The *C. elegans* 3'UTRome V2: an updated genomic resource to study 3'UTR biology

Steber HS[1, 2], Gallante C[3], O'Brien S[2], Chiu P.-L[4], Mangone M*[1, 2].

[1]Molecular and Cellular Biology Graduate Program, School of Life Sciences 427 East Tyler Mall Tempe, AZ 85287 4501.

[2]Virginia G. Piper Center for Personalized Diagnostics, The Biodesign Institute at Arizona State University, 1001 S McAllister Ave, Tempe, AZ, USA

[3]Barrett, The Honors College, Arizona State University, 751 E Lemon Mall, Tempe, AZ 85281

[4]Center for Applied Structural Discovery, The Biodesign Institute at Arizona State University, 1001 S McAllister Ave, Tempe, AZ 85287, USA

* To whom correspondence should be addressed. Tel: +1(480) 965-7957; Fax: +1(480) 965-3051; Email: mangone@asu.edu

Present Address: Marco Mangone, Arizona State University, Biodesign Institute Building A, 1001 S McAllister Ave, Tempe, AZ 85281 USA

Keywords: *C. elegans*, 3'Untranslated Regions, Alternative Polyadenylation, Cleavage and Polyadenylation Complex, miRNAs, 3'UTRome

Running Title: 3'UTRome in *C. elegans*

ABSTRACT

3'Untranslated Regions (3'UTRs) of mRNAs emerged as central regulators of cellular function as they contain important but poorly-characterized *cis*-regulatory elements targeted by a multitude of regulatory factors. The soil nematode *C. elegans* is an ideal model to study these interactions since it possesses a well-defined 3’UTRome. In order to improve its annotation, we have used a genomics approach to download raw transcriptome data for ~1,500 transcriptome datasets corresponding to the entire collection of *C. elegans* trancriptomes from 2015 to 2018 from the Sequence Read Archive at the NCBI. We then extracted and mapped high-quality 3’UTR data at ultra-deep coverage. Here we describe and release to the Community the updated version of the worm 3’UTRome, which we named 3’UTRome v2. This resource contains high-quality 3’UTR data mapped at single base ultra-resolution for 23,159 3’UTR isoforms variants corresponding to 14,808 protein-coding genes and is updated to the latest release of WormBase. We used this dataset to study and probe principles of RNA cleavage and polyadenylation in *C. elegans*. The worm 3’UTRome v2 represents the most comprehensive and high-resolution 3’UTR dataset available in *C. elegans*, and provides a novel resource to investigate the mRNA cleavage and polyadenylation reaction, 3’UTR biology and miRNA targeting in a living organism.

BACKGROUND

3'Untranslated Regions (3'UTRs) are the portions of mRNA located between the end of the coding sequence and the polyA tail of RNA polymerase II-transcribed genes. They contain *cis*-regulatory elements targeted by miRNAs and RNA binding proteins and modulate mRNA stability, localization, and overall translational efficiency (Bartel 2018). Because multiple 3'UTR isoforms of a particular mRNA can exist, differential regulation of 3'UTRs has been implicated in numerous diseases, and its discriminative processing influences development and metabolism (Mayr and Bartel 2009; Zhu et al. 2018). 3'UTRs are processed to full maturity through cleavage of the nascent mRNA and subsequent polyA tail addition to its 3' end by the nuclear polyA polymerase enzyme (PABPN1) (Kuhn and Wahle 2004). The mRNA cleavage step is a dynamic regulatory process directly involved in the control of gene expression in Eukaryotes. The reaction depends on the presence of a series of sequence elements located within the end of the 3'UTRs. The most well-characterized sequence is the PolyA Signal (PAS) element, a hexameric motif located at ~19nt from the polyadenylation site in the 3'UTR of mature mRNAs. In metazoans, the PAS element is commonly 'AAUAAA', which accounts for more than half of all 3' end processing in eukaryotes (Mangone et al. 2010; Tian and Graber 2012) although alternative forms of the PAS elements exist (Sheets et al. 1990; Mangone et al. 2010). Previous studies have shown that single base substitutions in this sequence reduce the effectiveness of the cleavage and the polyadenylation of the mRNA transcript (Sheets et al. 1990; Chen et al. 1995). However, this canonical sequence is necessary and sufficient for efficient 3’ end polyadenylation *in vitro* (Clerici et al. 2018; Sun et al. 2018). A less defined ‘GU rich’ element is also known to be present downstream of the cleavage site to facilitate the cleavage and polyadenylation steps (Chen et al. 1995). Recently, studies in human cells identified an additional upstream ‘UGUA’ sequence that is not required for the cleavage process, but acts as a cleavage enhancer in the context of Alternative Polyadenylation (APA) (Zhu et al. 2018).

APA is a poorly understood mRNA maturation step that produces mRNAs with different 3’UTR lengths due to the presence of multiple PAS elements within the same 3’UTR. The usage of the most upstream element, termed the proximal PAS element, leads to the formation of shorter 3’UTR isoforms while the use of the distal PAS element results in a longer isoform. Importantly, these changes in size may include or exclude regions to which regulatory molecules such as microRNAs (miRNAs) and RNA-binding proteins (RBPs) can bind, substantially impacting gene expression (Matlin et al. 2005; Bartel 2009). While its function in eukaryotes is still not fully understood, a recent study revealed that APA may occur in a tissue-specific manner and, at least in the soil nematode *C. elegans*, is used in specific cellular contexts to evade miRNA-based regulatory networks in a tissue-specific manner (Blazie et al. 2015; Blazie et al. 2017).

The length of the 3’UTRs is defined during the cleavage and polyadenylation reaction, which is still poorly characterized in metazoans. Although it involves a multitude of proteins and is considered to be very dynamic, the order in which this process is executed and the role of each member of the complex is still not fully understood.

In humans, the Cleavage and Polyadenylation Complex (CPC) is composed of at least 17 members (**Figure 1A**) which immunoprecipitate into at least four large sub-complexes: the Cleavage and Polyadenylation Specificity Factor (CPSF), the Cleavage Stimulation Specificity Factor (CstF), the Cleavage Factor Im (CFIm) and the Cleavage Factor IIm (CFIIIm) sub-complexes (**Figure 1A**). CPSF forms the minimal core complex necessary and sufficient to recognize and bind the PAS element of the nascent mRNA *in vitro* (Tian and Manley 2017) (**Figure 1A**). In humans, the CPSF sub-complex is composed of CPSF160 (Clerici et al. 2017; Sun et al. 2018), CPSF100 (Mandel et al. 2006), CPSF73 (Mandel et al. 2006), CPSF30 (Clerici et al. 2017; Sun et al. 2018), Fip1 (Kaufmann et al. 2004) and Wdr33 (Clerici et al. 2017; Sun et al. 2018). Initial experiments assigned CPSF160 with the role of binding the PAS element, but it is now clear that Wdr33 and CPSF30 are the proteins that instead contact the PAS directly. CPSF160 has a scaffolding role in this process and keeps this sub-complex structured (Chan et al. 2014). The interaction between members of the CPSF core complex (Wdr33, CPSF30, and CPSF160) and the PAS element was recently revealed using single-particle cryo-EM (Clerici et al. 2017; Sun et al. 2018), showing a unique conformation where the PAS element twists to form an s-shaped structure with a non-canonical pairing between the U3 and the A6 in the PAS element (Sun et al. 2018).

CPSF73 is the endonuclease that performs the cleavage of the nascent mRNAs (Ryan et al. 2004; Mandel et al. 2006) (**Figure 1A**). CPSF73 possesses a Metallo-β-lactamase domain and a β-CASP domain used to recognize and cleave nucleic acids. Purified recombinant CPSF73 retains RNA endonuclease activity, and mutations that disrupt the zinc binding in the active site of the enzyme abolish this activity (Mandel et al.

2006), suggesting that this protein's role is to perform mRNA cleavage. Importantly, CPSF73 is also required in the cleavage of pre-histone mRNAs and is recruited on their cleavage site by the U7 SNP (Yang et al. 2009).

Fip1 is another member of the CPSF sub-complex. Fip1 interacts with PABPN1, which is the enzyme that performs the polyadenylation reaction on the cleavage site. Fip1 preferentially binds U-rich elements in the nascent mRNA and stabilizes the cleavage complex using its arginine-rich RNA-binding domain (Kaufmann et al. 2004). Together with CPSF160 and PABPN1, Fip1 forms a ternary complex *in vitro* (Kaufmann et al. 2004) capable of inhibiting endogenous PABPN1activity (Zhelkovsky et al. 1998; Helmling et al. 2001), suggesting a bridging role for this protein in the complex.

The CstF sub-complex is the second most well-characterized sub-complex involved in the cleavage and polyadenylation reaction (**Figure 1A**). CstF binds to GU rich elements located downstream of the cleavage site in the nascent mRNA and directly contacts the CPSF sub-complex using its conserved HAT-C domain (Bai et al. 2007; Yang et al. 2018) (**Figure 1A**). The CstF sub-complex is a dimer of heterotrimers composed of CstF77, CstF64 and CstF50 (Yang et al. 2018). CstF77 holds the complex together through its Pro-rich domain located on its C terminal region (Takagaki and Manley 2000) **(Figure 1A)**. CstF64 recognizes GU rich sequences through its N-terminal RRM domain (Perez Canadillas and Varani 2003; Yang et al. 2018) and interacts with the scaffolding protein Symplekin and CstF77 using its N-terminal hinge domain (**Figure 1A**) (Takagaki and Manley 2000).

The CFIm and CFIIm sub-complexes are unfortunately less characterized (**Figure 1A**). The CFIm sub-complex is composed of the CFIm68, CFIm59 and CFIm25 subunits, and it was recently shown to contribute to APA by influencing PAS selection (Martin et al. 2012; Hwang et al. 2016). CFIm25 binds a ‘UGUA’ RNA element upstream of the cleavage site and contributes to 3’processing by recruiting CFIm59 and CFIm68 (Yang et al. 2010; Yang et al. 2011; Zhu et al. 2018).

The Cleavage Factor IIm sub-complex is composed of two factors named Pcf11 and hClp1 (Schafer et al. 2018). Pcf11 binds RNA unspecifically through two zinc fingers in its C-terminal region and stimulates the RNA 5’ kinase activity of hClp1, which is not required for the cleavage reaction (Schafer et al. 2018). It has been suggested that hClp1 binds CPSF, although the exact interaction has not been determined (de Vries et al. 2000).

Despite the importance of this complex, the CPC remains poorly characterized in most species, including humans, and most of the research in this field is performed *in vitro*.

The round nematode *C. elegans* represents an attractive, novel system to study the cleavage and polyadenylation process *in vivo*. Most of the CPC is conserved between humans and nematodes, including known functional domains and protein interactions (**Figure 1B and Supplemental Figure S1**). *C. elegans* possess the most well-annotated 3’UTRome available so far in metazoans, with mapped 3’UTR boundaries for ~26,000 distinct *C. elegans* protein-coding genes (Mangone et al. 2010; Jan et al. 2011).

The *C. elegans* 3’UTRome was originally developed in 2011 within the modENCODE project (Mangone et al. 2008; Gerstein et al. 2010; Mangone et al. 2010) and represented a milestone in 3’UTR biology since it allowed the Community to study and identify important regulatory elements such as miRNAs and RBPs targets with great precision. A second 3'UTRome was later published using a different mapping pipeline (Jan et al. 2011), confirming most of the previous data such as isoforms numbers, PAS usage, etc. Other datasets were made available later, mostly focusing on tissue-specific 3'UTRs and alternative polyadenylation (Haenni et al. 2012; Blazie et al. 2015; Blazie et al. 2017; Chen et al. 2017; Diag et al. 2018; West et al. 2018).

Although refined and based on several available datasets, only a subset of *C. elegans* 3'UTRs in protein-coding genes are sufficiently annotated today, and the existing mapping tools do not yet reach the single-base resolution necessary to execute downstream analysis and study the cleavage and polyadenylation process in detail. Most of these 3'UTR datasets were developed using a gene model now considered obsolete (WS190), and the 3'UTR coordinates often do not match the new gene coordinates.

To address these and other issues, we developed a novel pipeline to bioinformatically extract 3'UTR data from the entire collection of *C. elegans* transcriptome datasets stored in the public repository SRA trace archive from 2015 to 2018. This blind approach produced a new saturated dataset we named 3'UTRome v2. This updated 3'UTRome contains 3'UTR data for 23,159 3'UTR isoforms variants corresponding to 14,808 protein-coding genes and is available to the Community as an additional gBrowse track in the *C. elegans* database WormBase (www.WormBase.org) (Stein et al. 2001) and in the 3'UTR-centric database 3'UTRome (www.UTRome.org) (Mangone et al. 2008; Mangone et al. 2010).

We have also used this dataset to study the PAS sequence requirement and the cleavage location of the CPC *in vivo* using transgenic *C. elegans* animals. We found that the canonical CPC can in principle bind different PAS sequences and that elements downstream of the PAS site can in turn influence the location of the cleavage.

RESULTS

Functional elements of the human cleavage and polyadenylation complex are conserved in nematodes.

To initially gain structural and functional information for the *C. elegans* CPC, we downloaded the protein sequences of the orthologs of the *C. elegans* CPC and aligned them to their human counterparts (**Figure 1B and Supplemental Figure S1**). Based on sequence similarity, *C. elegans* possess orthologs to all the known members of the human CPC, with many peaks of conservation interspersed in the subunits within known interaction domains. The amino acids that make direct contact with PAS elements are also conserved in *C. elegans*; 11 out of the 12 amino acids that form hydrogen bonds and salt bridges with the PAS element (Clerici et al. 2017) are present in both the CPSF30 and WDR33 worm orthologs *cpsf-4* and *pfs-2* (V67^{CPSF30} with V81^{cpsf-4}; K69^{CPSF30} with K83^{cpsf-4}; R73^{CPSF30} with R87^{cpsf-4}; E95^{CPSF30} with E109^{cpsf-4}; K77^{CPSF30} with K91^{cpsf-4}; S106^{CPSF30} with S120^{cpsf-4}; N107 ^{CPSF30} with N121^{cpsf-4}; R54^{WRD33} with R80^{pfs-2}; R47^{WRD33} with R71^{pfs-2}; R49^{WRD33} with R73^{pfs-2}) (**Figure 1B and Supplemental Figure S1**). The only exception is Y97^{CPSF30}, which is substituted with a Phenylalanine residue in the worm ortholog. In addition, 9 out of the 10 amino acids in CPSF30 and WDR33 that form the π-π stacking and hydrophobic interactions with the AAUAAA RNA element (Clerici et al. 2017) are also conserved in the CPSF30 and WDR33 worm orthologs *cpsf-4* and *pfs-2* (A1:K69$^{\text{CPSF30}}$ with K83$^{\text{cpsf-4}}$ and F84 $^{\text{CPSF30}}$ with F98$^{\text{cpsf-4}}$; A2: H70$^{\text{CPSF30}}$ with H84$^{\text{cpsf-4}}$; U3: I156$^{\text{WDR33}}$ with I181$^{\text{pfs-2}}$; A4: F112$^{\text{CPSF30}}$ with F126$^{\text{cpsf-4}}$ and F98$^{\text{CPSF30}}$ with F112$^{\text{cpsf-4}}$; A5: F98$^{\text{CPSF30}}$ with F112$^{\text{cpsf-4}}$; A6: F153$^{\text{WDR33}}$ with F178$^{\text{pfs-2}}$) (**Figure 1B and Supplemental Figure S1**). The only exception is a F43$^{\text{WDR33}}$ substitution to a Glycine residue that interacts with A6 in the worm ortholog.

CPSF73, the endonuclease that performs the cleavage reaction, has a *C. elegans* ortholog named *cpsf-3*. Both genes are conserved with an overall 57.61% identity that increases to 69.52% in the β-lactamase domain, which is the region required to perform the cleavage reaction (**Figure 1B and Supplemental Figure S1**). Specifically, all eight amino acids shown previously to form the zinc binding site required for the cleavage reaction (Mandel et al. 2006) are also conserved (D75$^{\text{CPSF73}}$ with D74$^{\text{cpsf-3}}$; H76$^{\text{CPSF73}}$ in H75$^{\text{cpsf-3}}$; H73$^{\text{CPSF73}}$ in D72$^{\text{cpsf-3}}$; H396$^{\text{CPSF73}}$ with H397$^{\text{cpsf-3}}$; H158$^{\text{CPSF73}}$ with H159$^{\text{cpsf-3}}$; D179$^{\text{CPSF73}}$ with D180$^{\text{cpsf-3}}$; H418$^{\text{CPSF73}}$ with H419$^{\text{cpsf-3}}$; E204$^{\text{CPSF73}}$ with E205$^{\text{cpsf-3}}$) (**Figure 1B and Supplemental Figure S1**). This overall similarity is also observed in most of the other members of the *bona fide C. elegans* CPC complex (**Supplemental Figure S1**), suggesting similar structure and function.

In addition, when subjected to RNAi analysis, each of the *C. elegans* CPC members produced a similar strong embryonic lethal phenotype, suggesting that each of these genes may act as a complex and is required for viability (**Figure 1C and Supplemental Figure S2)**.

An updated 3'end mapping strategy

Next, we used a blind genomic approach to improve the current version of the 3'UTRome. We refined a 3'UTR mapping pipeline we previously developed and used in the past (Blazie et al. 2015; Blazie et al. 2017). This approach uses raw transcriptome data as input material to identify and precisely map high-quality 3'UTR end clusters (**Figure 2 and Supplemental Figure S3**).

We wanted to obtain the most accurate, saturated and tissue-independent dataset possible. To achieve this goal we downloaded the entire collection from 2015 to 2018 of transcriptome datasets stored in the Sequence Read Archive (SRA) (**Supplemental Figure S1**), and processed them through our 3'UTR mapping pipeline. We reasoned that this blind approach would lead to the identification of as many 3'UTR isoforms as possible in an unbiased manner since these downloaded transcriptomes have been sequenced using both *wild-type* and mutant strains subjected to many different environmental conditions and covering all developmental stages with many replicates.

We downloaded a total of 1,094 *C. elegans* transcriptome datasets (~2TB of total raw data)(**Supplemental Table S1**). Most of these datasets have also been used in the past to map polyadenylation sites in *C. elegans*. Our 3'UTR mapping approach extracted from these datasets ~5M unique, high-quality polyA reads, which we then used for cluster preparation and mapping (see Methods). We implemented very restrictive parameters for cluster identification and 3'UTR end mapping to limit the unavoidable noise produced by using such diverse datasets as data sources (**Supplemental Figure S3**). Our approach led us to map 3'UTR clusters with ultra-deep coverage of several magnitudes (average cluster coverage ~220X) (**Figure 2A**), and the identification of 23,159 3'UTR isoforms corresponding to 14,808 protein-coding genes. When compared to the previous 3'UTRome v1 dataset (Mangone et al. 2010), we obtained 3'UTR information for an additional 4,638 new protein-coding genes (6,218 3'UTR isoforms) (73% of all protein-coding genes included in the WS250 release) (**Figure 2B-C**).

The *C. elegans* 3'UTRome v2

Our approach produced high-quality 3'UTR data for 14,808 *C. elegans* protein-coding genes (**Figure 2B**). The most abundant nucleotide in *C. elegans* 3'UTRs is a Uridine, which accounts for 40% of all nucleotides in 3'UTRs (**Figure 3A Top Left Panel**). Adenosine nucleotides are the second most represented nucleotide class with ~30% of incidence (**Figure 3A Top Left Panel**). Alternative polyadenylation is common but occurs at a lesser extent than what was previously published (Mangone et al. 2010; Jan et al. 2011). The majority of protein-coding genes (58%) are transcribed with only one 3'UTR isoform (**Figure 3A Bottom Left Panel**) in contrast with ~61% as it was reported in the past (Mangone et al. 2010; Jan et al. 2011). Genes with two 3'UTR isoforms are notably increased in occurrence when compared with past studies (32% vs 25%), while the occurrence of genes with three or more 3'UTRs is comparable with what was previously found (**Figure 3A Bottom Left Panel**) (Mangone et al. 2010; Jan et al. 2011).

Interestingly, in the case of genes with multiple 3'UTRs, the canonical AAUAAA PAS site is greater than two times more abundant in longer 3'UTR isoforms than in shorter

3'UTR isoforms, suggesting that the preparation of shorter 3'UTR isoforms may be subject to regulation (**Supplemental Figure S4**).

The average 3'UTR length in the 3'UTRome v2 is 215nt (**Figure 3A Top Right Panel**), and the occurrence of more 3'UTR isoforms per gene correlates with an overall extension in length (**Figure 3A Top Right Panel**). We also note a slight correlation between 3'UTR length and PAS element usage, with longer 3'UTRs more frequently containing variant PAS elements (**Figure 3A Bottom Right Panel**). The most common PAS element in *C. elegans* protein-coding genes is consistently the hexamer 'AAUAAA', which is present in 58.4% of all the 3'UTRs mapped in this study (**Figure 3B Left Panel**). This element is ~20% more abundant than what was previously identified in past studies (Mangone et al. 2010; Jan et al. 2011). The PAS sequence is located ~ 18nt from the cleavage site (**Figure 3B Right Panel**), and a *buffer region* of ~12nt is present between the PAS element and the cleavage site (**Figure 3C**). The cleavage site occurs almost invariably at an Adenosine nucleotide, which is often preceded by a Uridine nucleotide (**Figure 3C**).

An RRYRRR motif in 3'UTRs with variant PAS elements

We could not detect any enrichment for the UGUA motif near the cleavage site (**Supplemental Figure S5**), and perhaps this element is either not used in *C. elegans* or the CFIm complex may recognize a variant motif not yet identified in this organism. Importantly, when we aligned the 3' ends of 3'UTRs which contain variant PAS elements, we noticed an enrichment of an 'RRYRRR' motif where the canonical PAS element is generally located; this suggests that in *C. elegans,* an 'RRYRRR' element could be used instead when the AAUAAA hexamer is absent (**Figure 4A**).

To better understand the molecular details of the interaction between CPSF and the PAS element, we built a pseudo-atomic homology model of the worm CPSF core complex containing *cpsf-1* (CPSF160), *pfs-2* (Wdr33), and *cpsf-4* (CPSF30) (**Figure 4B and Supplemental Figure S6)**. Most of this model can be superimposed to the cryo-EM structure of the human CPSF core complex (**Figure 4B and Supplemental Figure S6**).

The nucleotide-binding pocket can also be fitted into our homology model, which may implicate a conserved binding region in the *C. elegans* complex (**Figure 4B Right Panel)**. From the structural details of the human CPSF core complex, the interactions between the RNA nucleotides and CPSF30 or WDR33 are not specific. The nucleotide binding is mainly established by the π-π ring stacking force between the nucleotide bases and the residues with aromatic side chains, such as phenylalanine and tyrosine (**Supplemental Figure S6**). Also, the binding pockets of the Adenine base do not seem to have a steric hindrance for Guanine base to bind. It is similar for the Uridine base to the Cytosine base (**Supplemental Figure S6**). Thus, at least in *C. elegans*, the selectivity of the nucleotide binding in *C. elegans* may be only at a level to the nucleotide bases, that is, Pyrimidines or Purines.

An enrichment of Adenosine nucleotide at the cleavage site

We were intrigued by the almost invariable presence of Adenosine nucleotides near the cleavage site. This enrichment becomes more evident when we sort 3'UTRs with canonical PAS elements by the length of their respective *buffer regions* (**Figure 5A**). In the case of the largest group with a *buffer region* of 12-13nt, more than 2,000 3'UTRs terminate with ~70% occurrence of Adenosine nucleotides at the cleavage site. Since we bioinformatically removed the polyA sequences from the sequencing reads during our cluster preparation step, we do not have direct evidence that this last Adenosine nucleotide is indeed present in the mature transcripts and used as a template for the polymerization of the polyA tail, or that it is attached by PABPN1 during the polymerization of the polyA tail. Of note, the high abundance of this nucleotide at the cleavage site suggests that it is somehow important in the cleavage process.

We decided to investigate this issue further and study how precisely the raw reads produced by our cluster algorithm align to the genome. We noticed that in each gene, the cleavage rarely occurs at a unique position in the transcript. Instead, there are always slight fluctuations of the exact cleavage site, with a few percentages of reads ending a few nucleotides upstream and downstream of the most abundant cleavage site for a given gene (**Figure 5B**). Importantly, almost all the reads in each cluster terminate at an Adenosine nucleotide (**Figure 5B**). Also, if there are Adenosine nucleotides located within shorter *buffer regions*, the cleavage rarely occurs at these sites. Perhaps, the large size of the CPC does not allow for the docking and the cleavage of the pre mRNAs near the PAS element, which is optimally performed at 12-13nt downstream the PAS (**Figure 5A and Figure 5B**).

Next, we decided to study the role of the terminal Adenosine nucleotide in the cleavage process. We reasoned that if this Adenosine nucleotide indeed plays any role in the cleavage process, we should be able to alter the position of the mRNA cleavage site by mutating this residue with different Purines or Pyrimidines in the pre mRNAs of selected test genes.

We selected three test genes; *ges-1*, Y106G6H.9, and M03A1.3. These genes are processed only with a single 3'UTR isoform, use a single canonical PAS element, have a *buffer region* of 12, 13 and 14 nucleotides respectively and possess a terminal Adenosine nucleotide in their sequence. To capture their entire 3'UTR region, we cloned the genomic portions of these genes spanning from their translation STOP codons to ~200nt downstream of their cleavage sites. We then prepared several mutant *C. elegans* strains replacing their terminal Adenosine nucleotide at their cleavage site with other nucleotides. In the case of Y106G6H.9 we also prepared a double mutant removing an additional Adenosine nucleotide upstream of the first one located at the cleavage site (**Figure 5C and Supplemental Figure S7-S9**).

We cloned these *wt* and mutant 3'UTR regions downstream of a GFP reporter vector and prepared transgenic *C. elegans* strains that express them in the worm pharynx using the *myo-2* promoter. We opted to use the pharynx promoter since it is very strong and produces a robust expression of our constructs (**Supplemental Figure S7-S9**). We prepared transgenic worm strains expressing these constructs, recovered total RNAs, and tested using RT-PCR and a sequencing approach if the absence of the terminal Adenosine nucleotide in our mutants affects the position of the cleavage site (**Figure 5C and Supplemental Figure S7-S9**).

We observed an overall disruption of the cleavage process, in some case more pronounced than in others (**Figure 5C and Supplemental Figure S7-S9**). In the case of

M03A1.3, the absence of the terminal Adenosine nucleotide forces the cleavage complex to backtrack in 40% of the tested clones and cleave the mRNAs 3nt upstream of the original cleavage site, but still at an Adenosine nucleotide (**Figure 5C and Supplemental Figure S7**).

In the case of Y106G6H.9, the single mutant does not alter the position of the cleavage site, but interestingly activates a novel cryptic cleavage site 100 nucleotides upstream of the canonical cleavage site in 20% of the sequenced clones ~ (**Figure 5C and Supplemental Figure S8**). This new site also possesses a non-used PAS element containing the motif YRYRRR, which could still be recognized by the CPSF core complex, and a *buffer region* of 12nt. The Y106G6H.9 double mutant in one case skips the original cleavage site but still cut at the next Purine residue, which is not an Adenosine in this case (**Supplemental Figure S8**). In the case of *ges-1*, mutating the terminal Adenosine does not change the cleavage pattern, although it became more imprecise (**Supplemental Figure S9**).

Updated miRANDA prediction in *C. elegans*

Next, we used our new UTRome v2 dataset to update MiRanda miRNA target predictions. We downloaded and locally ran the miRanda prediction software (John et al. 2004) using our new 3'UTRome v2 as a target dataset. We have produced two sets of predictions; one generic, which contain the entire output produced by the software, and one more restrictive, in which we only output predictions with high scoring and with low Eenergy scores. These two tracks have been uploaded in both the 3’UTRome database (Mangone et al. 2008; Mangone et al. 2010) and the WormBase (Stein et al. 2001).

DISCUSSION

Here we have used a blind genome-wide approach to refine and study the 3’UTRome in the nematode *C. elegans*. We have identified 3’UTR data for 14,808 genes, corresponding to 23,159 3’UTR isoforms, improving their annotation. We now have 3’UTR data for 73% of all protein-coding genes included in the WS250 release. This dataset is not complete, since we could not assign 3’UTR data for the remaining 5,000 protein-coding genes present in WS250. Some of these genes may be transcribed at very low abundance and their mRNA is present below the sensitity of our approach, or their 3’UTRs data were discarded by our highly stringent filters used during our 3’UTR cluster preparation.

Alternative Polyadenylation is widespread in *C. elegans*, with ~42% of genes possessing at least two 3’UTR isoforms (**Figure 3A**). The PAS usage is still most commonly the hexamer ‘AAUAAA’ which is used to process ~58% of all *C. elegans* 3’UTRs (**Figure 3B**). Importantly, we found that the remaining 42% possess a variation of this canonical PAS element which indeed is very similar in chemical composition and contain an ‘RRYRRR’ motif at the same location where the PAS element is expected (**Figure 4A**). We do not have direct evidence that the CPC recognizes this motif, but since it is so conserved we hypothesize that in *C. elegans* it may provide a docking site in the absence of the canonical AAUAAA site during the cleavage reaction.

Our superimposition of the *C. elegans* CPSF ortholog to the human cryo-EM structure (Clerici et al. 2018; Sun et al. 2018) in **Figure 4B and Supplemental Figure S6** supports our hypothesis, suggesting that in worms the pocket used by this complex to bind the PAS element may accommodate other nucleotides as long as they have a similar chemical structure and can recapitulate the 'RRYRRR' motif. In humans, the second most abundant PAS element is 'AUUAAA' (Sun et al. 2018), which does not follow this guideline, suggesting that perhaps other factors can contribute to the cleavage of non-canonical PAS elements in other species.

Our analysis on the cleavage site found that the Cleavage and Polyadenylation machinery does not always cleave the same mRNA at the same position on the 3'UTR (**Figure 5B**). While a predominant site is often chosen for each gene, a slight variation of a few nucleotides upstream or downstream of the cleavage site is also possible. Importantly, this slight variation almost invariably ends at an Adenosine nucleotide in the genome, suggesting that this nucleotide is somehow 'sensed' in the cleavage process.

Our mutagenesis results also support an important role for the terminal Adenosine nucleotide during the cleavage reaction (**Supplemental Figures S7-S9**). In that experiments, the loss of this terminal Adenosine nucleotide disrupts in some cases the location of the cleavage, either activating cryptic cleavage sites or backtracking and using a different Adenosine nucleotide upstream the canonical cleavage site (**Supplemental Figures S7-S9**).

The concept of mRNAs terminating with an Adenosine nucleotide is not novel. Pioneering work using 269 vertebrate cDNA sequences has shown that ~71% of these genes terminate with a CA nucleotide element (Sheets et al. 1990). These experiments were biochemically validated a few years later using SV40 Late PolyA signal in mammalian cells in a more controlled environment (Chen et al. 1995). These experiments also showed that, at least for the case of this specific 3'UTR, the cleavage could not occur closer than 11nt from the PAS element and no farther of 23nt from it (Chen et al. 1995). In this context, these findings could explain why we do not detect a terminal Adenosine at the cleavage site with our double mutant Y106G6H.9, which is 27nt downstream the PAS element (**Supplemental Figure S8**). Of note, in the case of this gene, the cleavage still occurs at a Purine nucleotide, suggesting that perhaps another terminal Purine can compensate for the absence of an Adenosine nucleotide.

Overall, experiments in **Figure 5C and Supplemental Figures S7-9** support and expand both these initial results, showing that the altering nucleotide composition downstream the PAS element may influence the location of the cleavage.

Unfortunately, our study does not have the resolution to definitely verify if this Adenosine nucleotide is indeed included in the processed mRNAs or used by the CPC as a genomic mark of the cleavage site. More specifically we do not know if this nucleotide is read by the RNA polymerase II and incorporated in the nascent mRNAs or if the machinery somehow 'senses' its presence and cleaves the mRNA upstream of it. Another attractive hypothesis is that CPSF73 may cleave the mRNAs somewhere downstream of this terminal Adenosine nucleotide, and then unknown exonucleases degrade the mRNA molecule until the first Adenosine in a row is reached. Some insights may come from the process underlining histone 3'end formation, since CPSF73 also cleaves these polyA-lacking histone mRNAs. In this specific case, the enzyme is positioned near the cut site by the U7 snRNP, and interestingly cuts the nascent pre-mRNA just downstream of an Adenosine nucleotide (Yang et al. 2009). We speculate that perhaps CPSF73 is capable of either ‘sensing’ this terminal Adenosine nucleotide or is positioned next to it by either other members of the CPC or by a not yet identified factor.

If this terminal Adenosine is indeed incorporated in the pre-mRNAs, its functional requirement is unclear. It may be used by the polyA polymerase enzyme as a substrate to extend the polyA tail after the cleavage reaction has been completed, or perhaps has an unknown regulatory function. More experiments need to be performed to answer these questions.

Of note, while we observed a terminal Adenosine nucleotide in most of the mapped 3’UTRs, the Cytosine nucleotide previously identified upstream of the terminal Adenosine in humans is replaced with another Pyrimidine nucleotide in *C. elegans* (Thymidine) (**Figure 3C**), suggesting that other factors may contribute to the cleavage site decision by the CPC in higher eukaryotes.

MiRanda predictions were obsolete and needed to be updated since those present in the microrna.org database (www.microrna.org) were obtained using a 9-year-old 3’UTR dataset. Also, before this study, WormBase (Stein et al. 2001) did not include miRNA targeting predictions in its JBrowse software.

The number of predicted miRNA targets is now decreased from 34,186 to 23,160, mostly because several 3’UTR isoforms in the 3’UTRome v1 were discarded in this new 3’UTRome v2 release.

In conclusion, this new 3’UTR dataset, which we renamed 3’UTRome v2, has been uploaded to the WormBase (Stein et al. 2001) and it is shown as a new track in the JBrowse tool together with updated MiRanda miRNA target predictions. The 3’UTRome v2 expands the old 3’UTRome developed within the modENCODE Consortium, and together with updated MiRanda predictions provides the *C. elegans* Community with an important novel resource to investigate the RNA cleavage and polyadenylation reaction, 3'UTR biology and miRNA targeting.

METHODS

Comparative analysis of *C. elegans* members of the CPC

We have downloaded the protein sequences of each known member of the human CPC and used BLAT algorithm to identify *C. elegans* genes with high homology to their human counterparts. We then performed a Protein BLAST analysis using the tools available at the NCBI website to obtain the amino acid sequences for the fly, rat, and mouse orthologs. These amino acid sequences were then aligned using Clustal Omega Multiple Sequence Alignment with standard parameters. At the completion of the analysis, we used the Batch NCBI Conserved Domain Search (Batch CD-Search) against the database CDD- 52910 PSSMs using standard parameters to identify the conserved domains across the aligned protein sequences. We then used these results to populate the location of these elements within the alignment shown in **Supplemental Figure S1**.

3'UTR mapping pipeline

We have use the SRA toolkit from the NCBI to download raw reads from 1,094 transcriptome experiments. The complete list of datasets used in this study is shown in **Supplemental Table S1**. We restricted the analysis to sequences produced from *C. elegans* transcriptomes using the Illumina platform and with reads of at least 150nt in length. At the completion of the download step, the files were unzipped and stored in our servers. We then used custom-made Perl scripts to extract reads containing at least 23 consecutive Adenosine nucleotides at their 3'end or 23 consecutive Thymidine nucleotides at their 5'end. This filter produced 24,973,286 mappable 3'end reads. We then removed the terminal Adenosine or Thymidine nucleotides from these sequences, converted them to fastq files using the FASTX-Toolkit (CSHL), and mapped them to the WS250 release of the *C. elegans* genome using Bowtie2 algorithm with standard parameters (Langmead and Salzberg 2012). The Bowtie algorithm mapped 7,761,642 reads (31.08%), which were sorted and separated, based on their respective strand origin (positive or negative).

Cluster Preparations

PolyA clusters were prepared as follow. We stored the ID, genomic coordinates, and the strand orientation of each mapped read, and used this information throughout the pipeline. The BAM file produced by the aligners were sorted and converted to BED format using SAMtools software (Li et al. 2009). Contiguous genomic coordinates were merged using BEDTools software (Quinlan and Hall 2010) using the following command '`Bedtools merge -c 1 -o count -I > tmp.cluster`'. This new file produced the characteristic 'shark fin' graph visible in **Figure 2**. We used several stringed filters to eliminate as much as noise possible. 1) We ignored clusters composed of less than 6 reads. 2) We extracted genomic DNA sequences 20nt downstream the end of each cluster. If the number of Adenosine nucleotides was more than 65% in the genomic sequence, we ignored the corresponding cluster and marked it as caused by mispriming during the second strand synthesis in the RT reaction. 3) We ignored clusters overlapping with other clusters in the same orientation by 2nt or less were both ignored. 4) We attached clusters to the closest gene in the same orientation. If no gene could be identified within 2,000nt the cluster was ignored. 5) In cases with multiple 3'UTR isoforms identified, we calculated the frequency of occurrence for each isoform and ignored isoforms occurring at a frequency of less than 1% independently from the number of reads that form this cluster.

Plasmid DNA isolation, sequencing and visualization

All plasmids used in this study were prepared from cultures grown overnight in LB using the Wizard Plus SV Minipreps DNA Purification System (Promega) according to the manufacturer's instructions. DNA samples were sequenced with Sanger sequencing performed at the DNASU Sequencing Core Facility (The Biodesign Institute, ASU, Tempe, AZ).

RNAi experiments

RNAi experiments were performed in Standard NGM agar containing 1mM IPTG and 50 µg/ml Ampicillin. These plates were seeded with 75 µl of RNAi clone bacteria and allowed to induce for a minimum of 16 hours. 5 N2 *C. elegans* at the L1 stage were aliquoted for each RNAi clone tested. Three days after plating, the progeny was scored for embryonic lethality. Each RNAi experiment was performed in triplicate. The total number of hatched and not matched eggs was the following: *cpsf-1(CPSF160)* n=567*; cpsf-2(CPSF100)* n=557*; cpsf-4(CPSF30)* n=1,251*; cpf-2(CstF64)* n= 652*; cpf-1(CstF50)* n=801*; cfim-1(CFIm25)* n=644*; cfim-2(CFIm68)* n=739*; lrp-2(CFIm59)* n=1,120*; symk-1(Symplekin)* n=208*; tag-214(RBBP6)* n=753*; pcf-11(CPF11)* n=428*; clpf-1(CLP1*) n=841*.*

Extraction of 3'UTR regions from the *C. elegans* genome

The 3'UTRs used in the experiments described in **Figure 5C and Supplemental Figure S7-9** were initially cloned from N2 *wild type C. elegans* genomic DNA using PCR with Platinum Taq Polymerase (Invitrogen). Genomic DNA template was prepared as previously described (Blazie et al. 2017). Forward DNA primers were designed to include approximately 30 nucleotides upstream of the translation STOP codon and include the endogenous translation STOP codon. We used the Gateway BP Clonase II Enzyme Mix (Invitrogen) to clone the 3'UTR region into Gateway entry vectors. The DNA primer was modified to include the attB Gateway recombination elements required for insertion into pDONR P2RP3 (Invitrogen). The reverse DNA primers were designed to end between 200 and 250 nucleotides downstream of the RNA cleavage site and to include the reverse recombination element attB for cloning into pDONR P2RP3 (Invitrogen). At the conclusion of the recombination step, the entry vectors containing the cloned 3'UTR regions were transformed into Top10 competent cells (Thermo Fisher Scientific), using agar plates containing 20mg/μL of Kanamycin. The plasmids were then recovered, and clones were confirmed using Sanger sequencing with the M13F primer. The list of primers used in this study is available in **Supplemental Table S2**.

Mutagenesis of 3'UTRs cleavage sites

The mutagenesis reactions to remove the Adenosine nucleotides near the cleavage sites were carried out using the QuikChange Site-Directed Mutagenesis Kit (Agilent). The mutagenesis DNA primers for the site mutation reactions are available in **Supplemental Table S2**. Each mutagenesis reaction was followed by DNA digestion using Dpn-1 enzyme and transformed in Top10 competent cells (Thermo Fisher Scientific) in agar plates containing 20mg/µL of Kanamycin. We validated the nucleotide mutation using Sanger sequencing approach. *Wild type* and mutant 3'UTRs cloned in pDONR P2RP3 were then shuttled into destination vectors using the Gateway LR Clonase II Plus Enzyme Mix (Invitrogen, Carlsbad, CA). The finalized destination vectors contained the *C. elegans* pharynx promoter (Pmyo-2) in the first position, a GFP sequence with a mutated STOP codon in the second position, and the *wt* or mutant 3'UTRs used in this study in the third position. The resultant recombined constructs were then transformed in Top10 competent cells (Thermo Fisher Scientific) and plated on 10mg/µL Ampicillin plates overnight. The success of the recombination reaction was confirmed using Sanger sequencing with the M13F DNA primer.

Preparation of transgenic worm lines

Eg6699 strain worms were kindly provided by Christian Frokjaer-Jensen (Frokjaer-Jensen et al. 2008). These worm strains were maintained at 18°C on nematode growth media (NGM) agar plates and propagated on plates seeded with OP50-1 bacteria. To synchronize worms for injections, Eg6699 worms were bleached with bleaching solution (1 M NaOH) four days before injections. Each construct was mixed with an injection master mix containing pCFJ601 (25 ng/µl), pgH8 (10 ng/µl), and pCFJ104 (5 ng/µl) vectors. Injection needles were loaded with the injection mixture and mounted to the Leica DMI300B microscope. The needle was pressurized with 22 psi through the FemtoJet (Eppendorf). Young adult Eg6699 worms were picked onto an agarose pad covered with mineral oil on a glass coverslip. Injected worms were rescued onto an NGM plate and rinsed with M9 buffer. Two days post-injections, the F1 progeny were screened with a Leica DMI3000B microscope for both *unc-119* rescues and expression of the red fluorescence produced by the co-injection marker and then isolated onto individual plates. These worms were allowed to lay eggs, and then the F2 progeny was screened for fluorescence. Once 75% of the progeny on a single plate were transgenic, the strains were used for further experimentation.

Worm genotype validation

Populations obtained from single worms from each of the seven strains were lysed using worm lysis buffer (EDTA, 0.1 M Tris, 10% Triton-X, Proteinase K, 20% Tween 20). These worms were subjected to heating in a Bio-Rad T100 Thermal Cycler. To confirm that the mutated cleavage site was present in the injected strains, we used PCR approach using Platinum Taq polymerase (Invitrogen) with a forward DNA primer binding the beginning of the GFP sequence and 3'UTR-specific reverse DNA primers. The PCR product was then sequenced using Sanger sequencing with a forward DNA primer binding to the GFP sequence present in the injected construct.

Detection of the 3'UTR cleavage skipping

Total RNA was extracted from transgenic strains using the Direct-zol RNA MiniPrep Plus kit (RPI) according to the manufacturer's instructions. We tested approximately 10 independent *wt* and mutant clones for each 3'UTR. Approximately 50 μL of worm pellet was used for extraction. cDNA was synthesized using a reverse transcription reaction using Superscript II enzyme (Invitrogen). The first strand reaction was performed using a reverse poly dT DNA primer containing two anchors and the attB Gateway BP recombination element (Invitrogen). The second strand of the cDNA was synthesized using a PCR with HiFi taq polymerase (Thermo Fisher Scientific) and the forward DNA primer containing the pDONR P2RP3 Gateway element (Invitrogen), which binds to GFP and the same reverse poly dT DNA primer used in the first strand reaction. The BP Gateway kit (Invitrogen) was once again used to clone the cDNA which contains the polyA tail into pDONR P2RP3. These constructs were then transfected into Top10 competent cells (Thermo Fisher Scientific) and plated on agar plates containing 20mg/μL of Kanamycin. About 8-10 colonies were then sequenced with Sanger sequencing using the M13F DNA primer to map the location of the cleavage site.

Updated MiRanda Predictions

We downloaded a complete list of *C. elegans* miRNAs from miRBase (Griffiths-Jones et al. 2006) and the miRanda algorithm v3.3a (John et al. 2004) from the microrna.org website. We queried the 3'UTRome v2 with the miRanda algorithm using both standard and stringent parameters. The stringent query used was '-strict -sc -1.2'. The standard query produced 58,330 putative miRNA targets; the stringent query produced 12,136 putative miRNA targets. Both these predictions are included in WormBase (Stein et al. 2001) as individual tracks.

Homology model building

Homology modeling was performed using SWISS_MODEL (Waterhouse et al. 2018) with a matched templated of human CPSF160-WDR33-CPSF30 complex

(PDB code: 6DNF) (Sun et al. 2018). The molecular graphics were prepared using the UCSF ChimeraX software (version 0.8) (Goddard et al. 2018).

Data Availability

Strains and plasmids are available upon request. The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures and supplemental figures, and tables and supplemental tables. The results of our analyses are available in the WormBase (www.WormBase.org) (Stein et al. 2001) and in our 3’UTR-centric website www.UTRome.org.

Author Contribution

HSS and MM designed the experiments. MM developed and executed the bioinformatic analysis and 3’UTR cluster preparation. HSS performed the rescue experiments in **Figure 5**. PLC performed the homology modeling in **Figure 4 and Supplemental Figure S6** and helped writing the manuscript. CG assisted with the experiments and performed the analysis in **Supplemental Figure S1**. SO contributed to the experiments in **Figure S7-S9**. MM uploaded the results to the WormBase and UTRome.org database. MM and HSS led the analysis and interpretation of the data, assembled the Figures, and wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Institute of Health grant number 1R01GM118796.

Conflict of Interest

The authors declare that they have no competing interests.

Acknowledgements

We thank Heather Hrach for insights and review of the manuscript. We thank Gabrielle Richardson for maintaining the *C. elegans* strains used in this manuscript.

FIGURE LEGENDS

Figure 1. The *C. elegans* members of the Cleavage and Polyadenylation Complex (CPC). A) The CPC is composed of at least 4 independent subcomplexes named Cleavage and Polyadenylation Specificity Complex (Blue), which canonically recognizes the PAS hexamer 'AAUAAA'; the Cleavage Stimulation Factor Complex (Green), which binds downstream of the cleavage site to GU rich elements; and the Cleavage Factor CFIm (Red) and CFIIm (Orange) Complexes. CFIm recognizes the element 'UGUA' located upstream of the PAS element. Other known required factors are the PolyA Polymerase enzyme, the scaffolding member Symplekin and RBBP6. The name of the *C. elegans* orthologs are shown in parenthesis. B) The human and *C. elegans* CPSF subcomplexes are similar in amino acid composition and structure. 2-species alignments between several members of the human and *C. elegans* CPSF members. Amino acids 100% conserved

between these two species are shown in red in the conservation bar. Yellow dotted boxes show the sequence of the proteins that interact with the PAS element. Functional domains are conserved. The two Kyte-Doolittle graphs in each panel indicate the hydrophobic amino acids in human and *C. elegans*. C) We have used RNAi to selectively silence most of the members of the CPC complex in *C. elegans.* We observed a strong embryonic lethality phenotype with all the RNAi experiments performed.

Figure 2. Cluster preparation and analysis. A) Screenshots showing several mapped 3'UTR clusters for genes with one or two 3'UTR isoforms. miRanda predicted miRNA targets are shown for a particular 3'UTR at the bottom of this Panel. B) Summary of the 3'UTRs in genes identified in this study along with the number of reads mapped and clustered for each 3'UTR. C) Comparison between the 3'UTRs for genes and total isoforms mapped in this study vs the UTRome v1 (Mangone et al. 2010) and the dataset from Jan et al., 2001.

Figure 3. The worm 3'UTRome v2. A) Top Left Panel. Nucleotide composition of 3'UTRs in the 3'UTRome v2. Uridine is the most abundant nucleotide within 3'UTRs for *C. elegans*. Bottom Left Panel. The number of 3'UTR isoforms in each gene. 42% of the genes in the 3'UTRome v2 possess multiple 3'UTR isoforms. Top Right Panel. 3'UTR length distribution in genes expressed with one, two, or three or more 3'UTR isoforms. The median 3'UTR length across these datasets is 122nt. Genes with multiple 3'UTR isoforms are on average longer than genes with one 3'UTR isoform. Bottom Right Panel. Median 3'UTR length in genes with Canonical (C) or

Variant (V) PAS elements. There is a slight increase in 3'UTR length in genes with variant PAS elements when compared to those with canonical PAS elements. This variation is still detected when increasing the stringency of the density of the clusters (cd) used in this analysis. B) PAS element usage in 3'UTRs. 58.4% of 3'UTRs use the canonical PAS element 'AAUAAA' while the most common variant PAS element is the hexamer 'AAUGAA', which occurs in 11% of genes. The distribution of canonical PAS elements within 3'UTRs. The average distance from the PAS element to the cleavage site is 18nt. C) Alignment of 3'UTRs at the cleavage site. This alignment in genes with both canonical and variant PAS elements reveals a region between the PAS element and the cleavage site we renamed the *buffer region* in which cleavage rarely occurs. The most abundant nucleotide at the cleavage site is an Adenosine nucleotide preceded by a Thymidine nucleotide.

Figure 4. The sequence requirements of the *C. elegans* CPSF core complex. A) PAS element usage of the RRYRRR motif. 3'UTRs from the 3'UTRome v2 aligned by their cleavage site in genes with canonical or variant PAS element. The motif RRYRRR is highlighted in yellow, and its spatial conservation is very strong in single 3'UTR isoforms with canonical PAS elements and is enriched in those with variant PAS elements. This RRYRRR element is maintained in 3'UTRs that have at least two isoforms but is not strongly represented in human 3'UTR data due to the lack of their annotation. R= Purine, Y= Pyrimidine. B) Superimposition of the cryo-EM structure of the previously published human CPSF core complex (Clerici et al.

2018; Sun et al. 2018) to the worm CPSF core complex: cpsf-1 (CPSF-160) in blue, pfs-2 (Wdr33) in pink, and cpsf-4 (CPSF30) in green. The PAS element binding pocket can be fitted into the homology model. The PAS element of the RNA is represented in yellow. The size and the selectivity of the nucleotide binding pocket can fit other nucleotides as long as the motif is RRYRRR.

Figure 5. The Adenosine nucleotide is required at the cleavage site for correct cleavage. A) Sequence Logos produced from 3'UTRs from genes with only 3'UTR isoforms containing the canonical PAS element 'AAUAAA' and aligned by their respective *buffer region* length (n=4,374). Two extra nucleotides are included downstream of each cut site (triangle). The nucleotide distribution of the distance between the PAS element and the cleavage site is shown in the bar chart below. B) Example of slight variability in the cleavage site for the gene C09G9.8. While prevalent forms are observed, the exact cleavage site can vary on several occasions but predominantly occurs at a different Adenosine nucleotide. C) Test of the role of the terminal Adenosine nucleotide in the cleavage reaction. The 3'end regions of several test genes where cloned and used to prepare transgenic *C. elegans* strains expressing this region with or without mutated terminal Adenosine nucleotides (Red, see below). The top sequence shows the test 3'end region (Cyan=ORF, Green=translation STOP signal, Grey=3'UTR, Red=Terminal Adenosine nucleotide. The PAS element is underscored). The Sanger trace files show the outcome of the cleavage site location in selected clones. Two genes are shown (M03A1.3 and Y106G6H.9). In the case of M03A1.3, the loss of the terminal

Adenosine nucleotide sometimes forces the CPC to backtrack and cleave the mRNAs upstream of the regular cleavage site but still at the closest Adenosine nucleotide available. In the case of the gene Y106G6H.9, the loss of the terminal Adenosine nucleotide forces the complex to skip the cleavage site, which sometimes occurs at the next Purine nucleotide. Additional clones and more test genes are shown in the **Supplemental Figure S7-S9**.

REFERENCES

Bai Y, Auperin TC, Chou CY, Chang GG, Manley JL, Tong L. 2007. Crystal structure of murine CstF-77: dimeric association and implications for polyadenylation of mRNA precursors. *Mol Cell* **25**: 863-875.

Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* **136**: 215-233.

Bartel DP. 2018. Metazoan MicroRNAs. *Cell* **173**: 20-51.

Blazie SM, Babb C, Wilky H, Rawls A, Park JG, Mangone M. 2015. Comparative RNA-Seq analysis reveals pervasive tissue-specific alternative polyadenylation in Caenorhabditis elegans intestine and muscles. *BMC Biol* **13**: 4.

Blazie SM, Geissel HC, Wilky H, Joshi R, Newbern J, Mangone M. 2017. Alternative Polyadenylation Directs Tissue-Specific miRNA Targeting in Caenorhabditis elegans Somatic Tissues. *Genetics* **206**: 757-774.

Chan SL, Huppertz I, Yao C, Weng L, Moresco JJ, Yates JR, 3rd, Ule J, Manley JL, Shi Y. 2014. CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3' processing. *Genes Dev* **28**: 2370-2380.

Chen F, Chisholm AD, Jin Y. 2017. Tissue-specific regulation of alternative polyadenylation represses expression of a neuronal ankyrin isoform in C. elegans epidermal development. *Development* **144**: 698-707.

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

Clerici M, Faini M, Aebersold R, Jinek M. 2017. Structural insights into the assembly and polyA signal recognition mechanism of the human CPSF complex. *Elife* **6**.

Clerici M, Faini M, Muckenfuss LM, Aebersold R, Jinek M. 2018. Structural basis of AAUAAA polyadenylation signal recognition by the human CPSF complex. *Nat Struct Mol Biol* **25**: 135-138.

de Vries H, Ruegsegger U, Hubner W, Friedlein A, Langen H, Keller W. 2000. Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. *EMBO J* **19**: 5895-5904.

Diag A, Schilling M, Klironomos F, Ayoub S, Rajewsky N. 2018. Spatiotemporal m(i)RNA Architecture and 3' UTR Regulation in the C. elegans Germline. *Dev Cell* **47**: 785-800 e788.

Frokjaer-Jensen C, Davis MW, Hopkins CE, Newman BJ, Thummel JM, Olesen SP, Grunnet M, Jorgensen EM. 2008. Single-copy insertion of transgenes in Caenorhabditis elegans. *Nat Genet* **40**: 1375-1383.

Gerstein MB Lu ZJ Van Nostrand EL Cheng C Arshinoff BI Liu T Yip KY Robilotto R Rechtsteiner A Ikegami K et al. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**: 1775-1787.

Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, Ferrin TE. 2018. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci* **27**: 14-25.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**: D140-144.

Haenni S, Ji Z, Hoque M, Rust N, Sharpe H, Eberhard R, Browne C, Hengartner MO, Mellor J, Tian B et al. 2012. Analysis of C. elegans intestinal gene expression and polyadenylation by fluorescence-activated nuclei sorting and 3'-end-seq. *Nucleic Acids Res* **40**: 6304-6318.

Helmling S, Zhelkovsky A, Moore CL. 2001. Fip1 regulates the activity of Poly(A) polymerase through multiple interactions. *Mol Cell Biol* **21**: 2026-2037.

Hwang HW, Park CY, Goodarzi H, Fak JJ, Mele A, Moore MJ, Saito Y, Darnell RB. 2016. PAPERCLIP Identifies MicroRNA Targets and a Role of CstF64/64tau in Promoting Non-canonical poly(A) Site Usage. *Cell Rep* **15**: 423-435.

Jan CH, Friedman RC, Ruby JG, Bartel DP. 2011. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* **469**: 97-101.

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. 2004. Human MicroRNA targets. *PLoS Biol* **2**: e363.

Kaufmann I, Martin G, Friedlein A, Langen H, Keller W. 2004. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *EMBO J* **23**: 616-626.

Kuhn U, Wahle E. 2004. Structure and function of poly(A) binding proteins. *Biochim Biophys Acta* **1678**: 67-84.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

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

Mandel CR, Kaneko S, Zhang H, Gebauer D, Vethantham V, Manley JL, Tong L. 2006. Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature* **444**: 953-956.

Mangone M, Macmenamin P, Zegar C, Piano F, Gunsalus KC. 2008. UTRome.org: a platform for 3'UTR biology in C. elegans. *Nucleic Acids Res* **36**: D57-62.

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

Martin G, Gruber AR, Keller W, Zavolan M. 2012. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Rep* **1**: 753-763.

Matlin AJ, Clark F, Smith CW. 2005. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* **6**: 386-398.

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

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

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Ryan K, Calvo O, Manley JL. 2004. Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. *RNA* **10**: 565-573.

Schafer P, Tuting C, Schonemann L, Kuhn U, Treiber T, Treiber N, Ihling C, Graber A, Keller W, Meister G et al. 2018. Reconstitution of mammalian cleavage factor II involved in 3' processing of mRNA precursors. *RNA* **24**: 1721-1737.

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

Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J. 2001. WormBase: network access to the genome and biology of Caenorhabditis elegans. *Nucleic Acids Res* **29**: 82-86.

Sun Y, Zhang Y, Hamilton K, Manley JL, Shi Y, Walz T, Tong L. 2018. Molecular basis for the recognition of the human AAUAAA polyadenylation signal. *Proc Natl Acad Sci U S A* **115**: E1419-E1428.

Takagaki Y, Manley JL. 2000. Complex protein interactions within the human polyadenylation machinery identify a novel component. *Mol Cell Biol* **20**: 1515-1525.

Tian B, Graber JH. 2012. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip Rev RNA* **3**: 385-396.

Tian B, Manley JL. 2017. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18**: 18-30.

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L et al. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **46**: W296-W303.

West SM, Mecenas D, Gutwein M, Aristizabal-Corrales D, Piano F, Gunsalus KC. 2018. Developmental dynamics of gene expression and alternative polyadenylation in the Caenorhabditis elegans germline. *Genome Biol* **19**: 8.

Yang Q, Coseno M, Gilmartin GM, Doublie S. 2011. Crystal structure of a human cleavage factor CFI(m)25/CFI(m)68/RNA complex provides an insight into poly(A) site recognition and RNA looping. *Structure* **19**: 368-377.

Yang Q, Gilmartin GM, Doublie S. 2010. Structural basis of UGUA recognition by the Nudix protein CFI(m)25 and implications for a regulatory role in mRNA 3' processing. *Proc Natl Acad Sci U S A* **107**: 10062-10067.

Yang W, Hsu PL, Yang F, Song JE, Varani G. 2018. Reconstitution of the CstF complex unveils a regulatory role for CstF-50 in recognition of 3'-end processing signals. *Nucleic Acids Res* **46**: 493-503.

Yang XC, Sullivan KD, Marzluff WF, Dominski Z. 2009. Studies of the 5' exonuclease and endonuclease activities of CPSF-73 in histone pre-mRNA processing. *Mol Cell Biol* **29**: 31-42.

Zhelkovsky A, Helmling S, Moore C. 1998. Processivity of the Saccharomyces cerevisiae poly(A) polymerase requires interactions at the carboxyl-terminal RNA binding domain. *Mol Cell Biol* **18**: 5942-5951.

Zhu Y, Wang X, Forouzmand E, Jeong J, Qiao F, Sowd GA, Engelman AN, Xie X, Hertel KJ, Shi Y. 2018. Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation. *Mol Cell* **69**: 62-74 e64.

Figure 1

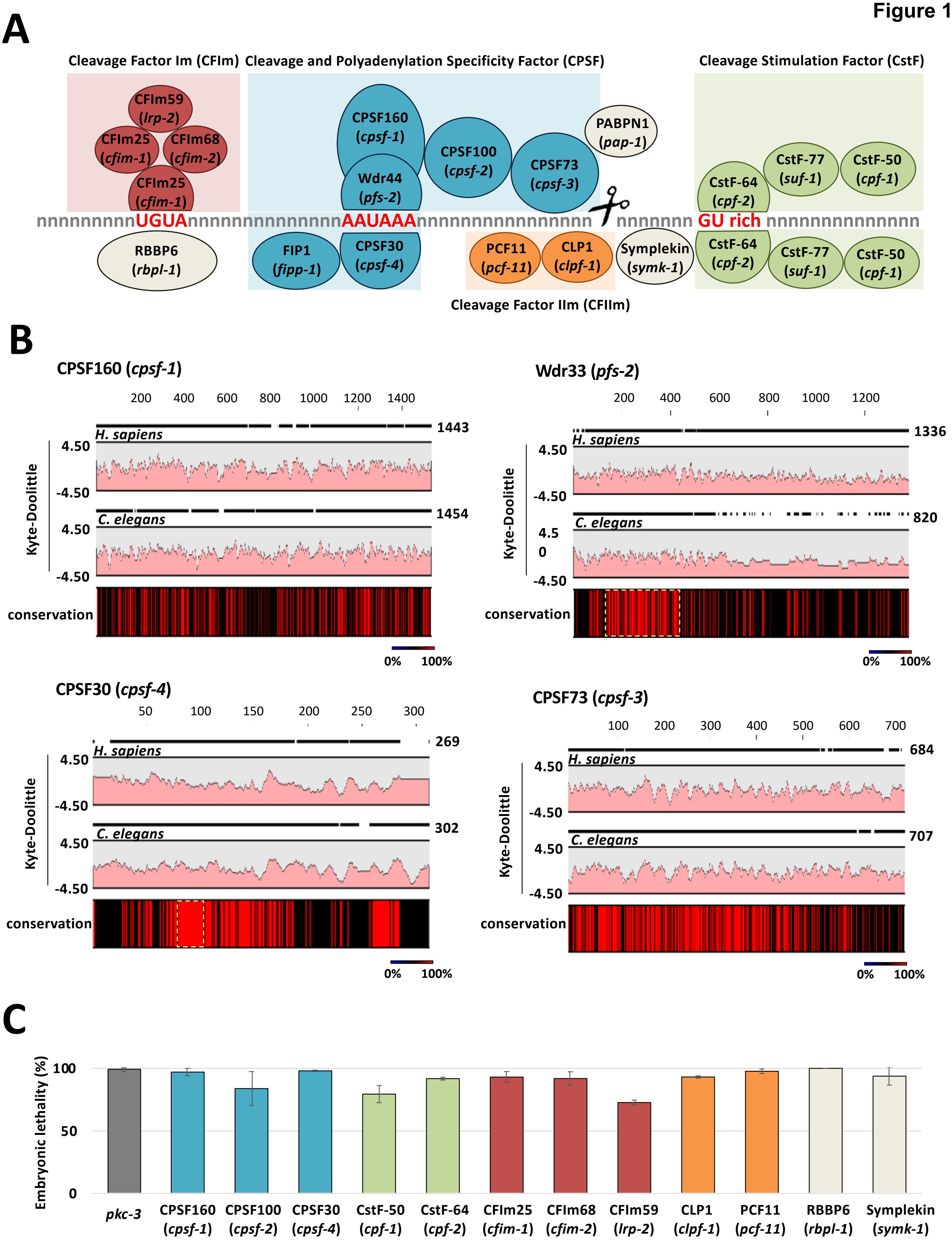
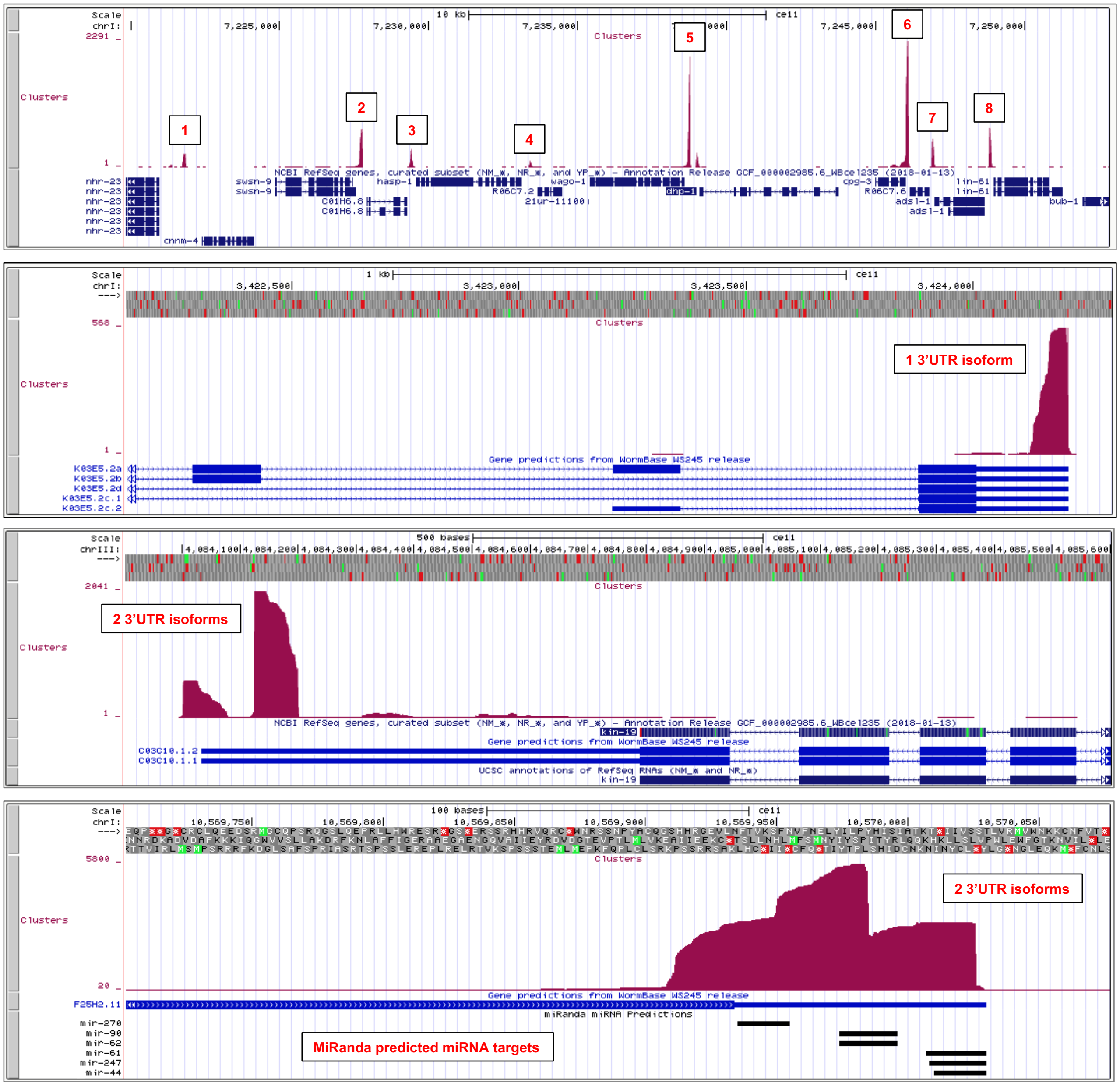

A
Cleavage Factor Im (CFIm)
Cleavage and Polyadenylation Specificity Factor (CPSF)
Cleavage Stimulation Factor (CstF)
CFIm59 (lrp-2)
CFIm25 (cfim-1)
CFIm68 (cfim-2)
CFIm25 (cfim-1)
CPSF160 (cpsf-1)
Wdr44 (pfs-2)
CPSF100 (cpsf-2)
CPSF73 (cpsf-3)
PABPN1 (pap-1)
CstF-64 (cpf-2)
CstF-77 (suf-1)
CstF-50 (cpf-1)
nnnnnnnnnUGUAnnnnnnnnnnnnnnnAAUAAAnnnnnnnnnnnnnnnnnn nnnnnnnn GU rich nnnnnnnnnnnnnn
RBBP6 (rbpl-1)
FIP1 (fipp-1)
CPSF30 (cpsf-4)
PCF11 (pcf-11)
CLP1 (clpf-1)
Symplekin (symk-1)
CstF-64 (cpf-2)
CstF-77 (suf-1)
CstF-50 (cpf-1)
Cleavage Factor IIm (CFIIm)
B
CPSF160 (cpsf-1)
200 400 600 800 1000 1200 1400
H. sapiens 1443
C. elegans 1454
Kyte-Doolittle
4.50
-4.50
4.50
-4.50
conservation
0% 100%
Wdr33 (pfs-2)
200 400 600 800 1000 1200
H. sapiens 1336
C. elegans 820
Kyte-Doolittle
4.50
-4.50
4.5
0
-4.50
conservation
0% 100%
CPSF30 (cpsf-4)
50 100 150 200 250 300
H. sapiens 269
C. elegans 302
Kyte-Doolittle
4.50
-4.50
4.50
-4.50
conservation
0% 100%
CPSF73 (cpsf-3)
100 200 300 400 500 600 700
H. sapiens 684
C. elegans 707
Kyte-Doolittle
4.50
-4.50
4.50
-4.50
conservation
0% 100%
C
Embryonic lethality (%)
100
50
0
pkc-3
CPSF160 (cpsf-1)
CPSF100 (cpsf-2)
CPSF30 (cpsf-4)
CstF-50 (cpf-1)
CstF-64 (cpf-2)
CFIm25 (cfim-1)
CFIm68 (cfim-2)
CFIm59 (lrp-2)
CLP1 (clpf-1)
PCF11 (pcf-11)
RBBP6 (rbpl-1)
Symplekin (symk-1)

Figure 2

A



B

# 3'UTRs in genes (isoforms)			# reads used	
total	in 3'UTRome v1	new	mapped	in each cluster
14,808 (23,159)	11,377 (16,941)	4,638 (6,218)	4,838,905	220 (median)

clusters >=10%

C

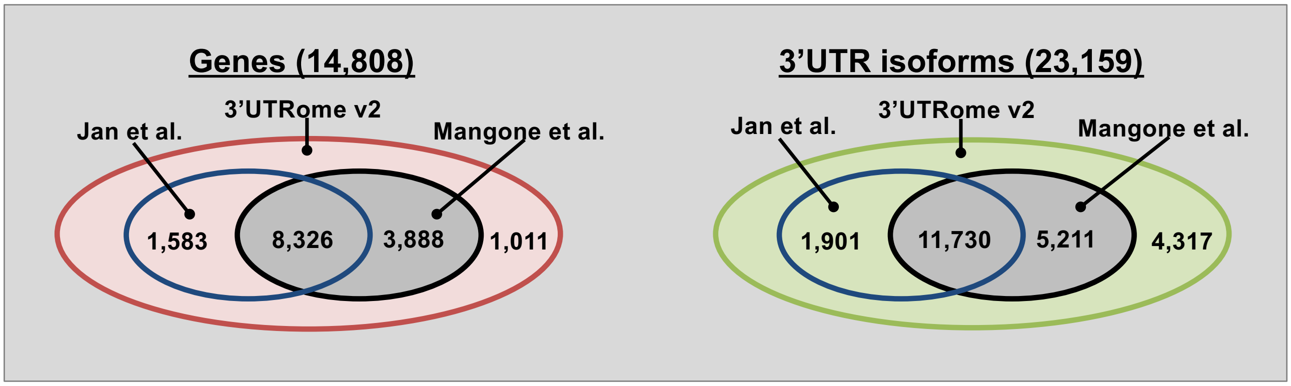


Figure 3

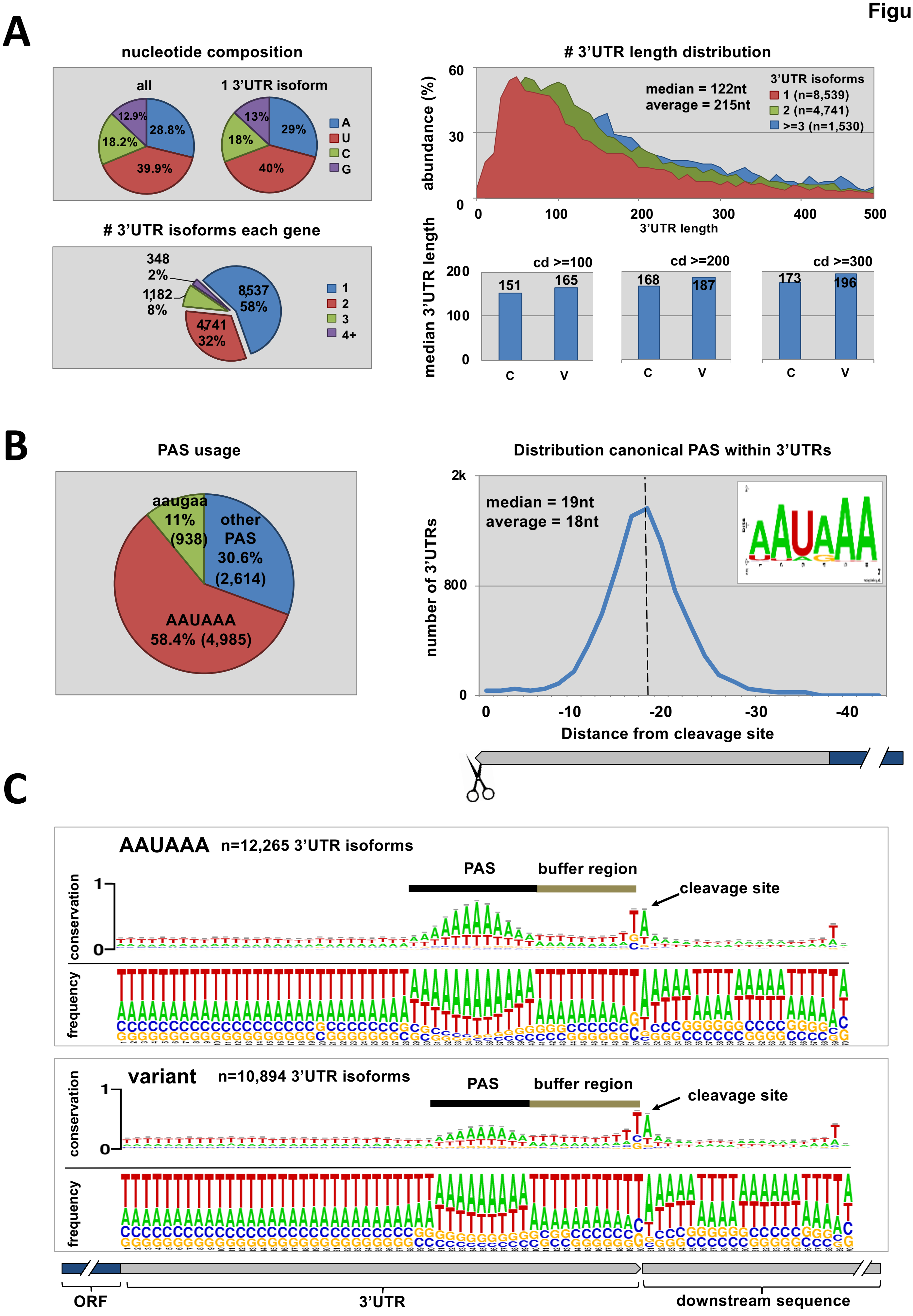

A
nucleotide composition
all
1 3'UTR isoform
12.9%
28.8%
18.2%
39.9%
13%
29%
18%
40%
A
U
C
G
# 3'UTR isoforms each gene
348
2%
1,182
8%
8,537
58%
4,741
32%
1
2
3
4+
# 3'UTR length distribution
abundance (%)
60
30
0
median = 122nt
average = 215nt
3'UTR isoforms
1 (n=8,539)
2 (n=4,741)
>=3 (n=1,530)
0
100
200
300
400
500
3'UTR length
median 3'UTR length
200
100
0
cd >=100
151
165
C
V
cd >=200
168
187
C
V
cd >=300
173
196
C
V
B
PAS usage
aaugaa
11%
(938)
other
PAS
30.6%
(2,614)
AAUAAA
58.4% (4,985)
Distribution canonical PAS within 3'UTRs
number of 3'UTRs
2k
800
0
median = 19nt
average = 18nt
0
-10
-20
-30
-40
Distance from cleavage site
C
AAUAAA n=12,265 3'UTR isoforms
PAS
buffer region
cleavage site
conservation
1
0
frequency
variant n=10,894 3'UTR isoforms
PAS
buffer region
cleavage site
conservation
1
0
frequency
ORF
3'UTR
downstream sequence

Figure 4

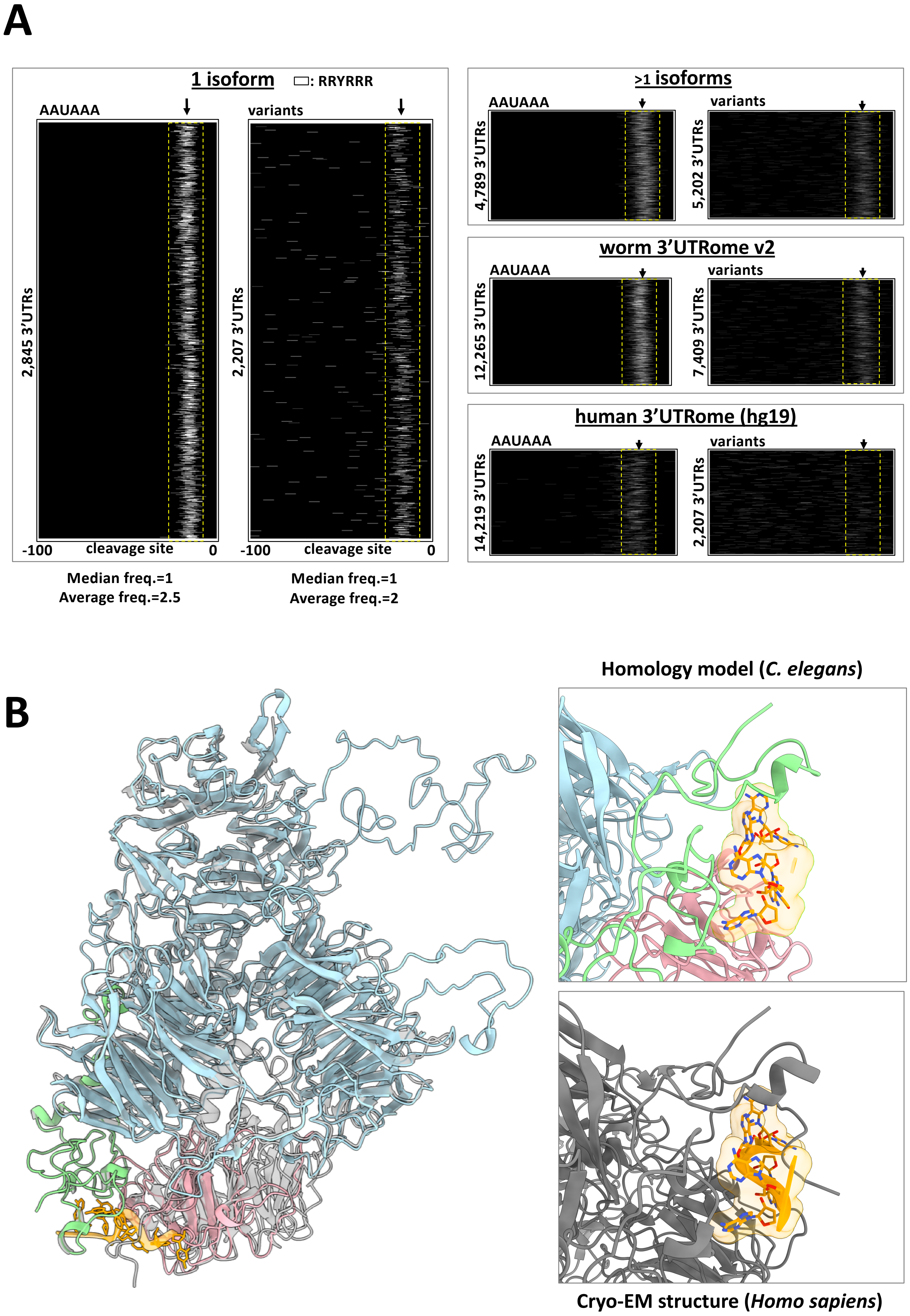
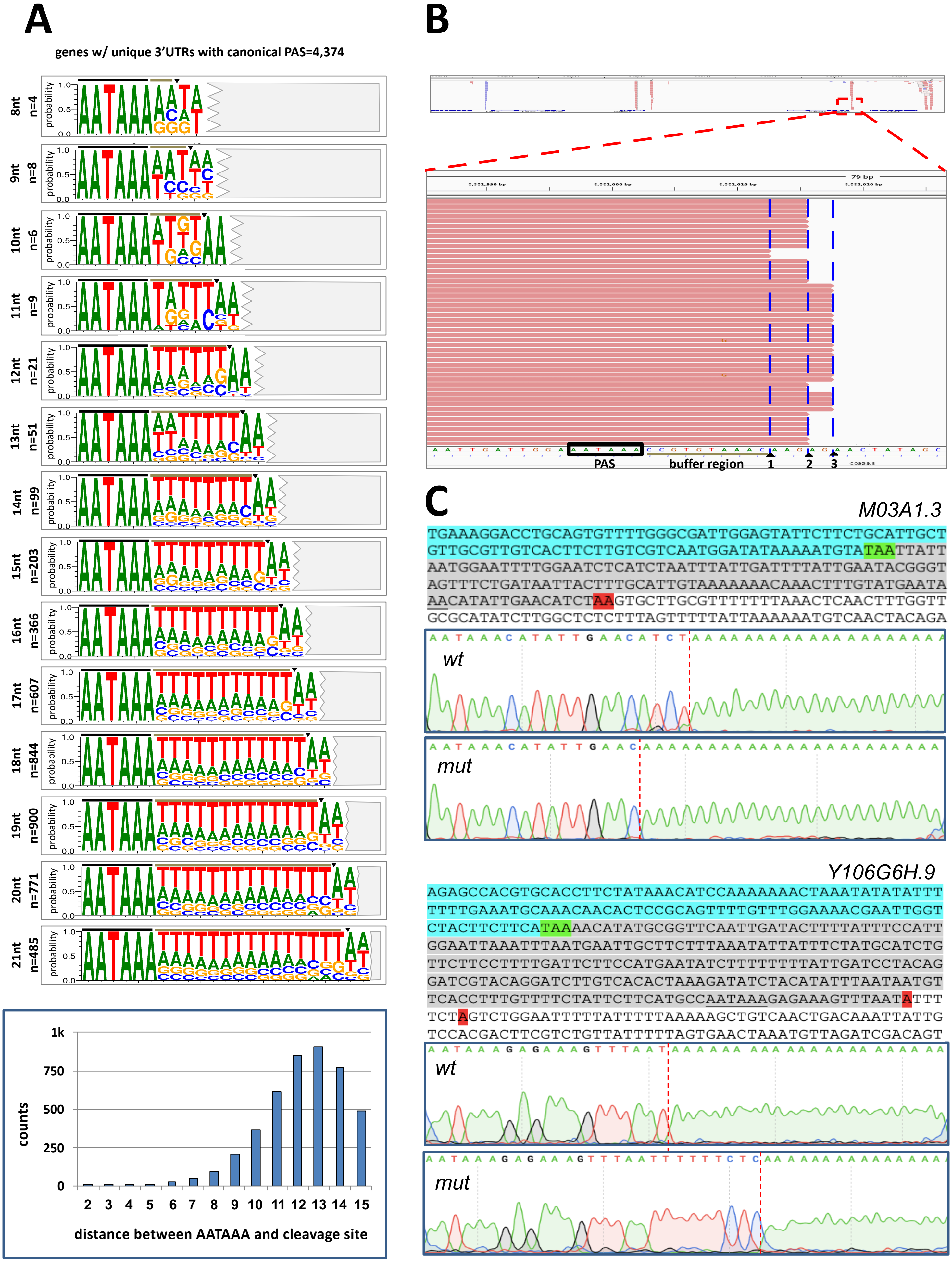

A
1 isoform
▭: RRYRRR
AAUAAA
variants
2,845 3'UTRs
2,207 3'UTRs
-100
cleavage site
0
-100
cleavage site
0
Median freq.=1
Average freq.=2.5
Median freq.=1
Average freq.=2
>1 isoforms
AAUAAA
variants
4,789 3'UTRs
5,202 3'UTRs
worm 3'UTRome v2
AAUAAA
variants
12,265 3'UTRs
7,409 3'UTRs
human 3'UTRome (hg19)
AAUAAA
variants
14,219 3'UTRs
2,207 3'UTRs
B
Homology model (*C. elegans*)
Cryo-EM structure (*Homo sapiens*)

Figure 5


A
genes w/ unique 3'UTRs with canonical PAS=4,374
8nt n=4
9nt n=8
10nt n=6
11nt n=9
12nt n=21
13nt n=51
14nt n=99
15nt n=203
16nt n=366
17nt n=607
18nt n=844
19nt n=900
20nt n=771
21nt n=485
probability
1.0
0.5
0.0
1k
750
500
250
0
counts
2 3 4 5 6 7 8 9 10 11 12 13 14 15
distance between AATAAA and cleavage site
B
79 bp
8,881,990 bp
8,882,000 bp
8,882,010 bp
8,882,020 bp
PAS
buffer region
1
2
3
C09G9.8
C
M03A1.3
TGAAAGGACCTGCAGTGTTTTGGGCGATTGGAGTATTCTTCTGCATTGCT
GTTGCGTTGTCACTTCTTGTCGTCAATGGATATAAAAATGTATAATTATT
AATGGAATTTTGGAATCTCATCTAATTTATTGATTTTATTGAATACGGGT
AGTTTCTGATAATTACTTTGCATTGTAAAAAAACAAACTTTGTATGAATA
AACATATTGAACATCTAAGTGCTTGCGTTTTTTTAAACTCAACTTTGGTT
GCGCATATCTTGGCTCTCTTTAGTTTTTATTAAAAAATGTCAACTACAGA
wt
mut
Y106G6H.9
AGAGCCACGTGCACCTTCTATAAACATCCAAAAAAACTAAATATATATTT
TTTTGAAATGCAAACAACACTCCGCAGTTTTGTTTGGAAAACGAATTGGT
CTACTTCTTCATAAAACATATGCGGTTCAATTGATACTTTTATTTCCATT
GGAATTAAATTTAATGAATTGCTTCTTTAAATATTATTTCTATGCATCTG
TTCTTCCTTTTGATTCTTCCATGAATATCTTTTTTTTATTGATCCTACAG
GATCGTACAGGATCTTGTCACACTAAAGATATCTACATATTTAATAATGT
TCACCTTTGTTTTCTATTCTTCATGCCAATAAAGAGAAAGTTTAATATTT
TCTAGTCTGGAATTTTTATTTTTAAAAAGCTGTCAACTGACAAATTATTG
TCCACGACTTCGTCTGTTATTTTTAGTGAACTAAATGTTAGATCGACAGT
wt
mut