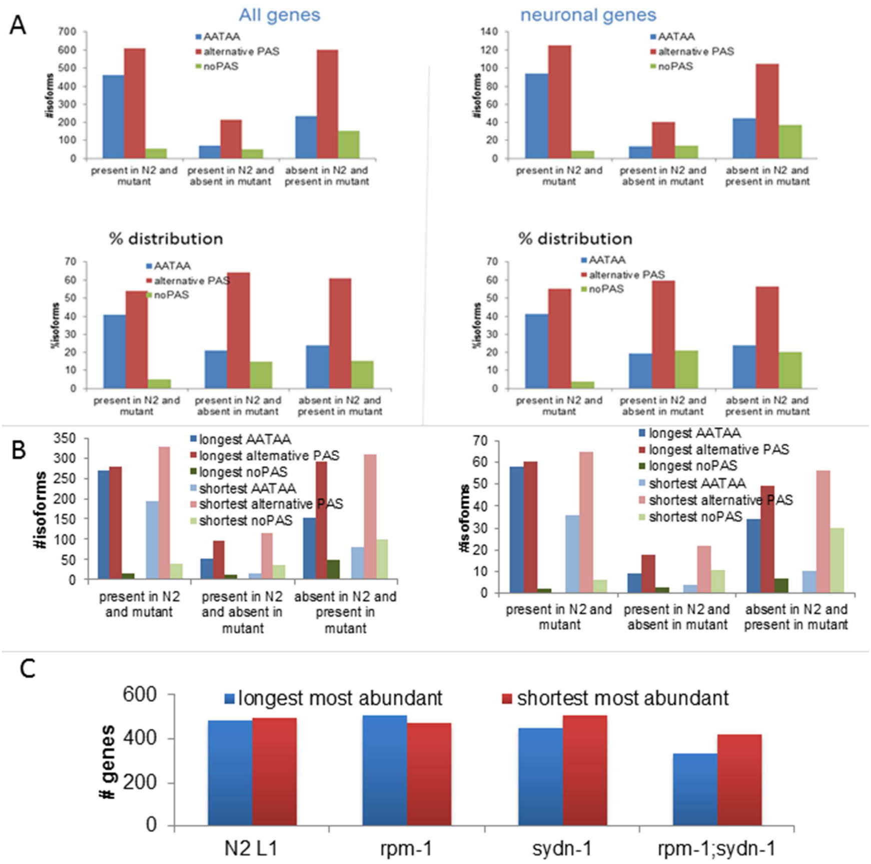


Figure 4.5: 3'UTR isoform analysis in two isoform genes

A: PAS distribution of 3'UTRs expressed in two isoform genes overlapping or unique between N2 and mutants for all genes (left) and neuronal genes (right). Bottom panel provides percentage PAS distribution of 3'UTRs expressed in two isoform genes overlapping or unique between N2 and mutants for all genes (left) and neuronal genes (right).

B: PAS distribution for longest and shortest 3'UTR expressed in two isoform genes overlapping or unique between N2 and mutants for all genes (left) and neuronal genes(right). The box highlights the significant variations.

C: Number of genes with two isoforms where the longest or the shortest isoform is the most abundant in the library.

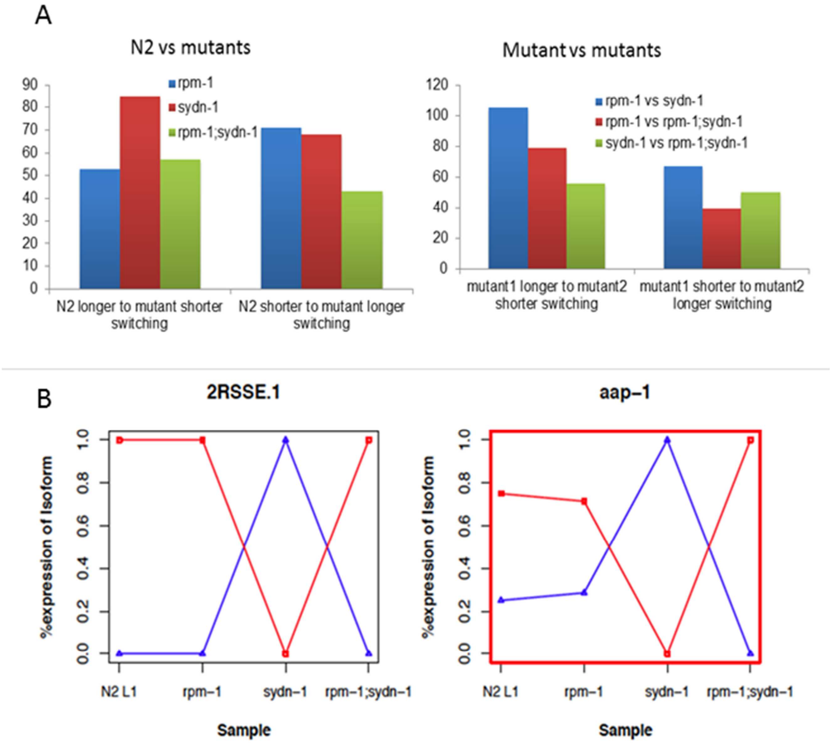


Figure 4.6: Alternate 3’UTR isoform expression in two isoform genes

A: Number of two 3’UTR isoform genes that exhibit alternate 3’UTR expression between N2 and mutants (left) and between mutants (right)

B: Examples of two 3’UTR isoform genes exhibiting alternate isoform expression. The longest isoform is given by red, the shortest by blue. Red outer boxes indicate neuronal expressed gene

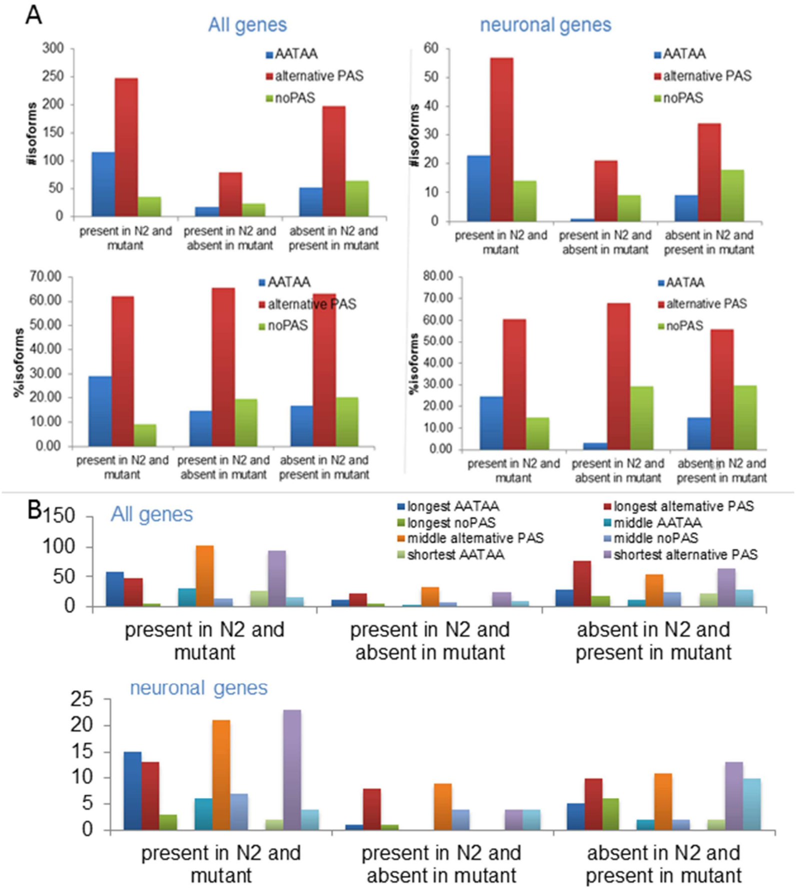


Figure 4.7: 3’UTR isoform analysis in three isoform genes

A: PAS distribution of 3’UTRs expressed in three isoform genes overlapping or unique between N2 and mutants for all genes (left) and neuronal genes (right). Bottom panel provides percentage PAS distribution of 3’UTRs expressed in three isoform genes overlapping or unique between N2 and mutants for all genes (left) and neuronal genes (right).

B: PAS distribution for longest, middle and shortest 3’UTR expressed in three isoform genes overlapping or unique between N2 and mutants for all genes (left) and neuronal genes(right). The arrows highlight the significant variations.

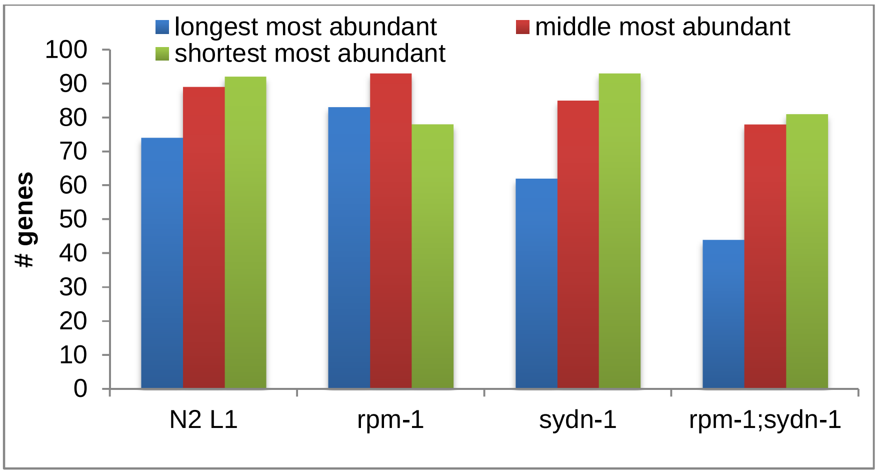


Figure 4.8: length dependent abundance distribution in 3 isoform genes
Number of genes with three isoforms where the longest, shortest or middle isoform is the most abundant in the library.

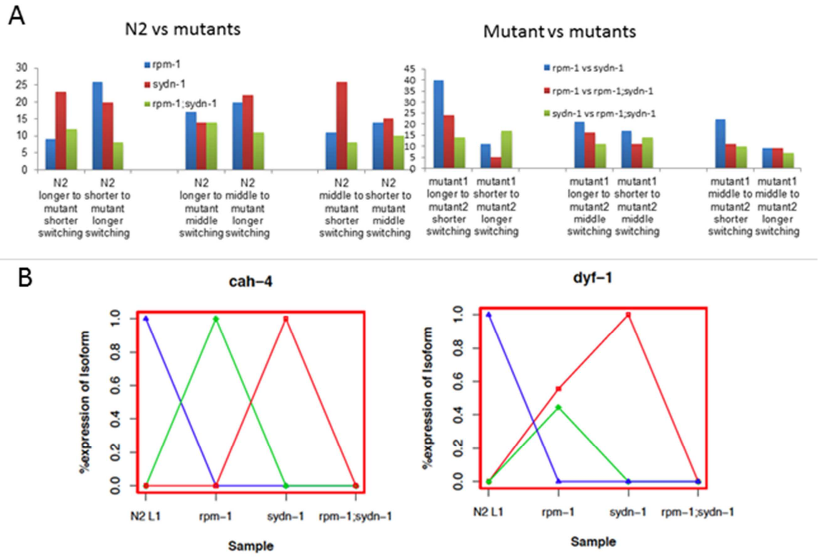


Figure 4.9: Alternate 3’UTR isoform expression in three isoform genes

A: Number of three 3’UTR isoform genes that exhibit alternate 3’UTR expression between N2 and mutants (left) and between mutants (right)

B: Examples of three 3’UTR isoform genes exhibiting alternate isoform expression. The longest isoform is given by red, the shortest by blue and middle by green. Red outer boxes indicate neuronal expressed gene

Chapter 5: 26G endo-siRNAs regulate spermatogenic and zygotic gene expression in *C. elegans*

5.1: Contribution

The primary goals of this project were to identify the small RNAs that play a role in the germline of an organism, especially in the gametes (sperm and oocytes), and how these small RNAs regulate transcription and translation in these germ cells. This work involved contributions from various researchers and necessitates proper acknowledgement. Ting Han and John Kim at the University of Michigan conceived the project. Colin Fitzpatrick and Diana Chu from SFSU provided germ cell samples. Ting Han performed the generation of the samples and small RNA libraries for sequencing. The libraries were pyrosequenced by Tim Harkins and Pascal Bouffard of Roche 454 life sciences. The Illumina sequencing of the libraries was performed at the British Columbia Genome Sequencing Center. Ting Han and John Kim designed the experiments and Ting Han performed all the experiments discussed in the paper. Jean and Danielle Thierry-Mieg at NIH performed parallel bioinformatics analysis of the data. Their conclusions provided independent validation of my analysis. Their analysis and curation from AceView database developed by them was instrumental in determining the nature of the 26G RNA targets. Ting Han and John Kim wrote the manuscript submitted to *PNAS* which is provided in the sections below.

My contribution to the project was the computational analysis of the high throughput sequencing libraries including preprocessing of the sequences to remove the linkers, mapping to the WS190 genome, and functional classification of the mapped sequences into known and new class of small RNAs. I also computationally identified and characterized the 26GRNAs based on length and first nucleotide distribution, genes targeted by these small RNAs and their predominant antisense nature. I also classified the two classes of the 26GRNAs targeting spermatogenesis and oogenesis specific genes and its depeletion in the *glp-4* and *eri-1* mutant libraries indicating germline expression and endo siRNA pathway dependency.

5.2: Abstract

Endogenous small interfering RNAs (endo-siRNAs) regulate diverse gene expression programs in eukaryotes by either binding and cleaving mRNA targets or mediating heterochromatin formation; however, the mechanisms of endo-siRNA biogenesis, sorting, and target regulation remain poorly understood. Here we report the identification and function of a specific class of germline-generated endo-siRNAs in *C. elegans* that are 26nt in length and contain a guanine at the first nucleotide position (i.e. 26G RNAs). 26G RNAs regulate gene expression during spermatogenesis and zygotic development, and their biogenesis requires the ERI-1 exonuclease and the RRF-3 RNA-dependent RNA polymerase. Remarkably, we identified two non-overlapping subclasses of 26G RNAs that sort into specific RNA-induced silencing complexes (RISCs) and differentially regulate distinct mRNA targets. Class I 26G RNAs target genes expressed during spermatogenesis, whereas Class II 26G RNAs are maternally inherited and silence gene expression during zygotic development. These findings

implicate a novel class of endo-siRNAs in the global regulation of transcriptional programs required for fertility and development.

5.3: Introduction

In eukaryotes, small RNAs (20-30nt) regulate gene expression and genome organization via nucleic acid sequence homology [1, 2]. Usually processed from double stranded RNA precursors by the RNase III-like enzyme Dicer, small RNAs are incorporated into a ribonucleoprotein complex called the RNA-induced silencing complex (RISC) that contains a core protein belonging to the Argonaute/Piwi protein family [3-9]. Through base pairing, small RNAs guide RISC to recognize cognate targets and elicit silencing activities.

Small RNAs are classified by their means of biogenesis, Argonaute/Piwi associations, and biological functions. MicroRNAs are processed from hairpin-bearing precursors by Drosha and Dicer, bind Argonaute proteins, and mediate translational repression or degradation of mRNAs [10]. Piwi-interacting RNAs (piRNAs), in contrast, are generated by a partially identified Dicer-independent self-amplification pathway, associate with Piwi proteins, and protect genome integrity by silencing transposons [11]. A third emerging class is endogenous small interfering RNA (endo-siRNAs), which fine-tune host gene expression [12].

Endo-siRNAs were first described and characterized in *C. elegans*. By cDNA cloning, Ambros *et al.* identified over 700 small (~20nt) antisense RNAs, which are perfectly complementary to protein-coding genes [13]. The biogenesis of these endo-siRNAs requires the *C. elegans* Dicer, *dcr-1,* the RNA-dependent RNA polymerase *rrf-3,* and the exonuclease *eri-1* [14, 15]. Mutants defective in endo-siRNAs exhibit elevated target mRNA levels, suggesting that endo-siRNAs repress host gene expression [14-16]. Ruby *et al.* employed a large-scale sequencing approach and identified thousands of endo-siRNAs that preferentially target transcripts associated with spermatogenesis and transposons [17]. Based on these studies, a complex picture of *C. elegans* endo-siRNAs is emerging, with different functions and genetic requirements. Recently, several groups reported the discovery of endo-siRNAs in *D. melanogaster* and *M. musculus* [18-23]. These endo-siRNAs are derived from transposable elements, natural antisense transcripts, and hairpin RNAs. Their biogenesis requires Dicer and Ago2 and their depletion results in target up-regulation, further supporting the notion that endo-siRNAs negatively regulate endogenous gene expression.

Proper maintenance of the germline and generation of healthy gametes are crucial for sexual reproduction. Many mechanisms have evolved to ensure germline stability and reproductive success [24-26]. Recent studies show that mutations affecting small RNA pathways frequently are associated with defective gametogenesis [7, 27]. For example, in *C. elegans*, the *dcr-1* null mutant is defective in microRNA and siRNA (small interfering RNA) biogenesis, displays impaired fertility, and accumulates malformed unfertilized oocytes [4, 6, 7]. Similarly, mutation of *prg-1* (*p*iwi *r*elated *g*ene) abrogates the expression of 21U RNAs (a piwi-interacting class of small RNAs) and results in severely impaired germline proliferation and sterility at elevated temperatures [28-30]. Small RNAs also can serve as heritable parental silencing factors to regulate filial gene expression. In *D. melanogaster*, misregulation of maternally inherited piRNAs results in activation of transposons and hybrid dysgenesis [31]. These observations underscore the essential functions of small RNAs in germline development and cross-generational epigenetic regulation.

To decipher the role of small RNAs in the germline of *C. elegans*, we employed high-throughput deep sequencing to characterize small RNAs expressed in purified male sperm, hermaphrodite oocytes, and embryos. We identified two subclasses of germline-generated endo-siRNAs (sperm 26G RNAs and oocyte/embryo 26G RNAs) that regulate gene expression during spermatogenesis and zygotic development. Genetic analyses revealed that the ERI-1 exonuclease and the RRF-3 RNA-dependent RNA polymerase are required for 26G RNA biogenesis. Interestingly, the two subclasses of 26G RNAs require different Argonautes for their expression, suggesting differential RISC loading and mRNA targeting. Recent evidence indicates that piRNAs are maternally inherited to silence transposons in the subsequent generation [31]. Our findings indicate that the 26G RNAs not only exert a profound influence over male gametogenesis, but are also maternally inherited to act as epigenetic agents to control gene expression during zygotic development in the progeny.

5.4: Results

Deep sequencing revealed germline-enriched, *eri-1*-dependent 26G endo-siRNAs

Small RNAs expressed in purified male sperm, hermaphrodite oocytes, and embryos of *C. elegans* were size selected (18-32nt), ligated to adaptors, and sequenced by high-throughput deep sequencing (Roche/454 and Illumina/Solexa). After excluding sequences corresponding to microRNAs, 21U RNAs, and putative degradation products derived from abundant noncoding RNAs (e.g. rRNAs) (Fig. 5.S1; supplemental methods), we identified 2.45 million putative endo-siRNA reads (14.8% of the total sequences). These endo-siRNAs display a bimodal length distribution with one peak clustered at ~21nt and the second at 26nt (Fig. 5.1A). Notably, while ~21nt endo-siRNAs do not have a strong first nucleotide bias, the 26nt endo-siRNAs preferentially start with a guanine nucleotide (Fig. 5.1B). Therefore, we refer to them as 26G RNAs (Table S1).

Although 26G RNAs previously have been identified by high-throughput sequencing of small RNAs isolated from mixed-stage worms, little is known about their biogenesis or their potential role in gene regulation [17]. Mapping to the genome reveals that 26G RNAs are largely derived from protein coding genes (i.e. exons, introns, and UTRs) (77%) and exhibit a strong antisense bias (73% of the total mapped to antisense vs. 4% of the total mapped sense of known and predicted genes) (Fig. 5.1C; also see supplemental computational methods). In addition, the majority of 26G RNAs derived from exons or introns of coding transcripts target exons (97.2%) or span exon-exon junctions (0.7%), suggesting that mature mRNAs are the main targets of 26G RNAs. (Fig. 5.S3).

We next used deep sequencing to compare the endo-siRNA profiles of N2, *glp-4(bn2)*, and *eri-1(mg366)* whole animals. The *glp-4(bn2)* mutant fails to proliferate its germline at non-permissive temperature (25°C) and therefore lacks germline-derived small RNAs; consequently, *glp-4* mutants exhibit a decline in the expression of ~21nt small RNA population (Fig. 5.1D). The *eri-1(mg366)* mutant exhibits impaired biogenesis of several known endo-siRNAs and produces defective sperm at 25°C [14, 15, 32]. Consistent with these findings, we identified a small fraction of the ~21nt endo-siRNAs that appear to be *eri-1*-dependent. These small RNAs largely overlap with 26G RNAs (starting with the same 5' G), but their depletion in *eri-1* is not as severe as that seen for 26G RNAs (Fig. 5.1F). Overall, the expression of 21nt endo-siRNAs remains relatively unchanged in N2 versus *eri-1(mg366)*, suggesting that ~21nt endo-siRNAs constitute a genetically diverse population of small RNAs (Fig. 5.1D). In contrast, the 26G RNAs are profoundly depleted in *glp-4(bn2)* and *eri-1(mg366)* animals (Fig. 5.1D-

F). Thus, we conclude that 26G RNAs represent a novel class of germline-enriched endo-siRNAs that depend on *eri-1* for their expression.

Two subclasses of 26G RNAs with different expression patterns

Strikingly, hierarchical clustering reveals that 98.9% of the 26G RNAs fall into two distinct classes (Fig. 5.2A; Fig. 5.S1; supplemental methods). Class I 26G RNAs are present in purified sperm (1,102 unique sequences; 5,960 total reads), but are not present in oocytes or embryos. By comparison, class II 26G RNAs are highly enriched in oocytes and embryos (2,441 unique sequences; 148,594 total reads), but are absent in sperm. Both classes of 26G RNAs are present at lower levels in mixed-stage N2 and are severely depleted in *glp-4(bn2)* and *eri-1(mg366)* animals. We analyzed the expression profiles of 5 relatively abundant sperm 26G RNAs (26G-S1, -S3, -S4, -S5, -S6) and 4 oocyte/embryo 26G RNAs (26G-16, -O1, -O2, -O3) by northern blotting and/or RT-qPCR assays (Taqman, Applied Biosystems). By northern blotting, the expression of 26G RNAs shows *eri-1* dependence in purified oocytes and embryos, as well as in male animals (Fig. 5.2B). In addition, clear temporal separation in the expression of these two classes of 26G RNAs was observed (Fig. 5.2C-D). The class I sperm 26G RNAs (denoted 26G-S) (Fig. 5.2C-D) are only detectable in late larval (L4) and young adult stages in N2 hermaphrodites and males (Fig. 5.2C); furthermore, a finer time course revealed class I sperm 26G RNA expression occurs in a relatively narrow window, consistent with their expression during *C. elegans* spermatogenesis (Fig. 5.2D). Conversely, expression of class II oocyte/embryo 26G RNAs (denoted 26G-O) (Fig. 5.2C-D) initiates during oogenesis, peaks in embryos, and progressively declines throughout the four larval stages. Interestingly, northern blotting revealed several cases of cross-hybridization of the 26G RNA probes to a less abundant ~21nt species (Fig. 5.2B-C). While the 26G RNA signal is completely abolished in the *eri-1* mutant background, the ~21nt signals were either not altered (e.g. 26G-S1, Fig. 5.2B) or depleted to a lesser extent (e.g. 26G-O2 in Fig. 5.2B), suggesting that 26G RNAs and 21nt endo-siRNAs may have distinct genetic requirements for biogenesis, even though both classes of endo-siRNAs may target similar sequences.

Two subclasses of 26G RNAs silence distinct sets of targets

26G RNAs are perfectly complementary to their predicted gene targets, suggesting that they may act as canonical siRNAs to direct the cleavage of their mRNA targets. Importantly, 26G RNAs target a different set of genes from those targeted by shorter length (20-24nt) endo-siRNAs (Fig. 5.S4). Because the expression patterns of the two classes of 26G RNAs are mutually exclusive, we next asked if they differentially regulate non-overlapping, discrete classes of target genes. Indeed, based on existing germline gene expression profiles [33], we found that predicted targets of class I sperm 26G RNAs are enriched 7-fold for genes expressed during spermatogenesis, whereas targets of class II oocyte/embryo 26G RNAs are depleted of all three classes of germline genes (spermatogenesis, oogenesis, and germline-intrinsic) (Fig. 5.3B). Because mutations in *eri-1* abolish the expression of both classes of 26G RNAs, we used RT-qPCR to analyze the relative expression of putative 26G RNA targets in *eri-1(mg366)* and N2 at the following five developmental time points: embryos, and 8 hrs (L1), 30 hrs (L3), 42 hrs (L4), and 70 hrs (adult) post hatching (Fig. 5.3A). While transcript levels of genes not targeted by 26G RNAs were similar in *eri-1(mg366)* and N2 animals (Fig. 5.3A, bottom panel), transcripts corresponding to 11 of the 12 genes that are targeted by class I sperm 26G RNAs and all 11 genes targeted by class II oocyte/embryo 26G RNAs are significantly elevated in *eri-1(mg366)* animals relative to N2 controls (Fig. 5.3A; see supplemental method for target selection criteria). Consistent with the temporal expression pattern of class I sperm 26G RNAs, target silencing occurs in a relatively narrow window that corresponds to spermatogenesis through young adulthood (Fig. 5.2A; Fig. 5.S5). By comparison, although class II oocyte/embryo 26G RNA levels steadily decline during larval development, their silencing effect persists throughout development (Fig. 5.3A). Thus, both classes of 26G RNAs appear to silence the expression of their targets, yet with different kinetics: class I sperm 26G RNAs repress targets during spermatogenesis, while class II oocyte/embryo 26G RNAs are maternally deposited to silence gene expression during filial zygotic development.

We next asked if the *eri-1*-dependent regulation of 26G RNA targets could be observed at the whole-transcriptome level. Using previously reported whole-genome microarray data that compared transcript expression profiles of L4 stage *eri-1* and N2 worms *[16]*, we found that predicted targets of 26G RNAs are significantly up-regulated in *eri-1(mg366)* ($p<0.0001$, t-test) (Fig. 5.3C). Conversely, genes up regulated in the *eri-1* mutant background were also significantly enriched for 26G RNA targets (4-fold). Taken together, the highly correlated expression patterns between 26G RNAs and their putative targets at the whole-transcriptome level further support the hypothesis that 26G RNAs directly regulate target gene expression in an *eri-1*-dependent manner.

To determine if target de-repression in *eri-1(mg366)* results in misexpression of target mRNAs in inappropriate tissues, we performed RNA in situ hybridization for select, relatively abundant [33] targets (*C04G2.8* and *ssp-16*) in dissected gonads. The expression of these sperm 26G RNA targets was detected in the spermatogenic gonads in males of both the *him-8* and *eri-1; him-8* strains, but not in the oogenic gonads of N2 or *eri-1* hermaphrodite animals (Fig. 5.3D). Thus, target de-silencing by class I sperm 26G RNAs in the *eri-1* mutant remains restricted to the male gonad, indicating that 26G RNAs repress target expression in their cognate cell types.

Genetic requirements for 26G RNA biogenesis and function

Small RNAs that start with a guanine nucleotide are thought to be products of an RNA-dependent RNA polymerase (RdRP) [34]. Therefore, we asked if RdRPs could play a role in biogenesis of 26G RNAs. The *C. elegans* genome encodes four RdRPs (*rrf-1*, *2*, *3*, and *ego-1*) [35]. We examined 26G RNA expression in mutants for three viable RdRPs, *rrf-1(ok589)*, *rrf-2(ok210)*, and *rrf-3(pk1426*). As mutations in *ego-1* result in lethality [36], we used RNAi to deplete the *ego-1* transcript from N2 animals. While *rrf-1 (ok589)*, *rrf-2 (ok210)*, and *ego-1(RNAi)* express normal levels of 26G RNAs, we found that the expression of both classes of 26G RNAs is abolished in *rrf-3(pk1426*), resulting in significant up-regulation of both classes of targets (Fig. 5.4A; Fig. 5.S6). However, we note that RNAi-inactivation of *ego-1* does not completely abolish *ego-1* expression and therefore we cannot definitively conclude that the 26G RNAs are *ego-1*-independent.

If 26G RNAs are *bona fide* RdRP products, then transcripts they target should serve as templates for 26G RNA production. *deps-1* is a gene whose 3'UTR appears to be targeted by a class I sperm 26G RNA (26G-S4) (Fig. 5.4B). Two alleles of *deps-1* (*bn121* and *bn124*) introduce premature stop codons into the gene and destabilize *deps-1* transcripts (Fig. 5.4B) [37]. In both alleles, the expression of 26G-S4 is significantly depleted, while expression of other 26G RNAs that do not target *deps-1* (26G-S5, -S6) is not affected, supporting the requirement of *deps-1* transcript as a template for 26G-S4 production. We attempted to rescue 26G-S4 expressions by crossing *deps-1* into the *smg-1(r861)* background, which stabilizes transcripts with premature stop codons. We observed a noticeable increase in one (*bn121;r861*) but not the other (*bn124;r861*) of the alleles, likely because *deps-1* mRNA levels are still below WT levels. (Fig. 5.S7). In *C. elegans*, during exogenous RNAi, a similar RdRP-mediated process programmed by *rrf-1* generates secondary siRNAs to amplify the silencing signal [38-40]. These secondary siRNAs start with guanine and are triphosphorylated at the 5' end (5'-PPP). However, although 26G RNAs require the RRF-3 RdRP, they are suitable substrates for T4 RNA ligase-mediated 5' linker ligation (Fig. 5.S8), suggesting that some 26G RNAs may possess a 5' monophosphate group [38, 39].

The non-overlapping identities of the two classes of 26G RNAs and the disparate targets they regulate suggested that they might be sorted into distinct RISCs. Argonaute proteins are central components in the effector phase of RNAi and are defined by the presence of two conserved domains, PAZ and PIWI. Argonautes directly bind small RNAs (via both domains) and may possess target cleavage ("slicer") activity via the PIWI domain [41]. *C. elegans* encodes 27 potential Argonautes with diverse functions, several of which were found to be enriched during spermatogenesis or oogenesis [33, 42]. We found that an Argonaute encoded by *ergo-1 [42]*, whose transcript is enriched during oogenesis [33], is required for the expression of class II oocyte/embryo 26G RNAs, but not for class I sperm 26G RNAs (Fig. 5.4C). Consistent with this finding, only targets of class II oocyte/embryo 26G RNAs were up-regulated in the *ergo-1(tm1860)* mutant (Fig 5.S6). The expression of two Argonautes, T22B3.2 and its close paralog, ZK757.3 (93.1% amino acid sequence identity), are enriched during spermatogenesis [33]. Although the single mutant of either *t22b3.2(tm1155)* or *zk757.3(tm1184)* maintains wild-type expression levels of both classes of 26G RNAs, mutations in both *T22B3.2* and *ZK757.3* abrogate the expression of class I sperm 26G RNAs, but not class II oocyte/embryo 26G RNAs (Fig. 5.4C). Similarly, only targets of class I sperm 26G RNAs are de-repressed in the double mutant (Fig. 5.S6). ERGO-1, T22B3.2, and ZK757.3 all possess the Asp-Asp-His catalytic "slicer" motif [8, 42], suggesting that they are capable of directly mediating endonucleolytic cleavage of their targets. Taken together, our data suggest that distinct RISCs guide the class I and class II 26G RNAs to their cognate targets for silencing.

What are the biological functions of 26G RNA-mediated target regulation? *eri-1* and *rrf-3* mutants, which lack both class I and class II 26G RNAs, are temperature-sensitive (*ts*) sterile due to defective spermatogenesis [32, 43]. While the single Argonaute mutants of *T22B3.2* and *ZK757.3* exhibit near-wild-type levels of fertility, the double mutant, which is specifically defective in the expression of class I sperm 26G RNAs, is completely sterile at 25℃ and can be full y rescued by crossing with WT males (Fig. 5.5 A-C). In contrast, the *ergo-1* Argonaute mutant, which is defective in the expression of class II oocyte/embryo 26G RNAs, displays near wild-type fertility. These findings suggest that class I sperm 26G RNAs play an essential gene regulatory role during spermatogenesis. Loss of class II oocyte/embryo 26G RNAs does not result in any overt developmental phenotypes, as we did not observe any somatic defects in *eri-1*, *rrf-3*, and *ergo-1* mutant animals. This is consistent with the finding that endo-siRNAs recently identified in fly soma and mouse oocytes appear to be dispensable for viability and reproduction [18-22]. Interestingly, mutants of *eri-1*, *rrf-3*, and *ergo-1* all exhibit an enhanced response to exogenous RNAi [32, 42, 43], whereas the *t22b3.2; zk757.3* double mutant does not (Fig. 5.5 E), suggesting that class II 26G RNAs may compete with the exogenous RNAi pathway for limiting common factors [14, 15].

5.5: Discussion

In this study, we identified a class of germline-enriched endo-siRNAs that are generated by a template-dependent mechanism and require the RRF-3 RNA-dependent RNA polymerase and the ERI-1 exonuclease for their biogenesis. In our model, class I and class II 26G RNAs are sorted into distinct, gamete-specific RISCs during germline development and differentially target discrete classes of target genes (Fig. 5.4D). Class I 26G RNAs repress genes associated with spermatogenesis while Class II 26G RNAs are maternally loaded and appear to be responsible for the clearance of maternal transcripts during zygotic development.

The presence of a 5' monophosphate is a signature for DICER processing. DCR-1 biochemically interacts with ERI-1 and RRF-3, two proteins required for 26G RNA biogenesis [14]. Fischer *et al.* used northern blotting to examine the expression of endo-siRNAs that target *K02E2.6* [44]. They observed two bands in the WT background, with the top band within the 26G RNA length range; both bands were abolished in the *dcr-1(ok247)* background. The 26G RNAs that target *K02E2.6* belong to the class II 26G RNAs. Additionally, Fischer *et al.* showed that mutants in *eri-6* and *er-7*, which exhibit an enhanced RNAi phenotype but are wild-type for spermatogenesis, are defective in the expression of endo-siRNAs that target *K02E2.6*. Taken together, these studies suggest DICER may be involved in the biogenesis of 26G RNAs and support our findings that defects in class II 26G RNAs result in an enhanced RNAi response.

Class I 26G RNAs preferentially act on transcripts associated with spermatogenesis. Ruby *et al.* postulated that they might suppress spermatogenesis genes to facilitate the switch from spermatogenesis to oogenesis; they also proposed that the RdRP EGO-1 might be involved because the *ego-1* mutant shows a delayed switch from spermatogenesis to oogenesis. We found, however, that the RRF-3 RdRP instead of EGO-1 is involved in 26G RNA biogenesis and that 26G RNA deficiency does not result in a delay in the initiation of oogenesis based on egg-to-egg time (Fig. 5.5D). We speculate that expression of genes required for spermatogenesis must be exquisitely and rapidly regulated and that class I 26G RNAs perform this function.

Class II 26G RNAs are generated in the maternal germline, loaded into embryos, and perdure throughout larval development. We hypothesize that they are critical for clearing the maternal transcripts during zygotic development. In zebrafish, miR-340 clears hundreds of maternal mRNAs during maternal-zygotic transition [45]. In our model, the class II 26G RNAs not only begin to clear the maternal load during oogenesis (for a subset of target genes) but are then maternally inherited to ensure that the maternal load of mRNAs continues to be cleared during filial development. The fact that the loss of class II 26G RNAs leads to enhanced RNAi phenotypes suggests that the ongoing transcript clearance competes with exogenous RNAi for limiting factors.

Several questions remain unanswered. Why are certain genes targeted by 26G RNAs? How do ERI-1 and RRF-3 participate in the biogenesis of 26G RNAs? Why do the loss of sperm 26G RNAs and consequent up-regulation of targets lead to *ts* sterility? Further genetic and biochemical analysis may reveal additional factors and mechanisms that mediate the biogenesis, sorting, differential stability, target silencing, and developmental functions of the class I and class II 26G RNAs.

5.6: Materials and methods

Strains and maintenance

The Bristol N2 was used as the reference wild type strain. Mutant alleles used in this study include: LG I: *glp-4(bn2), fer-1(hc1), rrf-1(pk1417), rrf-2(ok210), deps-1(bn121), deps-1(bn124), smg-1(r861)*; LG II: *rrf-3(pk1426)*; LG III: *zk757.3(tm1184)*; LG IV: *him-8(e1489), eri-1(mg366), t22b3.2(tm1155)*; LG V: *ergo-1(tm1860)*. *C. elegans* genetics and culture were performed as described [46]. Unless otherwise specified, worms were grown at 20°C.

Sperm, oocyte, and embryo purifications

Sperm and oocytes were purified as described with some modifications [47-49]. For sperm isolation, we used the *him-8(e1489)* strain, which increases the percentage of XO males to ~37% of the population versus ~0.2% males in the N2 wild-type strain [50]. Male worms from the *him-8(e1489)* strain were further isolated from hermaphrodites by filtering through a 35 µm nylon mesh filter as described [47], resulting in >95% males in the final sample. Isolated *him-8(e1489)* males were then subjected to 20,000 psi for 1 min, 3 times, to extrude and increase the yield of purified sperm. We used the *fer-1(hc1)* strain, which produces nonfunctional sperm at 25℃ [51], to obtain purified unfertilized oocytes. The *fer-1(hc1)* worms grown at 25°C were disrupted briefly in a Waring blender to release more oocytes from the body cavity. Sample purity (>95%) was inspected by DAPI staining and microscopy. Isolation of embryos from gravid adult worms was performed as described [52].

Total RNA isolation

RNA isolation was carried out using TriReagent (Ambion) following the vendor's protocol with the following modification: 3 times freeze/thaw/vortex was included to increase worm lysis efficiency; isopropanol precipitation of RNA was carried out at -80°C for one hour.

Construction of small RNA sequencing library

5' monophosphate-bearing small RNA libraries were constructed as described [53]. RNA oligos were purchased from Dharmacon and DNA oligos from Integrated DNA Technologies. Six Solexa libraries were constructed and sequenced on the 1G Genome Analyzer (Solexa/Illumina): N2 (mixed stage), sperm, oocyte, embryo, *eri-1(mg366)* young adult (YA), and *glp-4(bn2)* (YA). Five 454 libraries (sperm, oocyte, N2 (YA), *eri-*

1(mg366) (YA), and *glp-4(bn2)* (YA)) were sequenced on the Genome Sequencer FLX system (454/Roche).

Northern blotting of small RNAs

Due to limitation in sensitivity, relatively abundant 26G RNAs were selected for northern blotting (26G--O1, -O2, -S1, and -S5). An improved northern blot method using 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC)-mediated chemical crosslinking was performed as described [54]. For each assay, 5-10 µg of total RNA was used. For small RNA detection, DNA probes labeled with the Starfire Oligos Kit (IDT) were used.

RT-qPCR analysis of small RNA and mRNA levels

Custom small RNA Taqman assays were designed and synthesized by Applied Biosystems [55]. For each reaction, 50ng of total RNA was converted into cDNA with Multiscribe Reverse Transcriptase (Applied Biosystems) following the vendor's protocol. The resulting cDNAs were analyzed by a Realplex^2 thermocycler (Eppendorf) with TaqMan Universal PCR Master Mix, No AmpErase UNG (Applied Biosystems). Relative expression levels of small RNAs were calculated based on 2^{-ct} method [56]. For oocyte/embryo 26G RNA quantifications, miR-35 was used for normalization. For sperm 26G RNA quantifications, miR-1 was used for normalization. Gene targets of each class of 26G RNAs were selected based on 26G RNA cluster analysis (described below in supplementary computational methods). For quantification of mRNA levels, 250ng-1µg of total RNAs was converted into cDNAs with Multiscribe Reverse Transcriptase (Applied Biosystems) following the vendor's protocol. cDNAs were analyzed by a Realplex^2 thermocycler (Eppendorf) using Power Sybr Green PCR master mix (Applied Biosystems). Relative mRNA levels were calculated based on 2^{-ct} method using *act-1* for normalization.

Germline RNA in situ hybridization

RNA in situ hybridization was performed with dissected gonads according to Lee and Schedl [57]. Antisense cDNA fragments labeled with DIG DNA labeling Mix (Roche) for *C4G2.8* (547bp) and *ssp-16* (102bp) were used as probes. Probe detection was performed with alkaline-phosphate-conjugated anti-DIG (Fab2 fragment) from Roche and Sigma Fast BCIP/NBT.

RNA interference

Feeding RNAi was performed as described [58]. RNAi clones were picked from the Ahringer RNAi library [59].

5.7: Acknowledgements:

The authors thank John Moran, Harrison Gabel, Chi Zhang, Tammy Wu, Kris Gunsalus, Scott Kennedy, and Patrick Hu for discussions and helpful comments on the manuscript. We are grateful to Kris Gunsalus for comments on our computational analysis and Allison Billi for extensive editing of our manuscript. We thank the Caenorhabditis Genetics Center and Shohei Mitani for strains, Marco Marra and Martin Hirst of the British Columbia Cancer Centre for the Solexa deep sequencing, and David Miller of Applied Biosystems for assistance with acquiring Taqman probes. This research was supported by an NSF CAREER Award, an NIH S06 GM52588 grant to D.S.C., and an NIH HG004276-01 grant to J.K.K.

5.8: Supplementary computational methods

Sequence processing

All raw sequences (consolidating both 454 and Solexa) were processed with a custom Perl script to remove linker sequences and then mapped against the WS190 *C. elegans* genome using BLAST [1]. Sequences matching the genome with 0-2 mismatches were retained. Reads not matching the genome were mapped against Expressed Sequence Tags (EST) using BLAST to identify sequences that span exon-exon junctions. For reads matching more than one genomic locus, counts were normalized according to Ruby *et al. [2]*. For example, if a sequence had 20 reads and matched 2 genomic loci, each locus was assigned 10 reads. For all endo-siRNA analyses, reads corresponding to microRNAs [2], 21U RNAs [3, 4], and putative degradation products of non-coding RNAs (i.e. rRNAs, tRNAs, snRNAs, snoRNAs) were identified and excluded.

Genomic mapping of 26G RNAs

As outlined in Fig. 5.S1, we applied sequential filters to retain 26G RNAs with ≥ 2 reads in the 11 sequenced libraries and mapped them sequentially to Wormbase (WS190) and predicted gene models (Twinscan and Genefinder in WS190). Because 3'UTR regions are not well annotated, reads immediately downstream (within 500bp) of stop codons were annotated as overlapping with 3'UTR, which agrees well with the distribution of known 3' UTR lengths of annotated genes in Wormbase (Fig. 5.S2).

Cluster analysis of 26G RNAs

26G RNAs (≥ 2 total reads) were clustered using Cluster 3.0 software using hierarchical clustering, Euclidean distance and complete linkage options(copyright

Stanford University, 1998-99) and visualized using Java TreeView (open source). Clusters of the class I sperm 26G RNAs and the class II oocyte/embryo 26G RNAs were extracted from Java TreeView.

Target analysis of 26G RNAs

Targets of class I sperm 26G RNAs and class II oocyte/embryo 26G RNAs (extracted from clustering analysis) were annotated as spermatogenesis-enriched, oogenesis-enriched, germline-intrinsic, and “others” according to Reinke *et al.* [5]. For microarray analyses, raw CEL data from Asikainen *et al.* [6] were downloaded from NCBI Gene Expression Omnibus (Series GSE8659) and processed with dChip software [7]. Probe intensities corresponding to targets of sperm 26G RNAs were extracted from the CEL data.

Oligos for RT-qPCR

Gene	Forward (5’ to 3’)	Reverse (5’ to 3’)
act-1	CCAGGAATTGCTGATCGTATGCAGAA	TGGAGAGGGAAGCGAGGATAGA
C04G2.8	CGTGCTTCGACTGCAAAGAAGA	TTCTGTTGGCTTCTGCTGCG
C32E8.4	GAGCAACTTCTGCCGAAGGAA	CTTCAGGTTCTCCTTGAGCG
C40A11.10	AATGGCTCCTTGAAAAGATCG	TACATTTCCGCCACGTTGAAA
deps-1	GAAGGCTATGGCCGAAGTTCG	CAATGCGGTAACGGACAGATTT
dlc-6	CCGAAGGTTAAGCCACGTCATT	CTGCCATTGTGTATCATAATCCG
E01G4.7	GCACAAGGTTTCGTTCTTGGTG	AGTGACATCCCTTCTGATCG
F39E9.7	CCCAGTGGCCCAATTAAACG	CCCACGGCTTGTTCTTTGACA
F43E2.6	TGTAGGCGACGAGACTGATCG	TGCCGATGTTTCTGAGATGTCTT
F55B11.1	TTGATCGAGTCTCACTTTCCG	AAAGTCCACTGGTTCGTGATGAAT
F55C9.5	ACCATTGGAGCACGTAAATCAA	GGTCCTAATAATAAAGTTGCGTCG
fbxa-65	ACTTACAAGGATCAAGAAAAGCG	CCTTGACCGCTATTCCGAGAAA
fbxb-37	ATCGAAAGATGGAATACAAACCG	GACAAACATCCATCACATTCTTCG
gska-3	CGAGCAGACGACTCTGTGGAA	TTATTGAAACGCACAGTCTTCTCG
iff-1	CGAAGACCATAGAGAGTATGTCCG	CGAGCATTGCTTCGGGAAAGTA
K02E2.6	CAGTGGTACAAGTGGGAGTAAACG	AATTGGCAAGTAACTGATTCCG
K03H1.12	CAAAATTGCCACTTGTGATTCG	TCCAGTGAAGAGTGTCAAGAA

		CCA
msp-49	ATTAACTCCTCGGCTCGCCG	AGCTTCCTTTGGGTCGAGGAC
snf-6	GGATTGTTGGCTACTGGCCG	TCAAGCCAAAGGAAGCAAAGAA
sod-1	GATCTATGGTTGTTCATGCCG	CTTCTGCCTTGTCTCCGACTCC
ssp-16	GTCATCAAACAACAATGAGTACCG	GCTCCAGCAGTGCGAGTGAT
ssp-19	GCACCGAAGGAAGACAAGCTG	GAGCCACTGCAACAAAAGCG
T05E12.8	TTCCATTTGAGGATTTTGCTACG	ATTATTTGGATGGCAGCCGATG
T08B2.12	GAAACCAATGCTCCAGTTGATAC	GATGAAAGCGATGGACGAGAAG
T25G12.11	ACGTGCTTTCTGATTCACTCCG	CATGGGTGGGATGAGAGCAC
tax-2	GATTAATCCAAGACAAGTTCCTAAATTGAT	TTCAATTCTTGAACTCCTTTGTTTTC
Tc1	AACCGTTAAGCATGGAGGTG	CACATGACGACGTTGAAACC
Tc3	GAGCGTTCACGGAGAAGAAG	AATAGTCGCGGGTTGAGTTG
tdc-1	GAACTTCGTCAGAGATTCCCG	TCTCAACGGAAGAATGGGCTTC
U6	TGGAACAATACAGAGAAGATTAGCA	CTTCACGAATTTGCGTGTCAT
W05H12.2	GCTCAAGACCAGATAATGCTTGGA	CAATCCCAAAGATTCAATACCG
Y37E11B.2	AATGGAGACTCTTCTTCCACCCG	AGCGAAGGCATTGATCTTGGTT
Y7A5A.11	CCATTACTTTCAACATGCCG	TCCTTGTTCCAGCACTAGCAGA
Y82E9BR.20	CTCCCGCTTTCTTGATGTATTG	AGTCCGAACTCATCCAAAGCAG
ZC168.6	GTCCAGTTTATGGGTTCGTGGATG	AGTCTCTTCGGCTGGCACTTC
ZC328.1	GGGCGGTCATTTCTATTGTTTG	GCCAAATTGGTCCGTAATCTT

		GT
ZK484.5	CCGTCAGACAACTGCTCTCCTC	GGTTGGGCTGCTTCAGAGTC

Oligos for small RNA cloning

5' RNA adaptor:	5' GUUCAGAGUUCUACAGUCCGACGAUC 3'
3' RNA adaptor:	5' pUCGUAUGCCGUCUUCUGCUUGidT 3' p = phosphate; idT = inverted deoxythymidine
RT-primer (DNA):	5' CAAGCAGAAGACGGCATACGA 3'
P7 primer (DNA):	5' CAAGCAGAAGACGGCATACGA 3'
P5 long primer (DNA):	5' AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA 3'

Oligos for northern blotting

21UR-1	5' GCACGGTTAACGTACGTACCA /3StarFire/ 3'
26G-O1	5' TTGAAAATAATCTACCGTTTCTGAGC /3StarFire/ 3'
26G-O2	5' CATTTGCTGCAATTATGAGTCATAAC /3StarFire/ 3'
26G-S1	5' AATTATGTATTCTCGTCCTCCATAGC /3StarFire/ 3'
26G-S5	5' TACCATGTCGCTCACTGCTGATCCAC /3StarFire/ 3'
cel-miR-35	5' ACTGCTAGTTTCCACCCGGTGA /3StarFire/ 3'
cel-miR-1	5' TACATACTTCTTTACATTCCA /3StarFire/ 3'

5.9: Reference

1. Mello CC, Conte D, Jr.: **Revealing the world of RNA interference**. *Nature* 2004, **431**(7006):338-342.
2. Malone CD, Hannon GJ: **Small RNAs as guardians of the genome**. *Cell* 2009, **136**(4):656-668.
3. Bernstein E, Caudy AA, Hammond SM, Hannon GJ: **Role for a bidentate ribonuclease in the initiation step of RNA interference**. *Nature* 2001, **409**(6818):363-366.
4. Grishok A, Pasquinelli AE, Conte D, Li N, Parrish S, Ha I, Baillie DL, Fire A, Ruvkun G, Mello CC: **Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing**. *Cell* 2001, **106**(1):23-34.
5. Hutvagner G, McLachlan J, Pasquinelli AE, Balint E, Tuschl T, Zamore PD: **A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA**. *Science* 2001, **293**(5531):834-838.
6. Ketting RF, Fischer SE, Bernstein E, Sijen T, Hannon GJ, Plasterk RH: **Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans**. *Genes Dev* 2001, **15**(20):2654-2659.
7. Knight SW, Bass BL: **A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans**. *Science* 2001, **293**(5538):2269-2271.
8. Liu J, Carmell MA, Rivas FV, Marsden CG, Thomson JM, Song JJ, Hammond SM, Joshua-Tor L, Hannon GJ: **Argonaute2 is the catalytic engine of mammalian RNAi**. *Science* 2004, **305**(5689):1437-1441.
9. Tabara H, Sarkissian M, Kelly WG, Fleenor J, Grishok A, Timmons L, Fire A, Mello CC: **The rde-1 gene, RNA interference, and transposon silencing in C. elegans**. *Cell* 1999, **99**(2):123-132.
10. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function**. *Cell* 2004, **116**(2):281-297.
11. Aravin AA, Hannon GJ, Brennecke J: **The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race**. *Science* 2007, **318**(5851):761-764.
12. Okamura K, Lai EC: **Endogenous small interfering RNAs in animals**. *Nat Rev Mol Cell Biol* 2008, **9**(9):673-678.
13. Ambros V, Lee RC, Lavanway A, Williams PT, Jewell D: **MicroRNAs and other tiny endogenous RNAs in C. elegans**. *Curr Biol* 2003, **13**(10):807-818.
14. Duchaine TF, Wohlschlegel JA, Kennedy S, Bei Y, Conte D, Jr., Pang K, Brownell DR, Harding S, Mitani S, Ruvkun G *et al*: **Functional proteomics reveals the biochemical niche of C. elegans DCR-1 in multiple small-RNA-mediated pathways**. *Cell* 2006, **124**(2):343-354.
15. Lee RC, Hammell CM, Ambros V: **Interacting endogenous and exogenous RNAi pathways in Caenorhabditis elegans**. *Rna* 2006, **12**(4):589-597.
16. Asikainen S, Storvik M, Lakso M, Wong G: **Whole genome microarray analysis of C. elegans rrf-3 and eri-1 mutants**. *FEBS Lett* 2007, **581**(26):5050-5054.
17. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP: **Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans**. *Cell* 2006, **127**(6):1193-1207.

18. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T *et al*: **Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes**. *Nature* 2008, **453**(7194):539-543.

19. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM *et al*: **Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes**. *Nature* 2008, **453**(7194):534-538.

20. Okamura K, Chung WJ, Ruby JG, Guo H, Bartel DP, Lai EC: **The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs**. *Nature* 2008, **453**(7196):803-806.

21. Kawamura Y, Saito K, Kin T, Ono Y, Asai K, Sunohara T, Okada TN, Siomi MC, Siomi H: **Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells**. *Nature* 2008, **453**(7196):793-797.

22. Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, Xu J, Kittler EL, Zapp ML, Weng Z *et al*: **Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells**. *Science* 2008, **320**(5879):1077-1081.

23. Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, Perrimon N, Kellis M, Wohlschlegel JA, Sachidanandam R *et al*: **An endogenous small interfering RNA pathway in Drosophila**. *Nature* 2008, **453**(7196):798-802.

24. Seydoux G, Strome S: **Launching the germline in Caenorhabditis elegans: regulation of gene expression in early germ cells**. *Development* 1999, **126**(15):3275-3283.

25. Kimble J, Crittenden SL: **Controls of germline stem cells, entry into meiosis, and the sperm/oocyte decision in Caenorhabditis elegans**. *Annu Rev Cell Dev Biol* 2007, **23**:405-433.

26. O'Donnell KA, Boeke JD: **Mighty Piwis defend the germline against genome intruders**. *Cell* 2007, **129**(1):37-44.

27. Cox DN, Chao A, Baker J, Chang L, Qiao D, Lin H: **A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal**. *Genes Dev* 1998, **12**(23):3715-3727.

28. Wang G, Reinke V: **A C. elegans Piwi, PRG-1, regulates 21U-RNAs during spermatogenesis**. *Curr Biol* 2008, **18**(12):861-867.

29. Batista PJ, Ruby JG, Claycomb JM, Chiang R, Fahlgren N, Kasschau KD, Chaves DA, Gu W, Vasale JJ, Duan S *et al*: **PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C. elegans**. *Mol Cell* 2008, **31**(1):67-78.

30. Das PP, Bagijn MP, Goldstein LD, Woolford JR, Lehrbach NJ, Sapetschnig A, Buhecha HR, Gilchrist MJ, Howe KL, Stark R *et al*: **Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the Caenorhabditis elegans germline**. *Mol Cell* 2008, **31**(1):79-90.

31. Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ: **An epigenetic role for maternally inherited piRNAs in transposon silencing**. *Science* 2008, **322**(5906):1387-1392.

32. Kennedy S, Wang D, Ruvkun G: **A conserved siRNA-degrading RNase negatively regulates RNA interference in C. elegans**. *Nature* 2004, **427**(6975):645-649.

33. Reinke V, Gil IS, Ward S, Kazmer K: **Genome-wide germline-enriched and sex-biased expression profiles in Caenorhabditis elegans**. *Development* 2004, **131**(2):311-323.

34. Makeyev EV, Bamford DH: **Cellular RNA-dependent RNA polymerase involved in posttranscriptional gene silencing has two distinct activity modes**. *Mol Cell* 2002, **10**(6):1417-1427.

35. Sijen T, Fleenor J, Simmer F, Thijssen KL, Parrish S, Timmons L, Plasterk RH, Fire A: **On the role of RNA amplification in dsRNA-triggered gene silencing**. *Cell* 2001, **107**(4):465-476.

36. Smardon A, Spoerke JM, Stacey SC, Klein ME, Mackin N, Maine EM: **EGO-1 is related to RNA-directed RNA polymerase and functions in germ-line development and RNA interference in C. elegans**. *Curr Biol* 2000, **10**(4):169-178.

37. Spike CA, Bader J, Reinke V, Strome S: **DEPS-1 promotes P-granule assembly and RNA interference in C. elegans germ cells**. *Development* 2008, **135**(5):983-993.

38. Pak J, Fire A: **Distinct Populations of Primary and Secondary Effectors During RNAi in C. elegans**. *Science* 2007, **315**(5809):241-244.

39. Sijen T, Steiner FA, Thijssen KL, Plasterk RHA: **Secondary siRNAs Result from Unprimed RNA Synthesis and Form a Distinct Class**. *Science* 2007, **315**(5809):244-247.

40. Aoki K, Moriguchi H, Yoshioka T, Okawa K, Tabara H: **In vitro analyses of the production and activity of secondary small interfering RNAs in C. elegans**. *EMBO J* 2007, **26**(24):5007-5019.

41. Faehnle CR, Joshua-Tor L: **Argonautes confront new small RNAs**. *Curr Opin Chem Biol* 2007, **11**(5):569-577.

42. Yigit E, Batista PJ, Bei Y, Pang KM, Chen CC, Tolia NH, Joshua-Tor L, Mitani S, Simard MJ, Mello CC: **Analysis of the C. elegans Argonaute family reveals that distinct Argonautes act sequentially during RNAi**. *Cell* 2006, **127**(4):747-757.

43. Simmer F, Tijsterman M, Parrish S, Koushika SP, Nonet ML, Fire A, Ahringer J, Plasterk RH: **Loss of the putative RNA-directed RNA polymerase RRF-3 makes C. elegans hypersensitive to RNAi**. *Curr Biol* 2002, **12**(15):1317-1319.

44. Fischer SE, Butler MD, Pan Q, Ruvkun G: **Trans-splicing in C. elegans generates the negative RNAi regulator ERI-6/7**. *Nature* 2008, **455**(7212):491-496.

45. Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, Enright AJ, Schier AF: **Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs**. *Science* 2006, **312**(5770):75-79.

46. Brenner S: **The genetics of Caenorhabditis elegans**. *Genetics* 1974, **77**(1):71-94.

47. L'Hernault SW, Roberts TM: **Cell biology of nematode sperm**. *Methods Cell Biol* 1995, **48**:273-301.

48. Chu DS, Liu H, Nix P, Wu TF, Ralston EJ, Yates JR, 3rd, Meyer BJ: **Sperm chromatin proteomics identifies evolutionarily conserved fertility factors**. *Nature* 2006, **443**(7107):101-105.

49. Aroian RV, Field C, Pruliere G, Kenyon C, Alberts BM: **Isolation of actin-associated proteins from Caenorhabditis elegans oocytes and their localization in the early embryo**. *EMBO J* 1997, **16**(7):1541-1549.

50. Brenner S, Hodgkin,J , Horvitz, R . **Nondisjunction mutants of the nematode C. elegans**. *Genetics* 1979(91):67-94.

51. Ward S, Argon Y, Nelson GA: **Sperm morphogenesis in wild-type and fertilization-defective mutants of Caenorhabditis elegans**. *J Cell Biol* 1981, **91**(1):26-44.

52. Stiernagle T: **Maintenance of C. elegans**. *WormBook* 2006:1-11.

53. Lu C, Meyers BC, Green PJ: **Construction of small RNA cDNA libraries for deep sequencing**. *Methods* 2007, **43**(2):110-117.

54. Pall GS, Hamilton AJ: **Improved northern blot method for enhanced detection of small RNA**. *Nat Protoc* 2008, **3**(6):1077-1084.

55. Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, Barbisin M, Xu NL, Mahuvakar VR, Andersen MR *et al*: **Real-time quantification of microRNAs by stem-loop RT-PCR**. *Nucleic Acids Res* 2005, **33**(20):e179.

56. Nolan T, Hands RE, Bustin SA: **Quantification of mRNA using real-time RT-PCR**. *Nat Protoc* 2006, **1**(3):1559-1582.

57. Lee MH, Schedl T: **RNA in situ hybridization of dissected gonads**. *WormBook* 2006:1-7.

58. Kamath RS, Ahringer J: **Genome-wide RNAi screening in Caenorhabditis elegans**. *Methods* 2003, **30**(4):313-321.

59. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M *et al*: **Systematic functional analysis of the Caenorhabditis elegans genome using RNAi**. *Nature* 2003, **421**(6920):231-237.

5.10: Supplementary Reference

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
2. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP: **Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans**. *Cell* 2006, **127**(6):1193-1207.
3. Batista PJ, Ruby JG, Claycomb JM, Chiang R, Fahlgren N, Kasschau KD, Chaves DA, Gu W, Vasale JJ, Duan S *et al*: **PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C. elegans**. *Mol Cell* 2008, **31**(1):67-78.
4. Das PP, Bagijn MP, Goldstein LD, Woolford JR, Lehrbach NJ, Sapetschnig A, Buhecha HR, Gilchrist MJ, Howe KL, Stark R *et al*: **Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the Caenorhabditis elegans germline**. *Mol Cell* 2008, **31**(1):79-90.
5. Reinke V, Gil IS, Ward S, Kazmer K: **Genome-wide germline-enriched and sex-biased expression profiles in Caenorhabditis elegans**. *Development* 2004, **131**(2):311-323.
6. Asikainen S, Storvik M, Lakso M, Wong G: **Whole genome microarray analysis of C. elegans rrf-3 and eri-1 mutants**. *FEBS Lett* 2007, **581**(26):5050-5054.
7. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection**. *Proc Natl Acad Sci U S A* 2001, **98**(1):31-36.

5.11: Figures

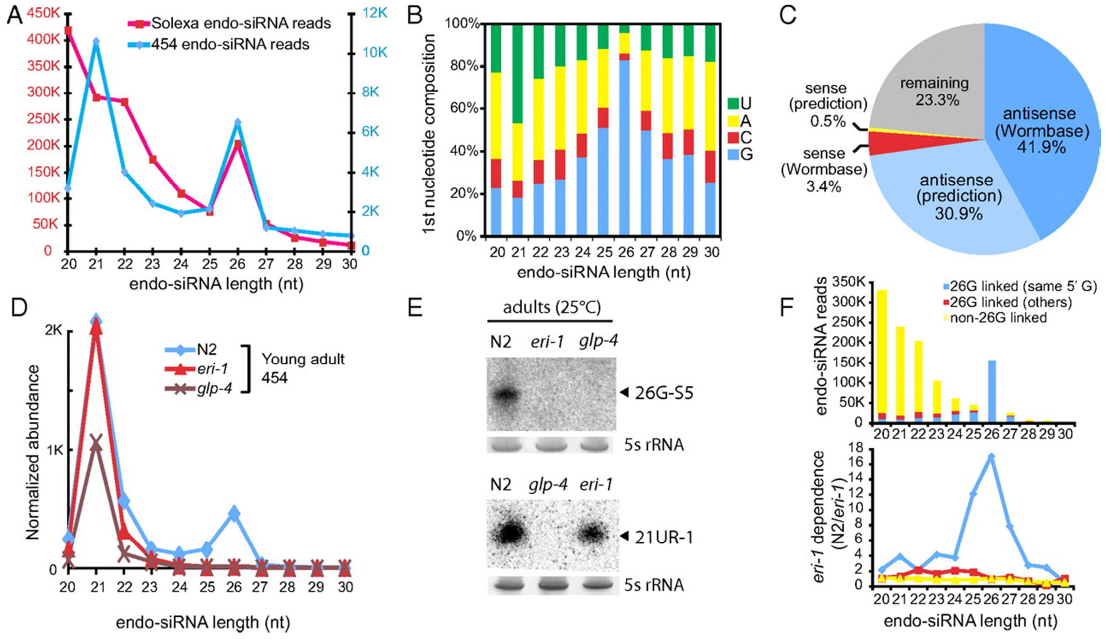


Figure 5.1: 26G RNAs are germline-enriched endogenous siRNAs.

A.) Length distribution of endo-siRNAs exhibits a bimodal pattern, peaking at both 21nt and 26nt length. Small RNA libraries were sequenced and combined to analyze the size distribution of endo-siRNAs. Small RNA libraries of mixed-stage N2 animals, purified male sperm (*him-8(e1489)*), purified oocytes (*fer-1(hc-1)*), and N2 embryos were sequenced by Solexa (Illumina). Small RNA libraries of N2 (young adult), sperm (*him-8(e1489)*), and oocytes (*fer-1(hc-1)*) were sequenced by 454 (Roche).

B.) First nucleotide identity of endo-siRNAs. 26nt endo-siRNAs have a strong preference for guanine as the first nucleotide (83%).

C.) The majority of 26G RNAs are anti-sense to known and predicted coding transcripts. Coding gene transcripts were defined by Wormbase gene annotations (WS190) and gene predictions (Twinscan and Genefinder predictions in Wormbase, WS190 coordinates), and assignment of the 3'UTRs as being up to 500bp downstream of CDS ends (see Fig. S2). The remaining intergenic 26G RNA sequences (23.3%) may also target genes yet to be identified by current Wormbase gene annotations and predictions.

D.) Normalized length distribution of endo-siRNAs in N2, *eri-1(mg366)*, and *glp-4(bn2)* young adult libraries sequenced by 454 (Roche). The abundance was normalized to 100K effective small RNA reads (total reads minus potential degradation products from rRNAs, tRNAs, snRNAs, and snoRNAs).

E.) Northern blotting validates the lack of 26G RNA expression in *eri-1(mg366)* and *glp-4(bn2)* mutants. Total RNA from N2, *eri-1(mg366)*, and *glp-4(bn2)* adult worms was probed for a 26G RNA (26G-S5) and a 21U RNA (21UR-1) by northern blotting. The expression of the germline-derived 21U RNA (21UR-1) is not *eri-1*-dependent (lower panel). Ethidium bromide (EtBr) stained 5s rRNA serves as the loading control.

F.) Endo-siRNAs were classified as 26G RNA-linked (targeting the same genes) or non-26G RNA-linked (targeting other genes or intergenic regions). Most 26G RNA-linked endo-siRNAs start with the same 5' G. A small fraction of shorter length (20-24) endo-siRNAs is 26G RNA-linked. The bottom panel plots the *eri-1* dependence as measured by the ratio of counts in N2 vs. *eri-1*.

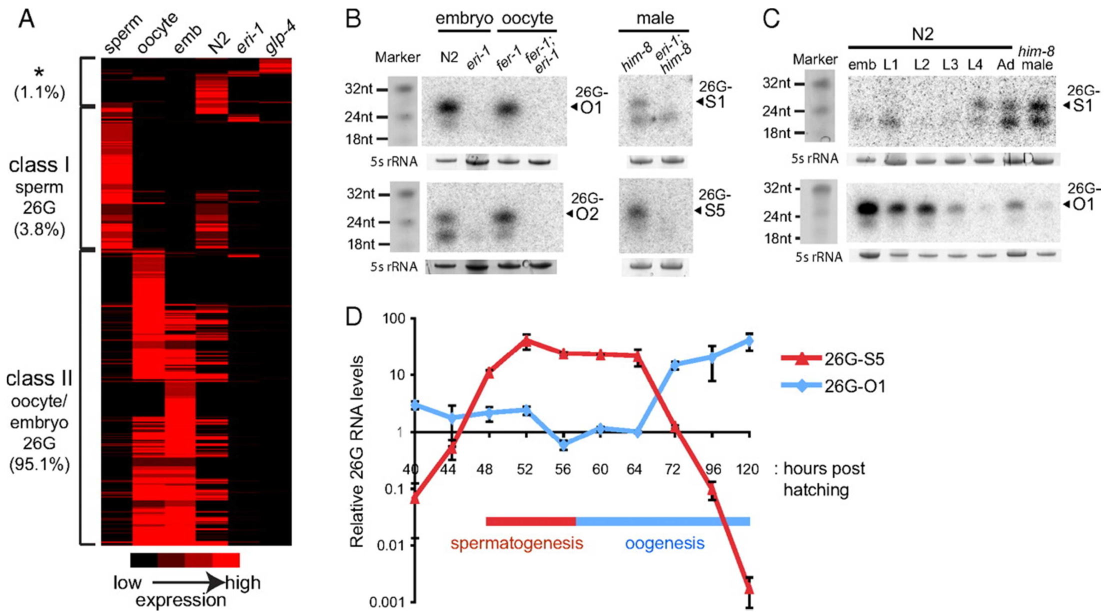


Figure 5.2: Two classes of 26G RNAs exhibit different expression patterns.

A.) Hierarchical clustering reveals two major classes of 26G RNAs: class I sperm 26G RNAs (3.8% of total reads) and class II oocyte/embryo 26G RNAs (95.1% total reads). 26G RNA reads matching to the *C. elegans* genome with at least two counts were included in the analysis (4,002 unique sequences; 156,204 total reads). The asterisk (*) indicates a small fraction (1.1%) of 26G RNA sequences that do not fall into either class I or II categories; these sequences are generally not abundant, with 3.6 total reads on average per unique 26G RNA sequence.

B.) Both classes of 26G RNAs are dependent on *eri-1* for their expression. Total RNA from embryos and oocytes of indicated genotypes was probed for two class II oocyte/embryo 26G RNAs (26G-O1, -O2); total RNA from *him-8(e1489)* and *eri-1(mg366);him-8(e1489)* adult males was probed for two class I sperm 26G RNAs (26G-S1, -S5). The 5s rRNA serves as a sample loading control.

C.) Class I and class II 26G RNAs are expressed in distinct periods during development. Total RNA from embryos (emb), four larval stages, adult hermaphrodites (Ad), and *him-8(e1489)* adult males was analyzed by northern blotting with probes for a class I sperm 26G RNA (26G-S1) and a class II oocyte/embryo 26G RNA (26G-O1). Synthetic RNA oligos stained with EtBr serve as size markers and 5s rRNA serves as a sample loading control.

D.) Analysis of 26G RNA levels during germline proliferation assayed by RT-qPCR. The expression of class I sperm 26G RNA (26G-S5) and class II oocyte/embryo 26G RNA (26G-O1) correlate with the time windows for spermatogenesis and oogenesis, respectively. The X-axis represents hours post hatching at 20°C.

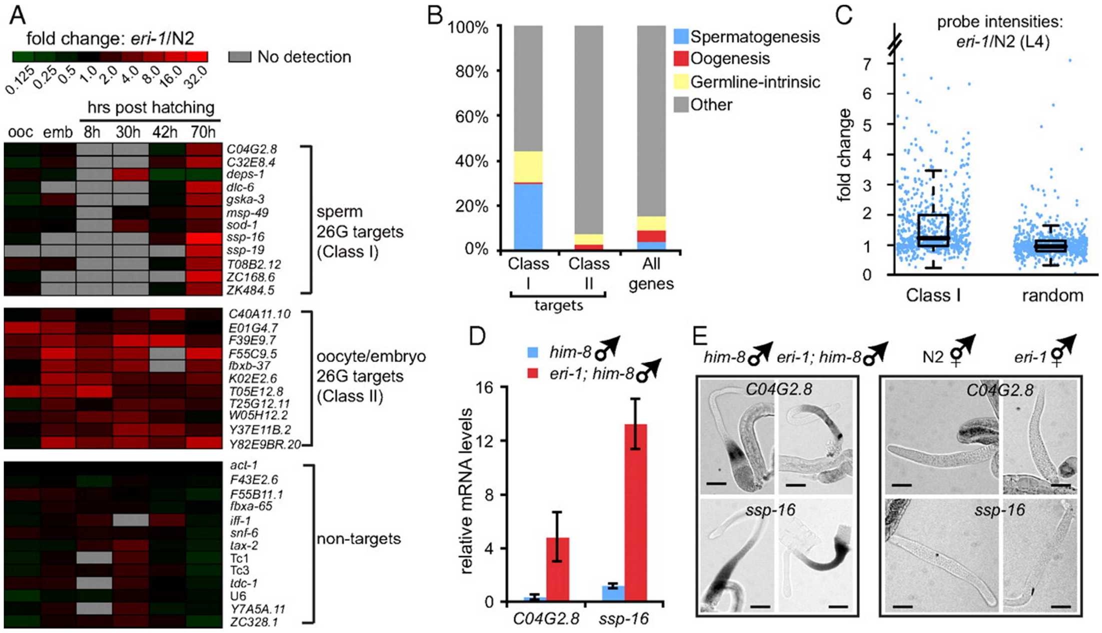


Figure 5.3: Two classes of 26G RNAs silence non-overlapping sets of mRNA transcripts.

A.) Gene targets of 26G RNAs are desilenced in the *eri-1(mg366)* background. Differential gene expression profiles between N2 and *eri-1(mg366)* for 12 targets of class I sperm 26G RNAs, 11 targets of class II oocyte/embryo 26G RNAs, and 13 non-targets. The level of fold up-regulation is represented according to the red-green color scheme indicated in the top panel. Abbreviations: ooc (oocytes), emb (embryo).

B.) Gene class analyses of class I sperm and class II oocyte/embryo 26G RNAs. Targets of class I sperm 26G RNAs (573 genes) are significantly overrepresented in genes expressed during spermatogenesis, while targets of class II oocyte/embryo 26Gs (243 genes) are depleted of germline enriched genes (i.e. spermatogenesis, oogenesis, and germline-intrinsic).

C.) Genes targeted by class I sperm 26G RNAs are up-regulated in the *eri-1(mg366)* mutant. Each point indicates the fold change in probe intensity corresponding to predicted targets of 26G RNAs (728 probes corresponding to 589 genes). Randomly selected probes do not show up-regulation in the *eri-1(mg366)* mutant.

D.) Loss of 26G RNA expression does not induce inappropriate ectopic expression of targets. RNA in situ hybridization of dissected gonads was performed with probes for the class I sperm 26G RNA targets *C02G2.8* and *ssp-16*. In both wild-type and *eri-1* backgrounds, expression of these two genes remained restricted to the spermatogenic gonad. No ectopic expression of the class I 26G RNA targets was observed in the hermaphrodite oogenic gonads. Bar, 50µm.

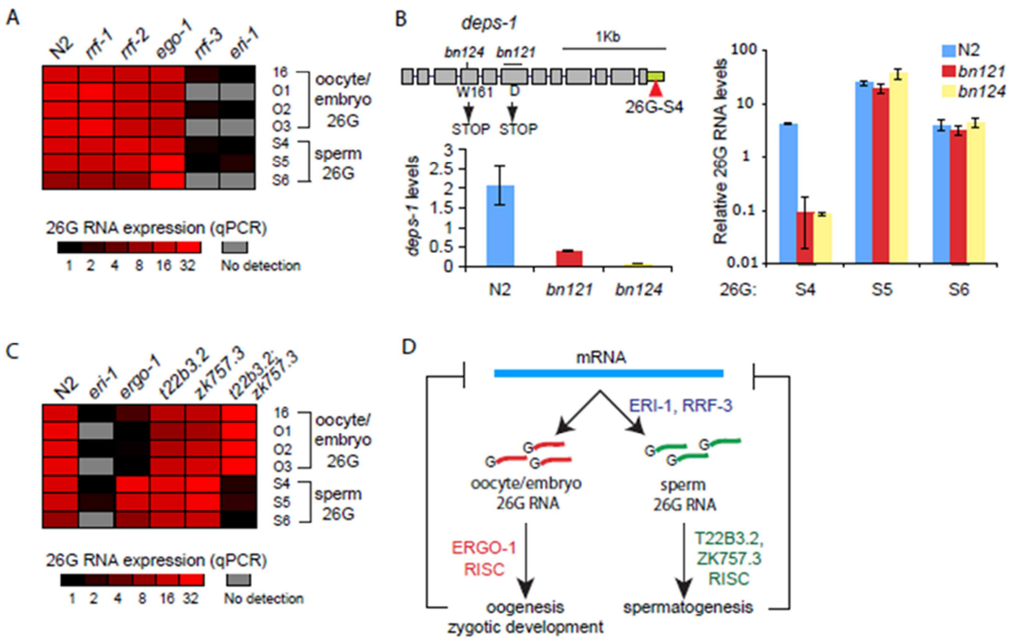


Figure 5.4: Genetic requirements for 26G biogenesis and function.
A.) RT-qPCR analysis of 26G RNA expression in *rrf-1(ok589), rrf-2(ok210), rrf-3(pk1426*), and *ego-1(RNAi).* Mutation of *rrf-3* abrogates the expression of both sperm and oocyte/embryo 26G RNAs, while the 26G RNAs are expressed at wild-type levels in the mutants of *rrf-1* and *rrf-2,* as well as in RNAi-inactivation of *ego-1.*

B.) Requirement of target mRNA transcript for 26G RNA biogenesis. Two *deps-1* mutant alleles (*bn121* and *bn124*) harbor premature stop codons that destabilize the *deps-1* transcript. The expression of the class I 26G RNA 26G-S4, which is antisense to the *deps-1* 3'UTR (green), is compromised in the *deps-1* mutants, while the expression of other sperm 26G RNAs that do not target *deps-1* remains unchanged. Both *deps-1* and 26G RNA levels were measured by RT-qPCR. Error bars indicate standard deviation for replicates.

C.) An oogenesis-enriched Argonaute encoded by *ergo-1* is required for class II oocyte/embryo 26G RNA expression, but dispensable for class I sperm 26G RNA expression. The *t22b3.2(tm1155)*; *zk757.3(tm1184)* double mutant is defective in sperm 26G RNAs, but expresses normal levels of oocyte/embryo 26G RNAs.

D.) Proposed model for 26G RNA biogenesis and function.

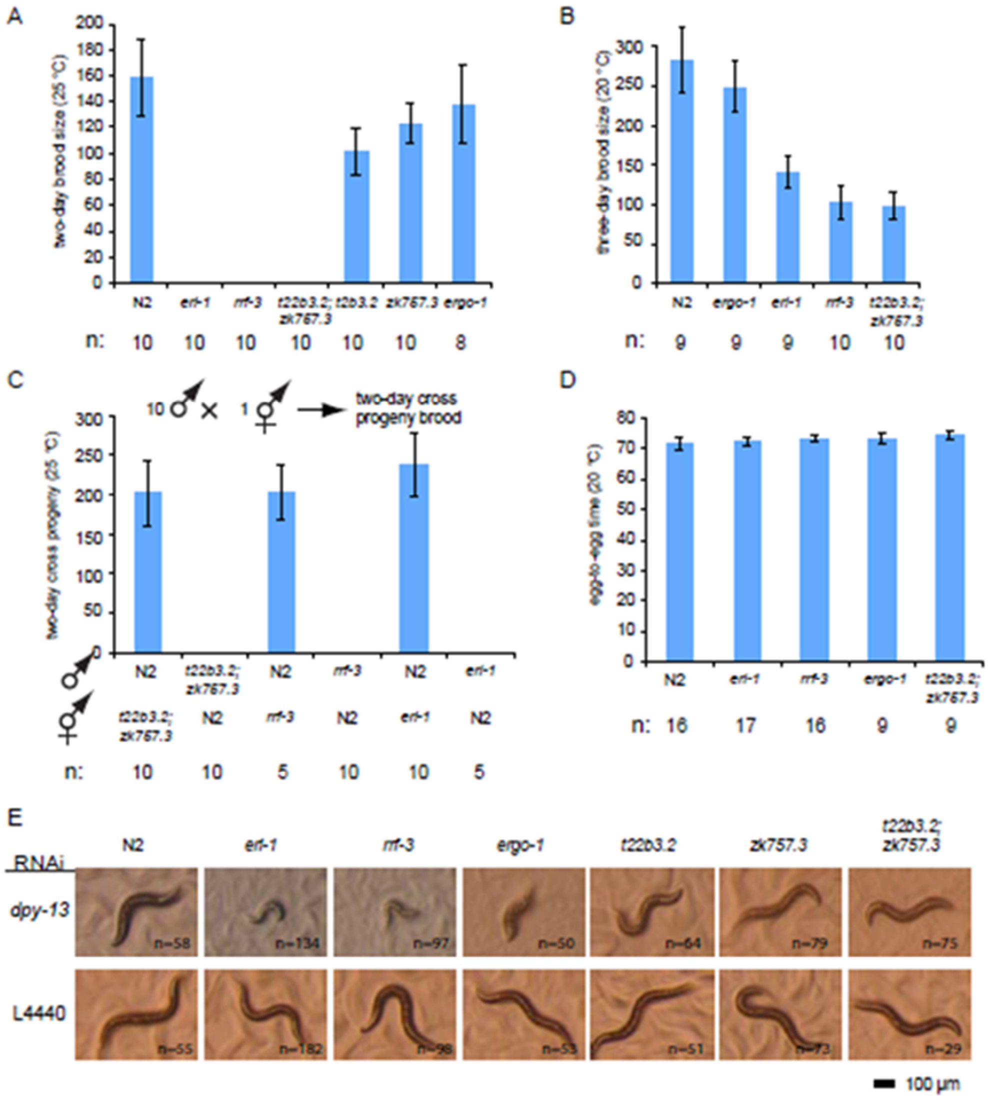


Figure 5.5: Phenotypes of mutants defective in 26G RNAs.

(A-B) The *t22b3.2(tm1155); zk757.3(tm1184)* double mutant is sterile at 25°C and exhibits significant loss of fertility at 20°C. Synchronized worms were singled at L4 stage and progeny brood size was counted for the subsequent two days. N is the number of parents assayed. Error bars represent standard deviation. Alleles used in this assay: *eri-1(mg366), rrf-3(pk1426)*, *ergo-1(tm1860)*, *t22b3.2(tm1155)*, *zk757.3(tm1184)*.

(C) The *ts* sterility of *t22b3.2*; *zk757.3*, *eri-1*, and *rrf-3* can be fully rescued by WT males. For each cross, 10 males were crossed with 1 hermaphrodite, and two day cross progeny brood was scored.

(D) N2, *t22b3.2(tm1155)*; *zk757.3(tm1184)*, *eri-1*, and *rrf-3* have similar egg-to-egg time at 20°C.

(E) The *t22b3.2(tm1155); zk757.3(tm1184)* double mutant does not display an enhanced RNAi phenotype. Synchronized L1 worms of indicated genotypes were subjected to feeding RNAi of *dpy-13* or control vector. L4 and young adult worms were examined for the severity of dumpy phenotype. A moderate dumpy phenotype was observed in N2, *t22b3.2(tm1155), zk757.3(tm1184),* and the *t22b3.2(tm1155); zk757.3(tm1184)* double mutant. In contrast, RNAi inactivation of *dpy-13* in *eri-1(mg366), rrf-3(pk1426)*, and *ergo-1(tm1860)* generated a severe dumpy phenotype, indicating hypersensitivity to exogenous RNAi of *dpy-13*.

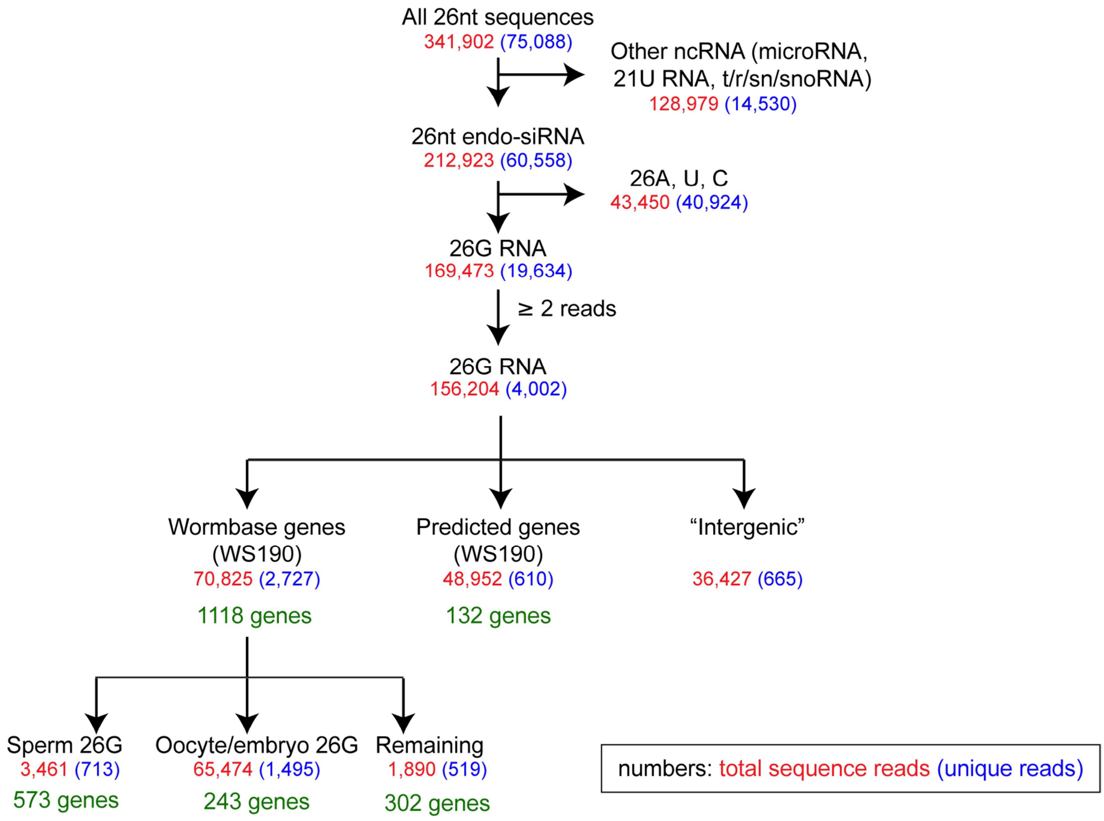


Figure 5.S1: Computational pipeline for 26G RNA annotations.
All 26nt genome BLAST hits were extracted from our datasets. Sequences matching noncoding RNAs (i.e. tRNAs, rRNAs, snRNAs, snoRNAs) and other classes of small RNAs (microRNAs, 21U RNAs) were identified and excluded from the analyses. Two additional filters were applied to retain sequences starting with guanine and having ≥ 2 sequence reads. 26G RNAs mapping within 500bp downstream of Wormbase gene annotations (WS190) and gene predictions (Twinscan, Genefinder predictions from Wormbase) were sequentially annotated. In sum, 1,118 Wormbase-annotated genes and 132 Wormbase-predicted genes were identified to be targets of 26G RNAs. 26G RNAs derived from Wormbase-annotated genes were further clustered into sperm 26G RNA (with 573 gene targets) and oocyte/embryo 26G RNA (with 243 gene targets).

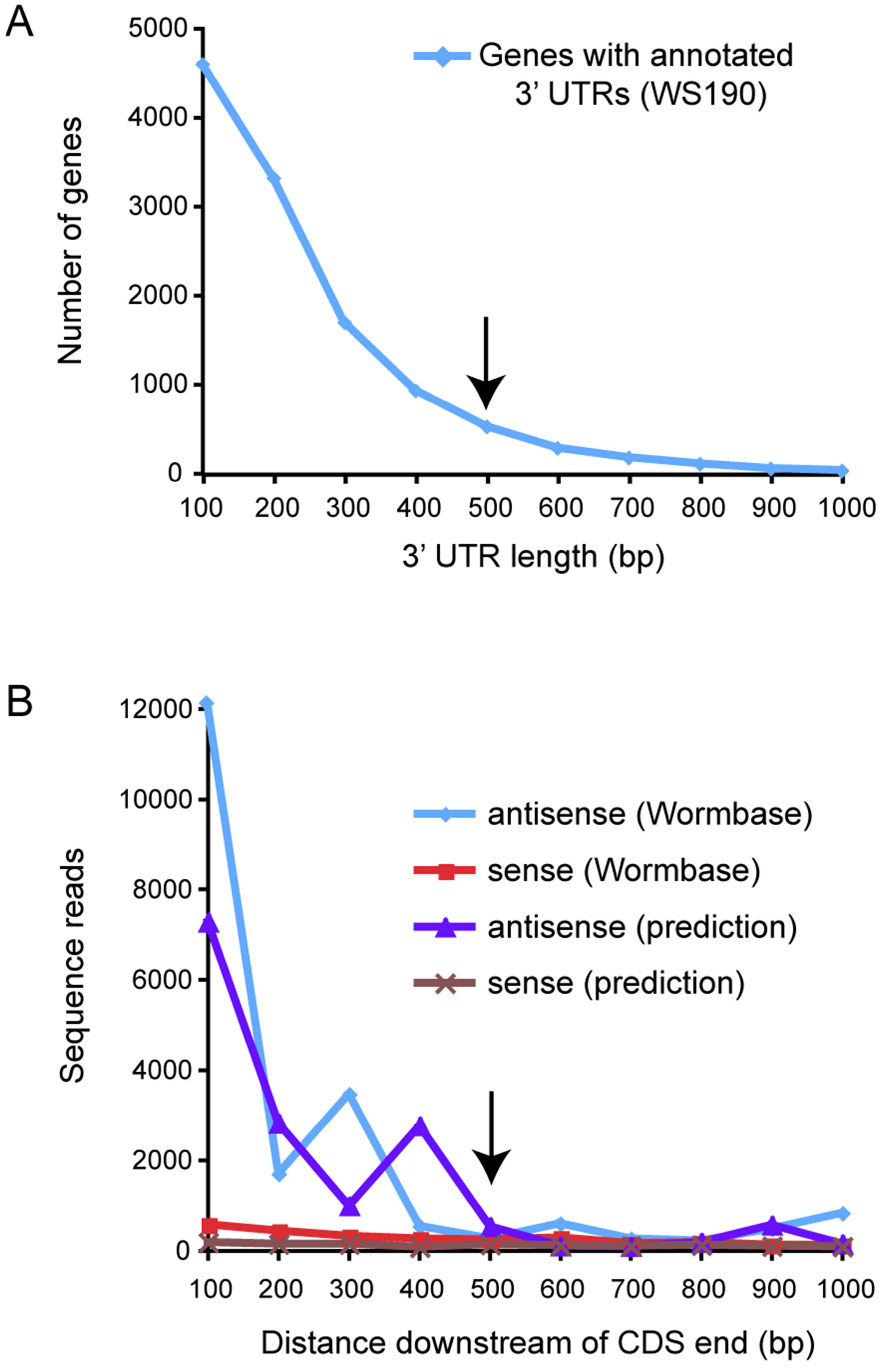

A
5000
4000
3000
2000
1000
0
Number of genes
Genes with annotated 3' UTRs (WS190)
100
200
300
400
500
600
700
800
900
1000
3' UTR length (bp)
B
12000
10000
8000
6000
4000
2000
0
Sequence reads
antisense (Wormbase)
sense (Wormbase)
antisense (prediction)
sense (prediction)
100
200
300
400
500
600
700
800
900
1000
Distance downstream of CDS end (bp)

Figure 5.S2: Distribution of 26G RNA reads originating from putative 3' UTR regions.

A.) The 3' UTR length distribution of genes in Wormbase. Arrow at 500nt indicates the 95% cutoff.

B.) Number of 26G RNA reads that mapped within every 100bp up to 1Kb downstream of the ends of the coding sequences (stop codons) was plotted. The majority of reads are antisense to mRNAs and map within 500 bp (arrow) downstream of stop codons.

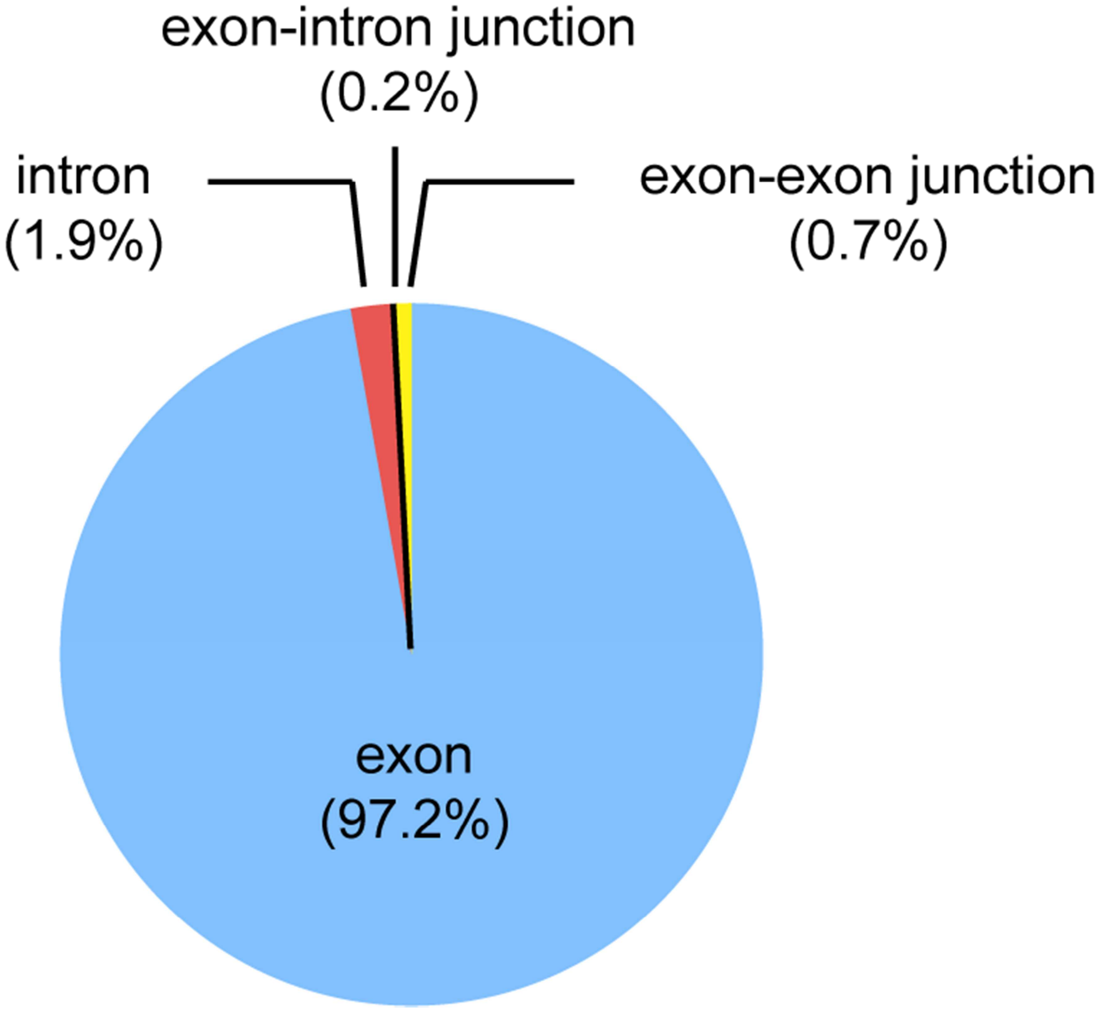


Figure 5.S3: 26G RNAs mapping to exons and introns.
26G RNA counts matching exons, introns, exon-intron junctions and exon-exon junctions of Wormbase genes were plotted. The majority of reads (97.9%) are derived from exons.

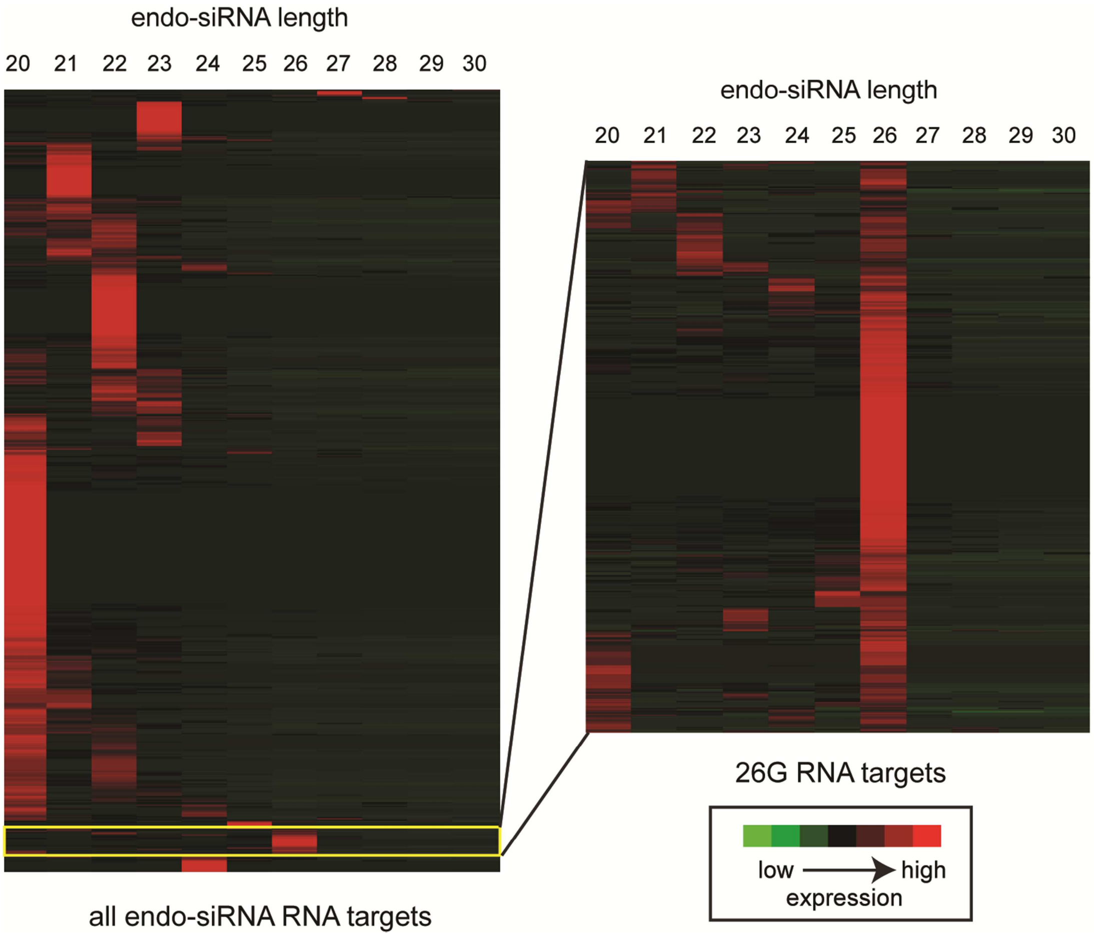


Figure 5.S4: 26G RNAs targets are a unique class of genes.
Endo-siRNA targets (Wormbase WS190) were clustered (left) based on the abundance of endo-siRNAs of different lengths. 26G RNAs targets are predominantly targeted by 26G RNAs (right).

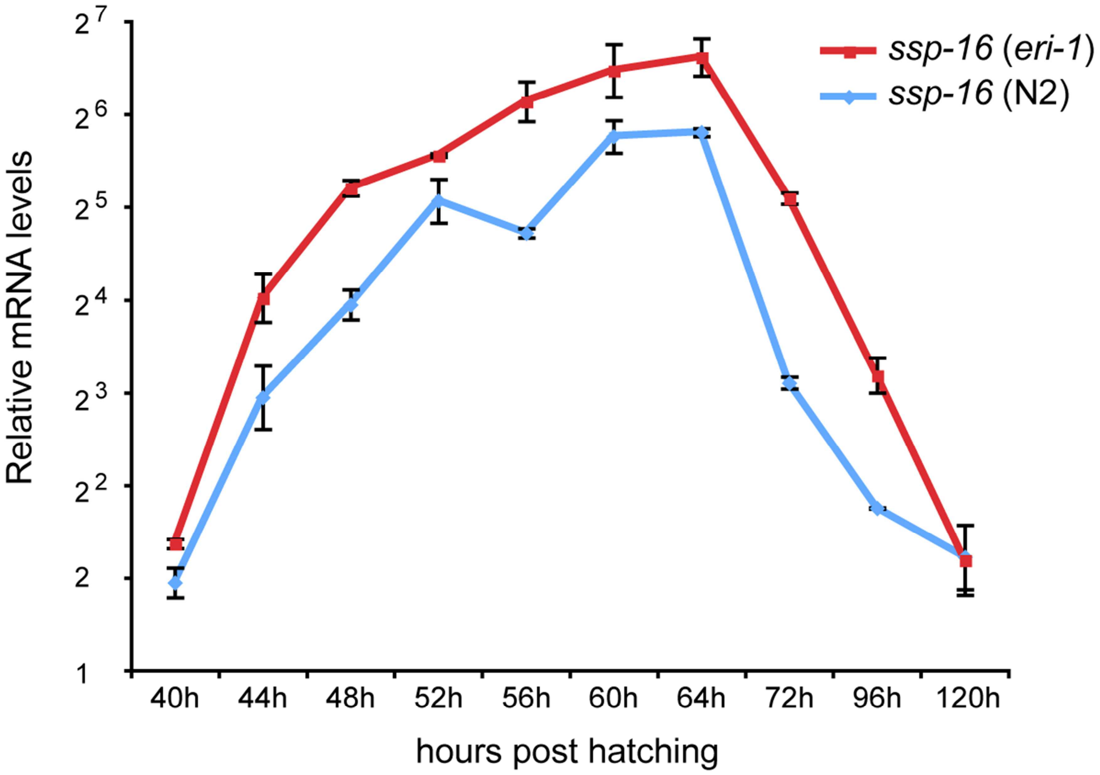


Figure 5.S5: ssp-16 (a target of sperm 26G RNA) is de-repressed starting from spermatogenesis until young adulthood in the eri-1 mutant.

The X-axis represents hours post hatching at 20°C; the Y-axis indicates relative mRNA abundance in log2 scale. Relative mRNA levels were examined by RT-qPCR and normalized to *act-1*.

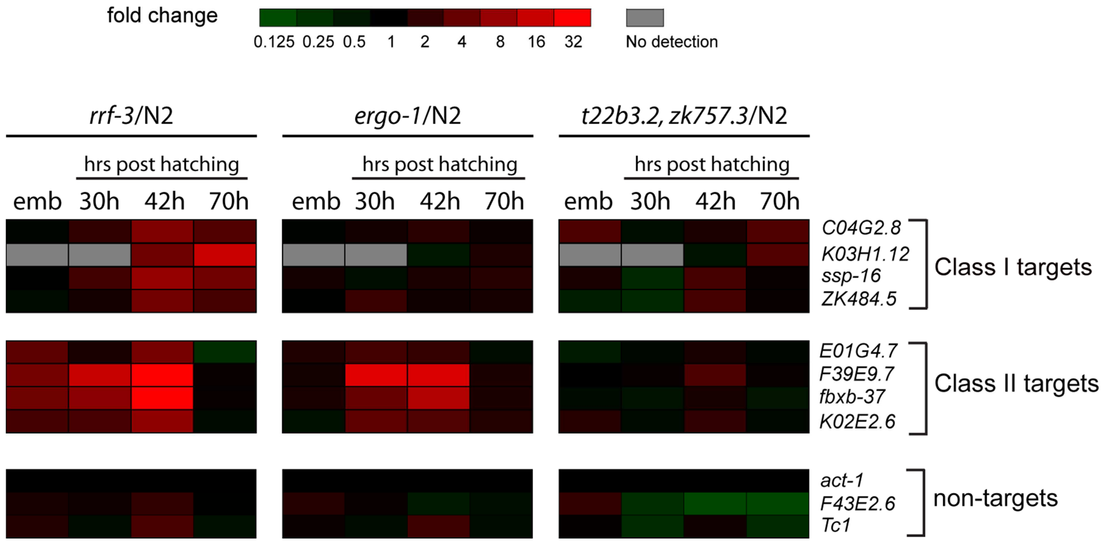


Figure 5.S6: Differential gene expression profiles of 26G RNA targets in N2, rrf-3(pk1426), ergo-1(tm1860), and the t22b3.2(tm1155); zk757.3(tm1184) double mutant.

The transcript levels of 4 targets of class I sperm 26G RNAs, 4 targets of class II oocyte/embryo 26G RNAs, and 3 non-targets were examined. The fold up-regulation was represented according to the red-green color scheme shown (top panel).

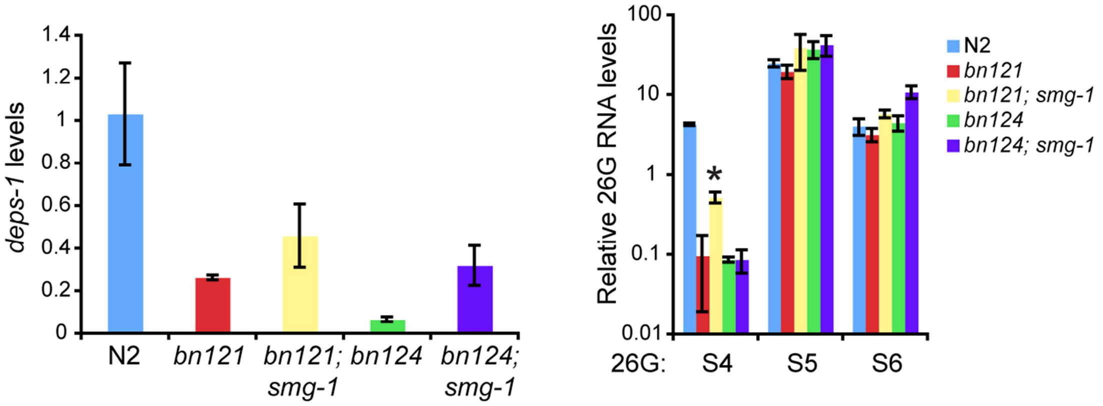


Figure 5.S7: deps-1 mRNA and 26G RNA levels in the deps-1; smg-1 double mutant.

Nonsense *deps-1* mRNA is stabilized by *smg-1*, but still below WT levels. A noticeable increase of 26G-S4 is seen for one of the alleles (*). Error bars represent standard deviation.

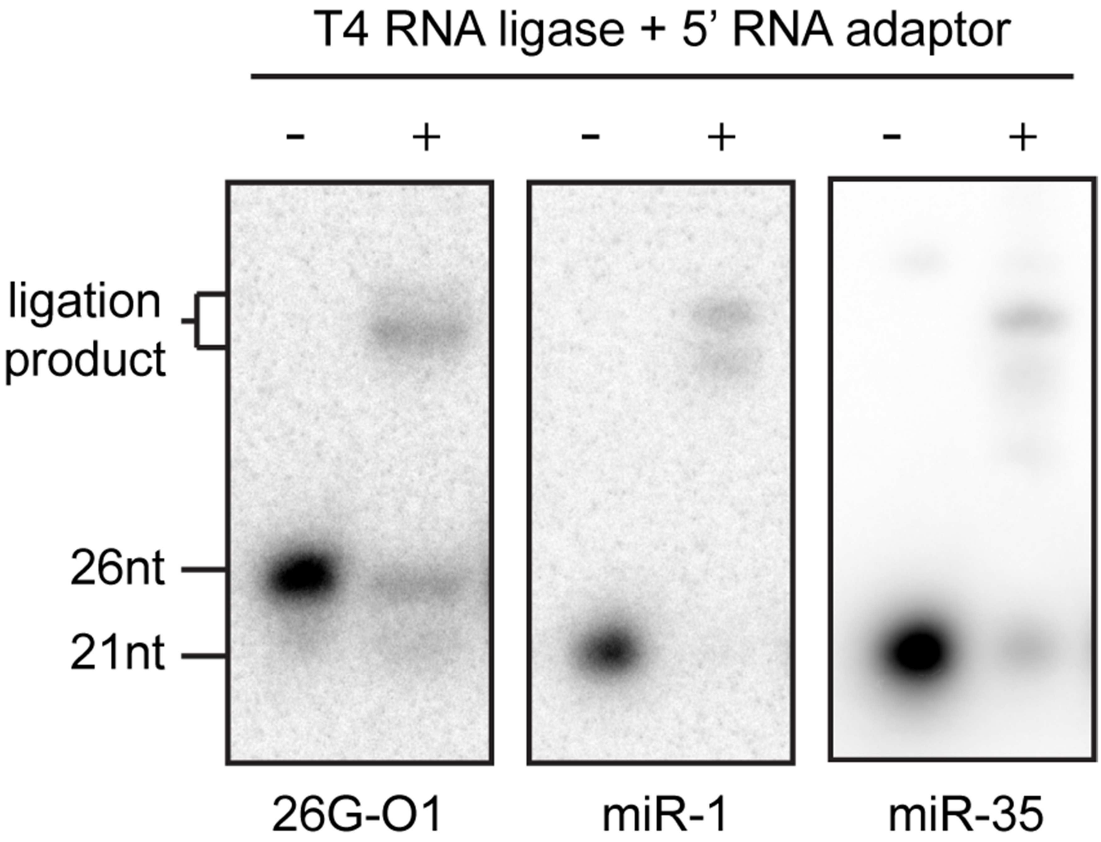


Figure 5.S8: 26G RNAs are suitable substrates for T4 RNA ligase-mediated ligation.
Small RNAs (18-32nt) were isolated by PAGE and ligated to the 5' RNA adaptor used in the small RNA cloning procedure. The ligation product was resolved on 11% Urea-PAGE and subjected to northern blotting analysis. The 26G RNA 26G-O1 shows similar levels of ligation compared to microRNAs miR-1 and miR-35, which are known to possess a 5' monophosphate.

Chapter 6: Conclusion and future studies

6.1: Conclusions

3'UTRome: The main aim of this dissertation was to elucidate the dynamic formation and expression of 3'UTRs in *C. elegans* at a transcriptome-wide scale through the acquisition of high quality 3'UTRome datasets by next generation sequencing methodologies. With the development of the novel "polyA capture" protocol by Ting Han, a graduate student in the Kim Lab, we were able to generate ~2,000,000 full-length or near full-length 3'UTR sequence reads at single nucleotide resolution across the major developmental stages of the worm. These efforts allowed me to annotate both abundant and rare 3'UTR isoforms, validate predicted gene models in *C. elegans* by providing evidence for their 3'UTRs, and, by combining the deep sequencing datasets with conventional cDNA sequence libraries generated by our collaborators, nearly double the number of annotated 3'UTRs in the *C. elegans* transcriptome.

Comparison with 3P-seq: Comparison with an independent parallel study [1] helped me to further derive conditions that can be used to filter for the inherent false priming artifacts which are known to occur in polydT based sequencing mechanisms. Importantly, my analyses showed that these potential false priming amplifications also exist in other 3'end sequencing methods such as 3P-seq, indicating that, at present, computational filtering is the most effective way to derive a robust dataset of genuine 3'UTR sequences.

Alternative polyadenylation: The mapping of 3'UTR ends at single nucleotide resolution highlighted interesting trends in polyadenylation of mRNAs. I found that ~40% of the genes in our dataset exhibited more than one 3'UTR. This made us wonder about the extent of alternative polyadenylation on a global scale. Recent genome-wide studies in other organisms also show extremely high levels of alternative polyadenylation. An estimated 70% of the genes in *Arabidopsis* [2], 52% of the genes in mouse [3], 70% of the genes in yeast [4] and 44% of the genes in humans [4] display alternative polyadenylation. Even protozoans such as trypanosomes display alternative polyadenylation [5]; and since transcriptional regulation is not a major source of gene regulation [6, 7], it suggests that trans-splicing and polyadenylation of the polycistronic genes represent the dominant forms of gene expression control. Recent studies in mammals [3] also highlight a few important characteristics of alternative polyadenylation (APA). Differentiation of stem cells results in a substantial change of the APA profile, resulting in longer 3'UTRs in the differentiated cells. In addition, the majority of the APA events (>13,000) were independent of splicing. APA has also been shown to play a role in protein output where shorter 3'UTRs express higher levels of the protein [8, 9]. Tissue specific APA, along with alternative splicing events, has been shown to increase protein diversity in humans [10]. This could alter protein functions as shown in the case of IgM protein in B cells [11]. Furthermore, APA can also regulate gene expression as shown in plants where APA of the antisense transcripts plays a role in the regulation of its corresponding sense transcripts [12, 13]. All of these studies suggest that polyadenylation of mRNAs is a complex and essential mechanism of gene regulation and alternative polyadenylation is a more pervasive mechanism with effects in a variety of biological processes than previously thought.

alternative PAS motifs: My analysis of PAS sites upstream of the polyA ends indicated that only 39% of our 3'UTRs expressed the canonical AAUAAA PAS, overturning the model of canonical PAS as the predominant signal. I identified 28 new PAS that have similar positional distribution as the AAUAAA site, peaking at 19nt upstream from the cleavage site. In addition, I also saw a significant number of 3'UTRs which did not have any recognizable PAS site. The biogenesis of these 3'UTRs could be from an alternative mechanism of 3'end processing. Our data showed that ~40% of our genes exhibited alternative polyadenylation and many of these sites showed conservation across nematodes. While the single 3'UTRs and the longest 3'UTR of genes favored the canonical PAS, the shorter 3'UTRs favored the alternative PAS motif. These results indicate an inherent sequence-based flexibility in 3' end formation that is pervasive in the post-transcriptional processing of messages.

Polyadenylation in operons: Since ~3,000 genes in *C. elegans* exist in polycistronic operons, we wanted to see how polyadenylation is regulated during trans-splicing. Surprisingly, I saw an increased level of alternative polyadenylation for genes inside operons than those outside, thus linking trans-splicing with alternative polyadenylation. Furthermore, the position of the gene inside the operon also affected its polyadenylation. The average length of the 3'UTR and the number of isoforms per gene progressively decrease as we travel down the operon. Comparing the genes trans-spliced by SL1 to those lying inside an operon, Jean Thierry-Mieg showed a reduced utilization of the AAUAAA signal compared to those outside an operon. The AAUAAA signal was more prevalent in the genes not trans-spliced. This shows an effect of trans-splicing in the usage of PAS sites showing an interaction between 5' splicing and 3' polyadenylation.

Developmental regulation of polyadenylation: Previous studies had shown few cases where 3’UTRs are regulated during development. My comprehensive analysis of the 3’UTR lengths showed a global trend of decreasing 3’UTR length over development. I also identified thousands of 3’UTR isoforms that were specific to individual developmental stages with embryos expressing the largest number of stage-specific isoforms, likely due to the maternal load of transcripts contributing to the diversity of 3’UTRs. Transitions between select developmental stages, namely L1 to the dauer stage, dauer to L3 dauer-exit stage, and from L4 to the adult stage, revealed a switch in particular 3’UTR isoform expression.

Updated miRNA target predictions: Our updated miRNA target predictions performed with the new 3’UTR annotations using the PicTar algorithm [14-16] from Niklaus Rajewsky’s laboratory showed that almost half of the previous predictions should be modified with new 3’UTR annotations. We had hypothesized that alternative polyadenylation was a mechanism to effectively exclude miRNA target sites in a 3’UTR when a shorter isoform lacking the miRNA binding site is expressed. While this is true in a case-by-case basis, we could not arrive at a generalizable set of rules that indicated that the distal regions of the longer 3’UTR isoforms were enriched for miRNA binding sites.

Polyadenylation of histone mRNAs: An important outcome of our studies pertains to the post-transcriptional processing of histone mRNAs. Histone genes, especially the replication dependent histones (H2a, H2b,H3 and H4), were considered to be processed differently than the other mRNAs through a mechanism employing a stem loop binding protein which recognizes a stem loop formed by a palindromic region at their 3’ends [17-20]. Due to this very specific processing mechanism, the prevailing conclusion was that histone transcripts were considered to be not polyadenylated [21]. However, my analysis indicated that most of the histone genes in *C. elegans* were expressed in a polyadenylated form and displayed the same properties of other coding genes, i.e. the use of canonical and variant PAS motifs. Similar results were also seen in mammals where 4.3% of the H2A histone transcripts appeared to be polyadenylated when detected by northern blot analysis [3]. Taken together, these data suggest that both the stem loop-mediated and PAS-mediated 3'end processing occur. There could be many reasons for two mechanisms and whether they occur in serial or as parallel mechanisms remains to be determined. In addition, many of the histones are clustered in tight loci, potentially leading to transcriptional read through past the stem-loop sequences and engage the downstream PAS signal. In such a scenario, the PAS-mediated polyadenylated 3'end formation may act as a by-pass mechanism for the canonical stem-loop processing event.

Alternative polyadenylation in synaptogenesis: My thesis also provides a specific example of where differential 3'UTR isoform expression may play an important biological role: synaptogenesis and neuronal development. We sequenced 3'UTRs from *rpm-1, sydn-1* and *rpm-1;sydn-1* mutants. *rpm-1* and *sydn-1* have been shown to participate in alternate pathways in the regulation of synapse and axon morphogenesis [22]. Mutating individual genes affects the synapse morphology but has minimal effect on locomotion. However, mutating both genes has been to shown to result in synapse and locomotion defects, suggesting a synthetic genetic interaction. Further *sydn-1* has been shown to interact with *pfs-2*, which encodes a member of the polyadenylation machinery. Defects on *sydn-1* can be suppressed by *pfs-2* mutation. These genetic data strongly suggest that defects in seemingly core biological processes such as polyadenylation can have tissue-specific phenotypic outcomes. While the global profile of the 3'UTRs shows no drastic change in polyadenylation in these neuronal mutants, subtle variations were seen in PAS usage when I examined 3'UTRs differentially expressed between N2 and the mutants and between mutants. I identified hundreds of 3'UTRs unique to each library and also identified many new 3'UTRs. I also saw evidence of differential isoform usage between the libraries. The extent to which these molecular phenotypes contribute to synaptogenesis and perhaps even to synaptic plasticity mechanisms remains to be determined.

Small RNAs in the germline: Small RNAs such as microRNAs and endogenous siRNAs play an important role in post-transcriptional gene regulation. In particular, miRNAs have been shown to target the 3'UTR regions of target transcripts to repress translation and/or induce target mRNA degradation. In contrast, endogenous siRNAs largely silence their target mRNAs by the canonical RNA interference mechanism. To identify novel classes of small noncoding RNAs in *C. elegans*, we sequenced the small RNAs from isolated gametes and the embryo and identified and characterized a new class of siRNAs called 26G RNAs [23], which were first reported in a large scale sequencing study in 2006 [24]. These 26G RNAs were germline specific and were absent in the *glp-4* mutant with defective germline. They were a uniform class of small RNAs that were 26nt in length and started with a guanosine nucleotide, hence the name 26G RNA. My bioinformatic analyses further differentiated the 26G RNAs into two non-overlapping subclasses based on their germ cell of origin and the genes that they targeted. Class I 26G RNAs target genes specific to spermatogenesis and were enriched in the sperm and Class II 26G RNAs are maternally derived from the oocyte and were shown by Ting Han to regulate genes throughout filial development. Genetic and biochemical analysis by Ting Han in the Kim Lab identified that they required ERI-1 endonuclease and RRF-3 RNA dependent RNA polymerase for their biogenesis. Further Ting identified that each class associated with a specific effector complex termed RISC (RNA Induced Silencing

Complex). Class I 26G RNAs associated with the AGO-3/AGO-4 RISC while class II 26G RNAs were bound to the ERGO-1 RISC. These endogenous siRNAs targeted both 3'UTRs and coding regions of their target transcripts and loss of 26G RNAs resulted in up regulation of their targets. Similar results were also later reported in other studies [25-27]. Further work identified *mut-16* as another player in the biogenesis [28]. However, what triggers the production of these endogenous siRNAs is currently not known. One intriguing model postulates that the 3' end structure of the target transcripts may provide a nucleating site to which the amplification machinery is recruited for antisense transcription and subsequent production of the 26G RNAs. Because the 3' ends of target transcripts do not possess any discernable primary sequence motif, such a model relies on the existence of a common secondary structure. One line of evidence that supports this model stems from recent studies that determined that the ERI-1 complex, which is essential for 26G RNA biogenesis, also recognizes a secondary structural motif in ribosomal RNAs to initiate their post-transcriptional processing [29]**.** It will be interesting to perform a comprehensive secondary structural analysis on the 3'UTRs of the 26G RNA target transcripts to determine if such a structural motif exists. Because of our comprehensive 3'UTRome assembly of the *C. elegans* transcriptome, these types of future bioinformatics projects are now possible.

Future work: Whole genome studies with the aid of high-throughput analysis methods are showing that alternative polyadenylation is a fundamental, ubiquitous process affecting 40-70% of the genes in an organism. Differential polyadenylation site usage is emerging as an important means of regulating gene expression during normal development as well as contributing to the organism's response to external stimuli. The comprehensive identification, at the transcriptome-level, of how 3'UTR isoform expression changes in the context of increasingly refined temporal and cell-type specific transitions, as well as identifying how 3'UTR isoform expression changes in disease states such as cancer [8] remain a promising areas of future research. The rapid developments in next-generation sequencing technologies will greatly facilitate the speed and depth of these types of transcriptome-wide studies. Another important question is to determine whether small RNAs that bind 3'UTRs have any direct role in influencing the selection of alternative polyadenylation sites. Again, recent high-throughput methods provide the technology to address these questions at ever-increasing depth, to the point where such questions can be answered with single-cell resolution [23, 30].

An important observation seen in Arabidopsis [2], yeast [4], mouse and human [4] 3'UTRomes was the presence of abundant antisense transcripts. Initial studies show 33% of *Arabidopsis* transcripts [2], 60% in yeast [4] and 30% of human transcripts [4] expressed antisense transcripts which could affect sense gene expression positively or negatively. However, when we looked for antisense transcripts in *C. elegans* [31], the percent was not as high as in the other species. However, this result could be biased due to the fact that we didn't sequence the whole transcriptome and only captured polyA ends or due to the pyrosequencing technology used. Hence more work is needed to address antisense transcripts in *C. elegans.* Furthermore, the sense/antisense overlap region had a peak at 20 nucleotides, or, intriguingly, less than the length of a small RNA. We proposed that this short overlap could be to prevent spawning of endogenous siRNAs from the resulting double strand during transcription. The parallel study [1] in *C. elegans*, proposed the sharing of *cis* elements between the sense/antisense transcripts which could be a mechanism for "genome compaction" to maintain the size of the small genome and reduce the need for long intergenic regions. Nevertheless, presence of sense/antisense transcript pairs suggests mechanisms for biogenesis of siRNAs at least in other organisms or sharing of *cis*-elements.

In my thesis, I showed specific examples of genes which exhibit switching of 3'UTRs to developmental cues [31]. Similar events have also been seen during immune response [9], differentiation [32], and cancer [8]. In these studies, differentiation correlates with longer 3'UTR expression while undifferentiated states seem to exhibit shorter 3'UTRs. In these studies, differential 3'UTR expression results in changes in gene expression. Shorter 3'UTRs seem to result in higher protein translation [8, 9] and is more favored in cases where quick turnover of proteins is needed, as in the case of an immune response, and longer 3'UTRs are favored in cases where translational machinery is needed to be shut down in case of stress [31], viral attack or turning off maternal transcripts in embryos[33]. While the biological significance of this 3'UTR switching mechanism remains to be determined, the fact that at least ~560 genes in *C. elegans* display such an expression pattern suggests that 3'UTR switching may represent an important facet of gene regulation during development.

Analyzing the results of our sequencing data, we saw evidence for polyA sites ending inside the coding regions even after accounting for artifacts due to false priming. Such a transcript would result in an mRNA that lacks a proper termination or stop codon and trigger mRNA degradation by nonsense-mediated decay mechanisms.[34-38]. A recent study has shown that the marking of mRNAs for degradation is done through binding of Upf1 to the 3'UTR regions in a length-dependent manner [39, 40]. Taken together, these findings suggest that the depth of sequencing coverage reveals the rare examples of inappropriate 3'UTR formation that are likely to be rapidly cleared by cellular surveillance mechanisms.

From our *C. elegans* data, we identified many islands in the genome where there were no previous gene annotations but had clear evidence of polyadenylation. We annotated ~1000 new genes based on this data. This number may not be saturated yet and many other new gene models may yet be discovered in other organisms. Furthermore, many non-coding RNAs are transcribed by RNA polymerase II [41-43] and these are also known to be polyadenylated. Using the polyA capture method will allow the comprehensive identification of these polyadenylated non-coding RNAs.

In summary, the research described in this thesis represents the initial salvo in the emerging area of gene regulation mediated by alternative 3'UTR isoform expression and by the small noncoding RNAs that interact with them. By deciphering the basic mechanisms of post-transcriptional regulation in the context of 3'UTR formation, we will then be able to determine if such processes are dys-regulated in disease states such as neuronal degeneration and cancer.

6.2: Reference

1. Jan CH, Friedman RC, Ruby JG, Bartel DP: **Formation, regulation and evolution of Caenorhabditis elegans 3′UTRs**. *Nature* 2010, **469**(7328):97-101.
2. Wu X, Liu M, Downie B, Liang C, Ji G, Li QQ, Hunt AG: **Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation**. *Proceedings of the National Academy of Sciences* 2011.
3. Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y: **Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq**. *Rna* 2011, **17**(4):761-772.
4. Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM: **Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation**. *Cell* 2010, **143**(6):1018-1029.
5. Jager AV, De Gaudenzi JG, Cassola A, D'Orso I, Frasch AC: **Inaugural Article: mRNA maturation by two-step trans-splicing/polyadenylation processing in trypanosomes**. *Proceedings of the National Academy of Sciences* 2007, **104**(7):2035-2042.
6. Campbell DA, Thomas S, Sturm NR: **Transcription in kinetoplastid protozoa: why be normal?** *Microbes Infect* 2003, **5**(13):1231-1240.
7. Palenchar JB, Bellofatto V: **Gene transcription in trypanosomes**. *Mol Biochem Parasitol* 2006, **146**(2):135-141.
8. Mayr C, Bartel DP: **Widespread Shortening of 3′UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells**. *Cell* 2009, **138**(4):673-684.
9. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB: **Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites**. *Science* 2008, **320**(5883):1643-1647.
10. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470-476.
11. Takagaki Y, Seipelt RL, Peterson ML, Manley JL: **The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation**. *Cell* 1996, **87**(5):941-952.
12. Hornyik C, Duc C, Rataj K, Terzi Lionel C, Simpson Gordon G: **Alternative polyadenylation of antisense RNAs and flowering time control**. *Biochemical Society Transactions* 2010, **38**(4):1077.
13. Liu F, Marquardt S, Lister C, Swiezewski S, Dean C: **Targeted 3' processing of antisense transcripts triggers Arabidopsis FLC chromatin silencing**. *Science* 2010, **327**(5961):94-97.
14. Grun D, Wang YL, Langenberger D, Gunsalus KC, Rajewsky N: **microRNA target predictions across seven Drosophila species and comparison to mammalian targets**. *PLoS Comput Biol* 2005, **1**(1):e13.
15. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M *et al*: **Combinatorial microRNA target predictions**. *Nat Genet* 2005, **37**(5):495-500.
16. Lall S, Grun D, Krek A, Chen K, Wang YL, Dewey CN, Sood P, Colombo T, Bray N, Macmenamin P *et al*: **A genome-wide map of conserved microRNA targets in C. elegans**. *Curr Biol* 2006, **16**(5):460-471.

17. Dominski Z, Marzluff WF: **Formation of the 3′ end of histone mRNA: Getting closer to the end**. *Gene* 2007, **396**(2):373-390.

18. Pettitt J, Crombie C, Schumperli D, Muller B: **The Caenorhabditis elegans histone hairpin-binding protein is required for core histone gene expression and is essential for embryonic and postembryonic cell division**. *J Cell Sci* 2002, **115**(Pt 4):857-866.

19. Williams AS, Marzluff WF: **The sequence of the stem and flanking sequences at the 3' end of histone mRNA are critical determinants for the binding of the stem-loop binding protein**. *Nucleic Acids Res* 1995, **23**(4):654-662.

20. Zhao J, Hyman L, Moore C: **Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis**. *Microbiol Mol Biol Rev* 1999, **63**(2):405-445.

21. Marzluff WF, Wagner EJ, Duronio RJ: **Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail**. *Nat Rev Genet* 2008, **9**(11):843-854.

22. Van Epps H, Dai Y, Qi Y, Goncharov A, Jin Y: **Nuclear pre-mRNA 3'-end processing regulates synapse and axon development in C. elegans**. *Development* 2010, **137**(13):2237-2250.

23. Han T, Manoharan AP, Harkins TT, Bouffard P, Fitzpatrick C, Chu DS, Thierry-Mieg D, Thierry-Mieg J, Kim JK: **26G endo-siRNAs regulate spermatogenic and zygotic gene expression in Caenorhabditis elegans**. *Proceedings of the National Academy of Sciences* 2009, **106**(44):18674-18679.

24. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP: **Large-Scale Sequencing Reveals 21U-RNAs and Additional MicroRNAs and Endogenous siRNAs in C. elegans**. *Cell* 2006, **127**(6):1193-1207.

25. Gent JI, Schvarzstein M, Villeneuve AM, Gu SG, Jantsch V, Fire AZ, Baudrimont A: **A Caenorhabditis elegans RNA-Directed RNA Polymerase in Sperm Development and Endogenous RNA Interference**. *Genetics* 2009, **183**(4):1297-1314.

26. Pavelec DM, Lachowiec J, Duchaine TF, Smith HE, Kennedy S: **Requirement for the ERI/DICER complex in endogenous RNA interference and sperm development in Caenorhabditis elegans**. *Genetics* 2009, **183**(4):1283-1295.

27. Conine CC, Batista PJ, Gu W, Claycomb JM, Chaves DA, Shirayama M, Mello CC: **Argonautes ALG-3 and ALG-4 are required for spermatogenesis-specific 26G-RNAs and thermotolerant sperm in Caenorhabditis elegans**. *Proc Natl Acad Sci U S A* 2010, **107**(8):3588-3593.

28. Zhang C, Montgomery TA, Gabel HW, Fischer SE, Phillips CM, Fahlgren N, Sullivan CM, Carrington JC, Ruvkun G: **mut-16 and other mutator class genes modulate 22G and 26G siRNA pathways in Caenorhabditis elegans**. *Proc Natl Acad Sci U S A* 2011, **108**(4):1201-1208.

29. Gabel HW, Ruvkun G: **The exonuclease ERI-1 has a conserved dual role in 5.8S rRNA processing and RNAi**. *Nature Structural &#38; Molecular Biology* 2008, **15**(5):531-533.

30. Stoeckius M, Maaskola J, Colombo T, Rahn HP, Friedlander MR, Li N, Chen W, Piano F, Rajewsky N: **Large-scale sorting of C. elegans embryos reveals the dynamics of small RNA expression**. *Nat Methods* 2009, **6**(10):745-751.

31. Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V *et al*: **The Landscape of C. elegans 3'UTRs**. *Science* 2010, **329**(5990):432-435.

32. Ji Z, Tian B: **Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types**. *PLoS One* 2009, **4**(12):e8419.

33. Ji Z, Lee JY, Pan Z, Jiang B, Tian B: **Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development**. *Proc Natl Acad Sci U S A* 2009, **106**(17):7028-7033.

34. Frischmeyer PA, van Hoof A, O'Donnell K, Guerrerio AL, Parker R, Dietz HC: **An mRNA surveillance mechanism that eliminates transcripts lacking termination codons**. *Science* 2002, **295**(5563):2258-2261.

35. van Hoof A, Frischmeyer PA, Dietz HC, Parker R: **Exosome-mediated recognition and degradation of mRNAs lacking a termination codon**. *Science* 2002, **295**(5563):2262-2264.

36. Ito-Harashima S, Kuroha K, Tatematsu T, Inada T: **Translation of the poly(A) tail plays crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization by proteasome in yeast**. *Genes Dev* 2007, **21**(5):519-524.

37. Atkinson GC, Baldauf SL, Hauryliuk V: **Evolution of nonstop, no-go and nonsense-mediated mRNA decay and their termination factor-derived components**. *BMC Evol Biol* 2008, **8**:290.

38. Isken O, Maquat LE: **Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function**. *Genes Dev* 2007, **21**(15):1833-1856.

39. Hogg JR: **This message was inspected by Upf1: 3'UTR length sensing in mRNA quality control**. *Cell Cycle* 2011, **10**(3):372-373.

40. Hogg JR, Goff SP: **Upf1 senses 3'UTR length to potentiate mRNA decay**. *Cell* 2010, **143**(3):379-389.

41. Tupy JL, Bailey AM, Dailey G, Evans-Holm M, Siebel CW, Misra S, Celniker SE, Rubin GM: **Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster**. *Proc Natl Acad Sci U S A* 2005, **102**(15):5495-5500.

42. Cai X, Hagedorn CH, Cullen BR: **Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs**. *Rna* 2004, **10**(12):1957-1966.

43. Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN: **MicroRNA genes are transcribed by RNA polymerase II**. *EMBO J* 2004, **23**(20):4051-4060.