

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

3 Steber HS^{1,2}, Gallante C³, O'Brien S², Chiu P.-L⁴, Mangone M^{*1,2}.

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

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

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

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

11
12 * To whom correspondence should be addressed. Tel: +1(480) 965-7957; Fax: +1(480) 965-3051;
13 Email: mangone@asu.edu

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

16

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

19

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

21

22

23

24 **ABSTRACT**

25

26 3'Untranslated Regions (3'UTRs) of mRNAs emerged as central regulators of
27 cellular function as they contain important but poorly-characterized *cis*-regulatory
28 elements targeted by a multitude of regulatory factors. The soil nematode *C.*
29 *elegans* is an ideal model to study these interactions since it possesses a well-
30 defined 3'UTRome. In order to improve its annotation, we have used a genomics
31 approach to download raw transcriptome data for ~1,500 transcriptome datasets
32 corresponding to the entire collection of *C. elegans* transcriptomes from 2015 to
33 2018 from the Sequence Read Archive at the NCBI. We then extracted and
34 mapped high-quality 3'UTR data at ultra-deep coverage. Here we describe and
35 release to the Community the updated version of the worm 3'UTRome, which we
36 named 3'UTRome v2. This resource contains high-quality 3'UTR data mapped at
37 single base ultra-resolution for 23,159 3'UTR isoforms variants corresponding to
38 14,808 protein-coding genes and is updated to the latest release of WormBase. We
39 used this dataset to study and probe principles of RNA cleavage and
40 polyadenylation in *C. elegans*. The worm 3'UTRome v2 represents the most
41 comprehensive and high-resolution 3'UTR dataset available in *C. elegans*, and
42 provides a novel resource to investigate the mRNA cleavage and polyadenylation
43 reaction, 3'UTR biology and miRNA targeting in a living organism.

44

45

46

47

48

49

BACKGROUND

50

51 3'Untranslated Regions (3'UTRs) are the portions of mRNA located between the
52 end of the coding sequence and the polyA tail of RNA polymerase II-transcribed genes.
53 They contain *cis*-regulatory elements targeted by miRNAs and RNA binding proteins and
54 modulate mRNA stability, localization, and overall translational efficiency (Bartel 2018).
55 Because multiple 3'UTR isoforms of a particular mRNA can exist, differential regulation of
56 3'UTRs has been implicated in numerous diseases, and its discriminative processing
57 influences development and metabolism (Mayr and Bartel 2009; Zhu et al. 2018). 3'UTRs
58 are processed to full maturity through cleavage of the nascent mRNA and subsequent
59 polyA tail addition to its 3' end by the nuclear polyA polymerase enzyme (PABPN1) (Kuhn
60 and Wahle 2004). The mRNA cleavage step is a dynamic regulatory process directly
61 involved in the control of gene expression in Eukaryotes. The reaction depends on the
62 presence of a series of sequence elements located within the end of the 3'UTRs. The most
63 well-characterized sequence is the PolyA Signal (PAS) element, a hexameric motif located
64 at ~19nt from the polyadenylation site in the 3'UTR of mature mRNAs. In metazoans, the
65 PAS element is commonly 'AAUAAA', which accounts for more than half of all 3' end
66 processing in eukaryotes (Mangone et al. 2010; Tian and Graber 2012) although
67 alternative forms of the PAS elements exist (Sheets et al. 1990; Mangone et al. 2010).
68 Previous studies have shown that single base substitutions in this sequence reduce the
69 effectiveness of the cleavage and the polyadenylation of the mRNA transcript (Sheets et

70 al. 1990; Chen et al. 1995). However, this canonical sequence is necessary and sufficient
71 for efficient 3' end polyadenylation *in vitro* (Clerici et al. 2018; Sun et al. 2018). A less
72 defined 'GU rich' element is also known to be present downstream of the cleavage site to
73 facilitate the cleavage and polyadenylation steps (Chen et al. 1995). Recently, studies in
74 human cells identified an additional upstream 'UGUA' sequence that is not required for the
75 cleavage process, but acts as a cleavage enhancer in the context of Alternative
76 Polyadenylation (APA) (Zhu et al. 2018).

77 APA is a poorly understood mRNA maturation step that produces mRNAs with
78 different 3'UTR lengths due to the presence of multiple PAS elements within the same
79 3'UTR. The usage of the most upstream element, termed the proximal PAS element, leads
80 to the formation of shorter 3'UTR isoforms while the use of the distal PAS element results
81 in a longer isoform. Importantly, these changes in size may include or exclude regions to
82 which regulatory molecules such as microRNAs (miRNAs) and RNA-binding proteins
83 (RBPs) can bind, substantially impacting gene expression (Matlin et al. 2005; Bartel 2009).
84 While its function in eukaryotes is still not fully understood, a recent study revealed that
85 APA may occur in a tissue-specific manner and, at least in the soil nematode *C. elegans*,
86 is used in specific cellular contexts to evade miRNA-based regulatory networks in a tissue-
87 specific manner (Blazie et al. 2015; Blazie et al. 2017).

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

92 In humans, the Cleavage and Polyadenylation Complex (CPC) is composed of at
93 least 17 members (**Figure 1A**) which immunoprecipitate into at least four large sub-
94 complexes: the Cleavage and Polyadenylation Specificity Factor (CPSF), the Cleavage
95 Stimulation Specificity Factor (CstF), the Cleavage Factor Im (CFIm) and the Cleavage
96 Factor IIm (CFIIm) sub-complexes (**Figure 1A**). CPSF forms the minimal core complex
97 necessary and sufficient to recognize and bind the PAS element of the nascent mRNA *in*
98 *vitro* (Tian and Manley 2017) (**Figure 1A**). In humans, the CPSF sub-complex is
99 composed of CPSF160 (Clerici et al. 2017; Sun et al. 2018), CPSF100 (Mandel et al.
100 2006), CPSF73 (Mandel et al. 2006), CPSF30 (Clerici et al. 2017; Sun et al. 2018), Fip1
101 (Kaufmann et al. 2004) and Wdr33 (Clerici et al. 2017; Sun et al. 2018). Initial experiments
102 assigned CPSF160 with the role of binding the PAS element, but it is now clear that Wdr33
103 and CPSF30 are the proteins that instead contact the PAS directly. CPSF160 has a
104 scaffolding role in this process and keeps this sub-complex structured (Chan et al. 2014).
105 The interaction between members of the CPSF core complex (Wdr33, CPSF30, and
106 CPSF160) and the PAS element was recently revealed using single-particle cryo-EM
107 (Clerici et al. 2017; Sun et al. 2018), showing a unique conformation where the PAS
108 element twists to form an s-shaped structure with a non-canonical pairing between the U3
109 and the A6 in the PAS element (Sun et al. 2018).

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

115 2006), suggesting that this protein's role is to perform mRNA cleavage. Importantly,
116 CPSF73 is also required in the cleavage of pre-histone mRNAs and is recruited on their
117 cleavage site by the U7 SNP (Yang et al. 2009).

118 Fip1 is another member of the CPSF sub-complex. Fip1 interacts with PABPN1,
119 which is the enzyme that performs the polyadenylation reaction on the cleavage site. Fip1
120 preferentially binds U-rich elements in the nascent mRNA and stabilizes the cleavage
121 complex using its arginine-rich RNA-binding domain (Kaufmann et al. 2004). Together with
122 CPSF160 and PABPN1, Fip1 forms a ternary complex *in vitro* (Kaufmann et al. 2004)
123 capable of inhibiting endogenous PABPN1 activity (Zhelkovsky et al. 1998; Helmling et al.
124 2001), suggesting a bridging role for this protein in the complex.

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

136 The CFIm and CFIm sub-complexes are unfortunately less characterized (**Figure**
137 **1A**). The CFIm sub-complex is composed of the CFIm68, CFIm59 and CFIm25 subunits,
138 and it was recently shown to contribute to APA by influencing PAS selection (Martin et al.
139 2012; Hwang et al. 2016). CFIm25 binds a 'UGUA' RNA element upstream of the cleavage
140 site and contributes to 3'processing by recruiting CFIm59 and CFIm68 (Yang et al. 2010;
141 Yang et al. 2011; Zhu et al. 2018).

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

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

149 The round nematode *C. elegans* represents an attractive, novel system to study the
150 cleavage and polyadenylation process *in vivo*. Most of the CPC is conserved between
151 humans and nematodes, including known functional domains and protein interactions
152 (**Figure 1B and Supplemental Figure S1**). *C. elegans* possess the most well-annotated
153 3'UTRome available so far in metazoans, with mapped 3'UTR boundaries for ~26,000
154 distinct *C. elegans* protein-coding genes (Mangone et al. 2010; Jan et al. 2011).

155 The *C. elegans* 3'UTRome was originally developed in 2011 within the
156 modENCODE project (Mangone et al. 2008; Gerstein et al. 2010; Mangone et al. 2010)
157 and represented a milestone in 3'UTR biology since it allowed the Community to study and

158 identify important regulatory elements such as miRNAs and RBPs targets with great
159 precision. A second 3'UTRome was later published using a different mapping pipeline (Jan
160 et al. 2011), confirming most of the previous data such as isoforms numbers, PAS usage,
161 etc. Other datasets were made available later, mostly focusing on tissue-specific 3'UTRs
162 and alternative polyadenylation (Haenni et al. 2012; Blazie et al. 2015; Blazie et al. 2017;
163 Chen et al. 2017; Diag et al. 2018; West et al. 2018).

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

170 To address these and other issues, we developed a novel pipeline to
171 bioinformatically extract 3'UTR data from the entire collection of *C. elegans* transcriptome
172 datasets stored in the public repository SRA trace archive from 2015 to 2018. This blind
173 approach produced a new saturated dataset we named 3'UTRome v2. This updated
174 3'UTRome contains 3'UTR data for 23,159 3'UTR isoforms variants corresponding to
175 14,808 protein-coding genes and is available to the Community as an additional gBrowse
176 track in the *C. elegans* database WormBase (www.WormBase.org) (Stein et al. 2001) and
177 in the 3'UTR-centric database 3'UTRome (www.UTRome.org) (Mangone et al. 2008;
178 Mangone et al. 2010).

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

183

184

RESULTS

185

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

188 To initially gain structural and functional information for the *C. elegans* CPC, we
189 downloaded the protein sequences of the orthologs of the *C. elegans* CPC and aligned
190 them to their human counterparts (**Figure 1B and Supplemental Figure S1**). Based on
191 sequence similarity, *C. elegans* possess orthologs to all the known members of the human
192 CPC, with many peaks of conservation interspersed in the subunits within known
193 interaction domains. The amino acids that make direct contact with PAS elements are also
194 conserved in *C. elegans*; 11 out of the 12 amino acids that form hydrogen bonds and salt
195 bridges with the PAS element (Clerici et al. 2017) are present in both the CPSF30 and
196 WDR33 worm orthologs *cpsf-4* and *pfs-2* (V67^{CPSF30} with V81^{cpsf-4}; K69^{CPSF30} with K83^{cpsf-4};
197 R73^{CPSF30} with R87^{cpsf-4}; E95^{CPSF30} with E109^{cpsf-4}; K77^{CPSF30} with K91^{cpsf-4}; S106^{CPSF30} with
198 S120^{cpsf-4}; N107^{CPSF30} with N121^{cpsf-4}; R54^{WDR33} with R80^{pfs-2}; R47^{WDR33} with R71^{pfs-2};
199 R49^{WDR33} with R73^{pfs-2}) (**Figure 1B and Supplemental Figure S1**). The only exception is
200 Y97^{CPSF30}, which is substituted with a Phenylalanine residue in the worm ortholog. In

201 addition, 9 out of the 10 amino acids in CPSF30 and WDR33 that form the π - π stacking
202 and hydrophobic interactions with the AAUAAA RNA element (Clerici et al. 2017) are also
203 conserved in the CPSF30 and WDR33 worm orthologs *cpsf-4* and *pfs-2* (A1:K69^{CPSF30}
204 with K83^{cpsf-4} and F84^{CPSF30} with F98^{cpsf-4}; A2: H70^{CPSF30} with H84^{cpsf-4}; U3: I156^{WDR33} with
205 I181^{pfs-2}; A4: F112^{CPSF30} with F126^{cpsf-4} and F98^{CPSF30} with F112^{cpsf-4}; A5: F98^{CPSF30} with
206 F112^{cpsf-4}; A6: F153^{WDR33} with F178^{pfs-2}) (**Figure 1B and Supplemental Figure S1**). The
207 only exception is a F43^{WDR33} substitution to a Glycine residue that interacts with A6 in the
208 worm ortholog.

209 CPSF73, the endonuclease that performs the cleavage reaction, has a *C. elegans*
210 ortholog named *cpsf-3*. Both genes are conserved with an overall 57.61% identity that
211 increases to 69.52% in the β -lactamase domain, which is the region required to perform
212 the cleavage reaction (**Figure 1B and Supplemental Figure S1**). Specifically, all eight
213 amino acids shown previously to form the zinc binding site required for the cleavage
214 reaction (Mandel et al. 2006) are also conserved (D75^{CPSF73} with D74^{cpsf-3}; H76^{CPSF73} in
215 H75^{cpsf-3}; H73^{CPSF73} in D72^{cpsf-3}; H396^{CPSF73} with H397^{cpsf-3}; H158^{CPSF73} with H159^{cpsf-3};
216 D179^{CPSF73} with D180^{cpsf-3}; H418^{CPSF73} with H419^{cpsf-3}; E204^{CPSF73} with E205^{cpsf-3}) (**Figure**
217 **1B and Supplemental Figure S1**). This overall similarity is also observed in most of the
218 other members of the *bona fide* *C. elegans* CPC complex (**Supplemental Figure S1**),
219 suggesting similar structure and function.

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

224

225 **An updated 3'end mapping strategy**

226 Next, we used a blind genomic approach to improve the current version of the
227 3'UTRome. We refined a 3'UTR mapping pipeline we previously developed and used in
228 the past (Blazie et al. 2015; Blazie et al. 2017). This approach uses raw transcriptome data
229 as input material to identify and precisely map high-quality 3'UTR end clusters (**Figure 2**
230 **and Supplemental Figure S3**).

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

239 We downloaded a total of 1,094 *C. elegans* transcriptome datasets (~2TB of total
240 raw data)(**Supplemental Table S1**). Most of these datasets have also been used in the
241 past to map polyadenylation sites in *C. elegans*. Our 3'UTR mapping approach extracted
242 from these datasets ~5M unique, high-quality polyA reads, which we then used for cluster
243 preparation and mapping (see Methods). We implemented very restrictive parameters for
244 cluster identification and 3'UTR end mapping to limit the unavoidable noise produced by
245 using such diverse datasets as data sources (**Supplemental Figure S3**). Our approach

246 led us to map 3'UTR clusters with ultra-deep coverage of several magnitudes (average
247 cluster coverage ~220X) (**Figure 2A**), and the identification of 23,159 3'UTR isoforms
248 corresponding to 14,808 protein-coding genes. When compared to the previous
249 3'UTRome v1 dataset (Mangone et al. 2010), we obtained 3'UTR information for an
250 additional 4,638 new protein-coding genes (6,218 3'UTR isoforms) (73% of all protein-
251 coding genes included in the WS250 release) (**Figure 2B-C**).

252

253 **The *C. elegans* 3'UTRome v2**

254 Our approach produced high-quality 3'UTR data for 14,808 *C. elegans* protein-
255 coding genes (**Figure 2B**). The most abundant nucleotide in *C. elegans* 3'UTRs is a
256 Uridine, which accounts for 40% of all nucleotides in 3'UTRs (**Figure 3A Top Left Panel**).
257 Adenosine nucleotides are the second most represented nucleotide class with ~30% of
258 incidence (**Figure 3A Top Left Panel**). Alternative polyadenylation is common but occurs
259 at a lesser extent than what was previously published (Mangone et al. 2010; Jan et al.
260 2011). The majority of protein-coding genes (58%) are transcribed with only one 3'UTR
261 isoform (**Figure 3A Bottom Left Panel**) in contrast with ~61% as it was reported in the
262 past (Mangone et al. 2010; Jan et al. 2011). Genes with two 3'UTR isoforms are notably
263 increased in occurrence when compared with past studies (32% vs 25%), while the
264 occurrence of genes with three or more 3'UTRs is comparable with what was previously
265 found (**Figure 3A Bottom Left Panel**) (Mangone et al. 2010; Jan et al. 2011).

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

268 3'UTR isoforms, suggesting that the preparation of shorter 3'UTR isoforms may be subject
269 to regulation (**Supplemental Figure S4**).

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

283

284 **An RRYRRR motif in 3'UTRs with variant PAS elements**

285 We could not detect any enrichment for the UGUA motif near the cleavage site
286 (**Supplemental Figure S5**), and perhaps this element is either not used in *C. elegans* or
287 the CFIm complex may recognize a variant motif not yet identified in this organism.
288 Importantly, when we aligned the 3' ends of 3'UTRs which contain variant PAS elements,
289 we noticed an enrichment of an 'RRYRRR' motif where the canonical PAS element is

290 generally located; this suggests that in *C. elegans*, an 'RRYRRR' element could be used
291 instead when the AAUAAA hexamer is absent (**Figure 4A**).

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

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

308

309 **An enrichment of Adenosine nucleotide at the cleavage site**

310 We were intrigued by the almost invariable presence of Adenosine nucleotides near
311 the cleavage site. This enrichment becomes more evident when we sort 3'UTRs with

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

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

332 Next, we decided to study the role of the terminal Adenosine nucleotide in the
333 cleavage process. We reasoned that if this Adenosine nucleotide indeed plays any role in
334 the cleavage process, we should be able to alter the position of the mRNA cleavage site

335 by mutating this residue with different Purines or Pyrimidines in the pre mRNAs of selected
336 test genes.

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

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

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

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

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

371

372 **Updated miRANDA prediction in *C. elegans***

373 Next, we used our new UTRome v2 dataset to update MiRanda miRNA target
374 predictions. We downloaded and locally ran the miRanda prediction software (John et al.
375 2004) using our new 3'UTRome v2 as a target dataset. We have produced two sets of
376 predictions; one generic, which contain the entire output produced by the software, and
377 one more restrictive, in which we only output predictions with high scoring and with low E-

378 energy scores. These two tracks have been uploaded in both the 3'UTRome database
379 (Mangone et al. 2008; Mangone et al. 2010) and the WormBase (Stein et al. 2001).

380

381

DISCUSSION

382

383 Here we have used a blind genome-wide approach to refine and study the
384 3'UTRome in the nematode *C. elegans*. We have identified 3'UTR data for 14,808 genes,
385 corresponding to 23,159 3'UTR isoforms, improving their annotation. We now have 3'UTR
386 data for 73% of all protein-coding genes included in the WS250 release. This dataset is
387 not complete, since we could not assign 3'UTR data for the remaining 5,000 protein-
388 coding genes present in WS250. Some of these genes may be transcribed at very low
389 abundance and their mRNA is present below the sensitivity of our approach, or their 3'UTRs
390 data were discarded by our highly stringent filters used during our 3'UTR cluster
391 preparation.

392 Alternative Polyadenylation is widespread in *C. elegans*, with ~42% of genes
393 possessing at least two 3'UTR isoforms (**Figure 3A**). The PAS usage is still most
394 commonly the hexamer 'AAUAAA' which is used to process ~58% of all *C. elegans*
395 3'UTRs (**Figure 3B**). Importantly, we found that the remaining 42% possess a variation of
396 this canonical PAS element which indeed is very similar in chemical composition and
397 contain an 'RRYRRR' motif at the same location where the PAS element is expected
398 (**Figure 4A**). We do not have direct evidence that the CPC recognizes this motif, but since
399 it is so conserved we hypothesize that in *C. elegans* it may provide a docking site in the
400 absence of the canonical AAUAAA site during the cleavage reaction.

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

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

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

421 The concept of mRNAs terminating with an Adenosine nucleotide is not novel.
422 Pioneering work using 269 vertebrate cDNA sequences has shown that ~71% of these
423 genes terminate with a CA nucleotide element (Sheets et al. 1990). These experiments
424 were biochemically validated a few years later using SV40 Late PolyA signal in

425 mammalian cells in a more controlled environment (Chen et al. 1995). These experiments
426 also showed that, at least for the case of this specific 3'UTR, the cleavage could not occur
427 closer than 11nt from the PAS element and no farther of 23nt from it (Chen et al. 1995). In
428 this context, these findings could explain why we do not detect a terminal Adenosine at the
429 cleavage site with our double mutant Y106G6H.9, which is 27nt downstream the PAS
430 element (**Supplemental Figure S8**). Of note, in the case of this gene, the cleavage still
431 occurs at a Purine nucleotide, suggesting that perhaps another terminal Purine can
432 compensate for the absence of an Adenosine nucleotide.

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

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

448 of either 'sensing' this terminal Adenosine nucleotide or is positioned next to it by either
449 other members of the CPC or by a not yet identified factor.

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

455 Of note, while we observed a terminal Adenosine nucleotide in most of the mapped
456 3'UTRs, the Cytosine nucleotide previously identified upstream of the terminal Adenosine
457 in humans is replaced with another Pyrimidine nucleotide in *C. elegans* (Thymidine)
458 (**Figure 3C**), suggesting that other factors may contribute to the cleavage site decision by
459 the CPC in higher eukaryotes.

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

464 The number of predicted miRNA targets is now decreased from 34,186 to 23,160,
465 mostly because several 3'UTR isoforms in the 3'UTRome v1 were discarded in this new
466 3'UTRome v2 release.

467 In conclusion, this new 3'UTR dataset, which we renamed 3'UTRome v2, has been
468 uploaded to the WormBase (Stein et al. 2001) and it is shown as a new track in the
469 JBrowse tool together with updated MiRanda miRNA target predictions. The 3'UTRome v2
470 expands the old 3'UTRome developed within the modENCODE Consortium, and together
471 with updated MiRanda predictions provides the *C. elegans* Community with an important

472 novel resource to investigate the RNA cleavage and polyadenylation reaction, 3'UTR
473 biology and miRNA targeting.

474

475

METHODS

476

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

478 We have downloaded the protein sequences of each known member of the human CPC
479 and used BLAT algorithm to identify *C. elegans* genes with high homology to their human
480 counterparts. We then performed a Protein BLAST analysis using the tools available at the
481 NCBI website to obtain the amino acid sequences for the fly, rat, and mouse orthologs.

482 These amino acid sequences were then aligned using Clustal Omega Multiple Sequence
483 Alignment with standard parameters. At the completion of the analysis, we used the Batch
484 NCBI Conserved Domain Search (Batch CD-Search) against the database CDD- 52910
485 PSSMs using standard parameters to identify the conserved domains across the aligned
486 protein sequences. We then used these results to populate the location of these elements
487 within the alignment shown in **Supplemental Figure S1**.

488

489 **3'UTR mapping pipeline**

490 We have use the SRA toolkit from the NCBI to download raw reads from 1,094
491 transcriptome experiments. The complete list of datasets used in this study is shown in
492 **Supplemental Table S1**. We restricted the analysis to sequences produced from *C.*
493 *elegans* transcriptomes using the Illumina platform and with reads of at least 150nt in
494 length. At the completion of the download step, the files were unzipped and stored in our

495 servers. We then used custom-made Perl scripts to extract reads containing at least 23
496 consecutive Adenosine nucleotides at their 3'end or 23 consecutive Thymidine nucleotides
497 at their 5'end. This filter produced 24,973,286 mappable 3'end reads. We then removed
498 the terminal Adenosine or Thymidine nucleotides from these sequences, converted them
499 to fastq files using the FASTX-Toolkit (CSHL), and mapped them to the WS250 release of
500 the *C. elegans* genome using Bowtie2 algorithm with standard parameters (Langmead and
501 Salzberg 2012). The Bowtie algorithm mapped 7,761,642 reads (31.08%), which were
502 sorted and separated, based on their respective strand origin (positive or negative).

503

504 **Cluster Preparations**

505 PolyA clusters were prepared as follow. We stored the ID, genomic coordinates, and the
506 strand orientation of each mapped read, and used this information throughout the pipeline.
507 The BAM file produced by the aligners were sorted and converted to BED format using
508 SAMtools software (Li et al. 2009). Contiguous genomic coordinates were merged using
509 BEDTools software (Quinlan and Hall 2010) using the following command '`Bedtools`
510 `merge -c 1 -o count -I > tmp.cluster`'. This new file produced the
511 characteristic 'shark fin' graph visible in **Figure 2**. We used several stringed filters to
512 eliminate as much as noise possible. 1) We ignored clusters composed of less than 6
513 reads. 2) We extracted genomic DNA sequences 20nt downstream the end of each
514 cluster. If the number of Adenosine nucleotides was more than 65% in the genomic
515 sequence, we ignored the corresponding cluster and marked it as caused by mispriming
516 during the second strand synthesis in the RT reaction. 3) We ignored clusters overlapping
517 with other clusters in the same orientation by 2nt or less were both ignored. 4) We
518 attached clusters to the closest gene in the same orientation. If no gene could be identified

519 within 2,000nt the cluster was ignored. 5) In cases with multiple 3'UTR isoforms identified,
520 we calculated the frequency of occurrence for each isoform and ignored isoforms
521 occurring at a frequency of less than 1% independently from the number of reads that form
522 this cluster.

523

524 **Plasmid DNA isolation, sequencing and visualization**

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

530

531 **RNAi experiments**

532 RNAi experiments were performed in Standard NGM agar containing 1mM IPTG and 50
533 µg/ml Ampicillin. These plates were seeded with 75 µl of RNAi clone bacteria and allowed
534 to induce for a minimum of 16 hours. 5 N2 *C. elegans* at the L1 stage were aliquoted for
535 each RNAi clone tested. Three days after plating, the progeny was scored for embryonic
536 lethality. Each RNAi experiment was performed in triplicate. The total number of hatched
537 and not matched eggs was the following: *cpsf-1(CPSF160)* n=567; *cpsf-2(CPSF100)*
538 n=557; *cpsf-4(CPSF30)* n=1,251; *cpf-2(CstF64)* n= 652; *cpf-1(CstF50)* n=801; *cfim-*
539 *1(CFIm25)* n=644; *cfim-2(CFIm68)* n=739; *lrp-2(CFIm59)* n=1,120; *symk-1(Symplekin)*
540 n=208; *tag-214(RBBP6)* n=753; *pcf-11(CPF11)* n=428; *clpf-1(CLP1)* n=841.

541

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

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

559

560 **Mutagenesis of 3'UTRs cleavage sites**

561 The mutagenesis reactions to remove the Adenosine nucleotides near the cleavage sites
562 were carried out using the QuikChange Site-Directed Mutagenesis Kit (Agilent). The
563 mutagenesis DNA primers for the site mutation reactions are available in **Supplemental**
564 **Table S2**. Each mutagenesis reaction was followed by DNA digestion using Dpn-1

565 enzyme and transformed in Top10 competent cells (Thermo Fisher Scientific) in agar
566 plates containing 20mg/ μ L of Kanamycin. We validated the nucleotide mutation using
567 Sanger sequencing approach. *Wild type* and mutant 3'UTRs cloned in pDONR P2RP3
568 were then shuttled into destination vectors using the Gateway LR Clonase II Plus Enzyme
569 Mix (Invitrogen, Carlsbad, CA). The finalized destination vectors contained the *C. elegans*
570 pharynx promoter (Pmyo-2) in the first position, a GFP sequence with a mutated STOP
571 codon in the second position, and the *wt* or mutant 3'UTRs used in this study in the third
572 position. The resultant recombined constructs were then transformed in Top10 competent
573 cells (Thermo Fisher Scientific) and plated on 10mg/ μ L Ampicillin plates overnight. The
574 success of the recombination reaction was confirmed using Sanger sequencing with the
575 M13F DNA primer.

576

577 **Preparation of transgenic worm lines**

578 Eg6699 strain worms were kindly provided by Christian Frokjaer-Jensen (Frokjaer-Jensen
579 et al. 2008). These worm strains were maintained at 18°C on nematode growth media
580 (NGM) agar plates and propagated on plates seeded with OP50-1 bacteria. To
581 synchronize worms for injections, Eg6699 worms were bleached with bleaching solution (1
582 M NaOH) four days before injections. Each construct was mixed with an injection master
583 mix containing pCFJ601 (25 ng/ μ l), pgH8 (10 ng/ μ l), and pCFJ104 (5 ng/ μ l) vectors.
584 Injection needles were loaded with the injection mixture and mounted to the Leica
585 DMI300B microscope. The needle was pressurized with 22 psi through the FemtoJet
586 (Eppendorf). Young adult Eg6699 worms were picked onto an agarose pad covered with
587 mineral oil on a glass coverslip. Injected worms were rescued onto an NGM plate and

588 rinsed with M9 buffer. Two days post-injections, the F1 progeny were screened with a
589 Leica DMI3000B microscope for both *unc-119* rescues and expression of the red
590 fluorescence produced by the co-injection marker and then isolated onto individual plates.
591 These worms were allowed to lay eggs, and then the F2 progeny was screened for
592 fluorescence. Once 75% of the progeny on a single plate were transgenic, the strains were
593 used for further experimentation.

594

595 **Worm genotype validation**

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

604

605 **Detection of the 3'UTR cleavage skipping**

606 Total RNA was extracted from transgenic strains using the Direct-zol RNA MiniPrep Plus
607 kit (RPI) according to the manufacturer's instructions. We tested approximately 10
608 independent *wt* and mutant clones for each 3'UTR. Approximately 50 μ L of worm pellet
609 was used for extraction. cDNA was synthesized using a reverse transcription reaction
610 using Superscript II enzyme (Invitrogen). The first strand reaction was performed using a

611 reverse poly dT DNA primer containing two anchors and the attB Gateway BP
612 recombination element (Invitrogen). The second strand of the cDNA was synthesized
613 using a PCR with HiFi taq polymerase (Thermo Fisher Scientific) and the forward DNA
614 primer containing the pDONR P2RP3 Gateway element (Invitrogen), which binds to GFP
615 and the same reverse poly dT DNA primer used in the first strand reaction. The BP
616 Gateway kit (Invitrogen) was once again used to clone the cDNA which contains the polyA
617 tail into pDONR P2RP3. These constructs were then transfected into Top10 competent
618 cells (Thermo Fisher Scientific) and plated on agar plates containing 20mg/μL of
619 Kanamycin. About 8-10 colonies were then sequenced with Sanger sequencing using the
620 M13F DNA primer to map the location of the cleavage site.

621

622 **Updated MiRanda Predictions**

623 We downloaded a complete list of *C. elegans* miRNAs from miRBase (Griffiths-Jones et al.
624 2006) and the miRanda algorithm v3.3a (John et al. 2004) from the microrna.org website.
625 We queried the 3'UTRome v2 with the miRanda algorithm using both standard and
626 stringent parameters. The stringent query used was '-strict -sc -1.2'. The standard query
627 produced 58,330 putative miRNA targets; the stringent query produced 12,136 putative
628 miRNA targets. Both these predictions are included in WormBase (Stein et al. 2001) as
629 individual tracks.

630

631 **Homology model building**

632 Homology modeling was performed using SWISS_MODEL (Waterhouse et al.
633 2018) with a matched template of human CPSF160-WDR33-CPSF30 complex

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

636

637 **Data Availability**

638 Strains and plasmids are available upon request. The authors affirm that all data
639 necessary for confirming the conclusions of the article are present within the article,
640 figures and supplemental figures, and tables and supplemental tables. The results
641 of our analyses are available in the WormBase (www.WormBase.org) (Stein et al.
642 2001) and in our 3'UTR-centric website www.UTRome.org.

643

644 **Author Contribution**

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

654

655 **Funding**

656 This work was supported by the National Institute of Health grant number
657 1R01GM118796.

658

659 **Conflict of Interest**

660 The authors declare that they have no competing interests.

661

662 **Acknowledgements**

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

665

666

FIGURE LEGENDS

667

668 **Figure 1.** The *C. elegans* members of the Cleavage and Polyadenylation Complex (CPC).

669 A) The CPC is composed of at least 4 independent subcomplexes named Cleavage and

670 Polyadenylation Specificity Complex (Blue), which canonically recognizes the PAS

671 hexamer 'AAUAAA'; the Cleavage Stimulation Factor Complex (Green), which binds

672 downstream of the cleavage site to GU rich elements; and the Cleavage Factor CFIm

673 (Red) and CFIIIm (Orange) Complexes. CFIm recognizes the element 'UGUA' located

674 upstream of the PAS element. Other known required factors are the PolyA Polymerase

675 enzyme, the scaffolding member Symplekin and RBBP6. The name of the *C. elegans*

676 orthologs are shown in parenthesis. B) The human and *C. elegans* CPSF subcomplexes

677 are similar in amino acid composition and structure. 2-species alignments between several

678 members of the human and *C. elegans* CPSF members. Amino acids 100% conserved

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

685

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

693

694 **Figure 3.** The worm 3'UTRome v2. A) Top Left Panel. Nucleotide composition of
695 3'UTRs in the 3'UTRome v2. Uridine is the most abundant nucleotide within 3'UTRs
696 for *C. elegans*. Bottom Left Panel. The number of 3'UTR isoforms in each gene.
697 42% of the genes in the 3'UTRome v2 possess multiple 3'UTR isoforms. Top Right
698 Panel. 3'UTR length distribution in genes expressed with one, two, or three or more
699 3'UTR isoforms. The median 3'UTR length across these datasets is 122nt. Genes
700 with multiple 3'UTR isoforms are on average longer than genes with one 3'UTR
701 isoform. Bottom Right Panel. Median 3'UTR length in genes with Canonical (C) or

702 Variant (V) PAS elements. There is a slight increase in 3'UTR length in genes with
703 variant PAS elements when compared to those with canonical PAS elements. This
704 variation is still detected when increasing the stringency of the density of the
705 clusters (cd) used in this analysis. B) PAS element usage in 3'UTRs. 58.4% of
706 3'UTRs use the canonical PAS element 'AAUAAA' while the most common variant
707 PAS element is the hexamer 'AAUGAA', which occurs in 11% of genes. The
708 distribution of canonical PAS elements within 3'UTRs. The average distance from
709 the PAS element to the cleavage site is 18nt. C) Alignment of 3'UTRs at the
710 cleavage site. This alignment in genes with both canonical and variant PAS
711 elements reveals a region between the PAS element and the cleavage site we
712 renamed the *buffer region* in which cleavage rarely occurs. The most abundant
713 nucleotide at the cleavage site is an Adenosine nucleotide preceded by a
714 Thymidine nucleotide.

715

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

725 2018; Sun et al. 2018) to the worm CPSF core complex: cpsf-1 (CPSF-160) in blue,
726 pfs-2 (Wdr33) in pink, and cpsf-4 (CPSF30) in green. The PAS element binding
727 pocket can be fitted into the homology model. The PAS element of the RNA is
728 represented in yellow. The size and the selectivity of the nucleotide binding pocket
729 can fit other nucleotides as long as the motif is RRYRRR.

730

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

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

754

755

REFERENCES

756

- 757 Bai Y, Auperin TC, Chou CY, Chang GG, Manley JL, Tong L. 2007. Crystal structure of murine CstF-77: dimeric
758 association and implications for polyadenylation of mRNA precursors. *Mol Cell* **25**: 863-875.
- 759 Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* **136**: 215-233.
- 760 Bartel DP. 2018. Metazoan MicroRNAs. *Cell* **173**: 20-51.
- 761 Blazie SM, Babb C, Wilky H, Rawls A, Park JG, Mangone M. 2015. Comparative RNA-Seq analysis reveals
762 pervasive tissue-specific alternative polyadenylation in *Caenorhabditis elegans* intestine and
763 muscles. *BMC Biol* **13**: 4.
- 764 Blazie SM, Geissel HC, Wilky H, Joshi R, Newbern J, Mangone M. 2017. Alternative Polyadenylation Directs
765 Tissue-Specific miRNA Targeting in *Caenorhabditis elegans* Somatic Tissues. *Genetics* **206**: 757-774.
- 766 Chan SL, Huppertz I, Yao C, Weng L, Moresco JJ, Yates JR, 3rd, Ule J, Manley JL, Shi Y. 2014. CPSF30 and
767 Wdr33 directly bind to AAUAAA in mammalian mRNA 3' processing. *Genes Dev* **28**: 2370-2380.
- 768 Chen F, Chisholm AD, Jin Y. 2017. Tissue-specific regulation of alternative polyadenylation represses
769 expression of a neuronal ankyrin isoform in *C. elegans* epidermal development. *Development* **144**:
770 698-707.
- 771 Chen F, MacDonald CC, Wilusz J. 1995. Cleavage site determinants in the mammalian polyadenylation
772 signal. *Nucleic Acids Res* **23**: 2614-2620.
- 773 Clerici M, Faini M, Aebersold R, Jinek M. 2017. Structural insights into the assembly and polyA signal
774 recognition mechanism of the human CPSF complex. *Elife* **6**.
- 775 Clerici M, Faini M, Muckenfuss LM, Aebersold R, Jinek M. 2018. Structural basis of AAUAAA polyadenylation
776 signal recognition by the human CPSF complex. *Nat Struct Mol Biol* **25**: 135-138.
- 777 de Vries H, Ruegsegger U, Hubner W, Friedlein A, Langen H, Keller W. 2000. Human pre-mRNA cleavage
778 factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. *EMBO J*
779 **19**: 5895-5904.
- 780 Diag A, Schilling M, Klironomos F, Ayoub S, Rajewsky N. 2018. Spatiotemporal m(i)RNA Architecture and 3'
781 UTR Regulation in the *C. elegans* Germline. *Dev Cell* **47**: 785-800 e788.
- 782 Frokjaer-Jensen C, Davis MW, Hopkins CE, Newman BJ, Thummel JM, Olesen SP, Grunnet M, Jorgensen EM.
783 2008. Single-copy insertion of transgenes in *Caenorhabditis elegans*. *Nat Genet* **40**: 1375-1383.

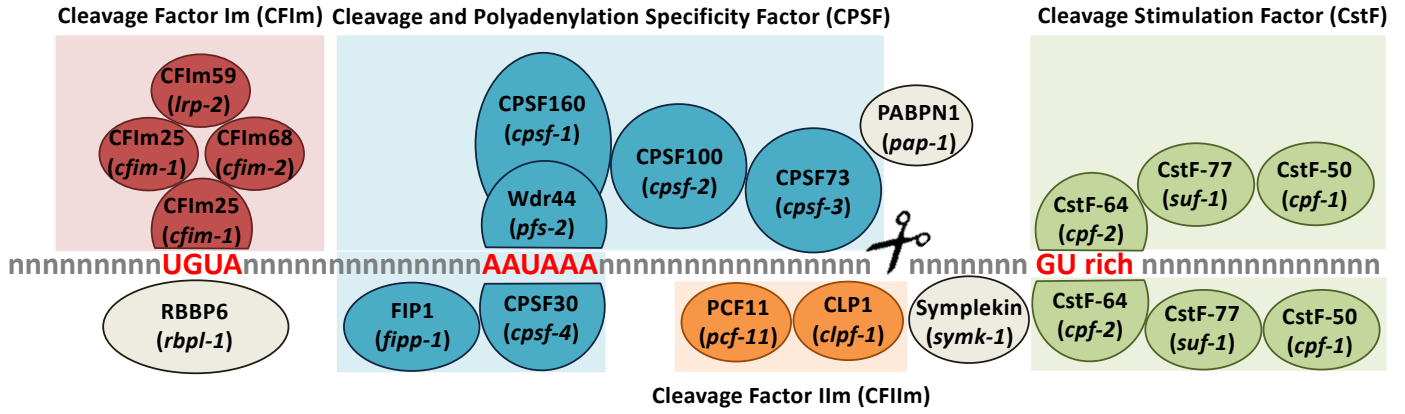
- 784 Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K et
785 al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project.
786 *Science* **330**: 1775-1787.
- 787 Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, Ferrin TE. 2018. UCSF ChimeraX:
788 Meeting modern challenges in visualization and analysis. *Protein Sci* **27**: 14-25.
- 789 Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences,
790 targets and gene nomenclature. *Nucleic Acids Res* **34**: D140-144.
- 791 Haenni S, Ji Z, Hoque M, Rust N, Sharpe H, Eberhard R, Browne C, Hengartner MO, Mellor J, Tian B et al.
792 2012. Analysis of *C. elegans* intestinal gene expression and polyadenylation by fluorescence-
793 activated nuclei sorting and 3'-end-seq. *Nucleic Acids Res* **40**: 6304-6318.
- 794 Helmling S, Zhelkovsky A, Moore CL. 2001. Fip1 regulates the activity of Poly(A) polymerase through
795 multiple interactions. *Mol Cell Biol* **21**: 2026-2037.
- 796 Hwang HW, Park CY, Goodarzi H, Fak JJ, Mele A, Moore MJ, Saito Y, Darnell RB. 2016. PAPERCLIP Identifies
797 MicroRNA Targets and a Role of CstF64/64tau in Promoting Non-canonical poly(A) Site Usage. *Cell*
798 *Rep* **15**: 423-435.
- 799 Jan CH, Friedman RC, Ruby JG, Bartel DP. 2011. Formation, regulation and evolution of *Caenorhabditis*
800 *elegans* 3'UTRs. *Nature* **469**: 97-101.
- 801 John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. 2004. Human MicroRNA targets. *PLoS Biol* **2**:
802 e363.
- 803 Kaufmann I, Martin G, Friedlein A, Langen H, Keller W. 2004. Human Fip1 is a subunit of CPSF that binds to
804 U-rich RNA elements and stimulates poly(A) polymerase. *EMBO J* **23**: 616-626.
- 805 Kuhn U, Wahle E. 2004. Structure and function of poly(A) binding proteins. *Biochim Biophys Acta* **1678**: 67-
806 84.
- 807 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.
- 808 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project
809 Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:
810 2078-2079.
- 811 Mandel CR, Kaneko S, Zhang H, Gebauer D, Vethantham V, Manley JL, Tong L. 2006. Polyadenylation factor
812 CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature* **444**: 953-956.
- 813 Mangone M, Macmenamin P, Zegar C, Piano F, Gunsalus KC. 2008. UTRome.org: a platform for 3'UTR
814 biology in *C. elegans*. *Nucleic Acids Res* **36**: D57-62.
- 815 Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein
816 MR, Khivansara V et al. 2010. The landscape of *C. elegans* 3'UTRs. *Science* **329**: 432-435.
- 817 Martin G, Gruber AR, Keller W, Zavolan M. 2012. Genome-wide analysis of pre-mRNA 3' end processing
818 reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Rep* **1**: 753-
819 763.
- 820 Matlin AJ, Clark F, Smith CW. 2005. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol*
821 *Cell Biol* **6**: 386-398.
- 822 Mayr C, Bartel DP. 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation
823 activates oncogenes in cancer cells. *Cell* **138**: 673-684.
- 824 Perez Canadillas JM, Varani G. 2003. Recognition of GU-rich polyadenylation regulatory elements by human
825 CstF-64 protein. *EMBO J* **22**: 2821-2830.
- 826 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
827 *Bioinformatics* **26**: 841-842.
- 828 Ryan K, Calvo O, Manley JL. 2004. Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing
829 endonuclease. *RNA* **10**: 565-573.

- 830 Schafer P, Tuting C, Schonemann L, Kuhn U, Treiber T, Treiber N, Ihling C, Graber A, Keller W, Meister G et
831 al. 2018. Reconstitution of mammalian cleavage factor II involved in 3' processing of mRNA
832 precursors. *RNA* **24**: 1721-1737.
- 833 Sheets MD, Ogg SC, Wickens MP. 1990. Point mutations in AAUAAA and the poly (A) addition site: effects
834 on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res* **18**: 5799-
835 5805.
- 836 Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J. 2001. WormBase: network access to the genome
837 and biology of *Caenorhabditis elegans*. *Nucleic Acids Res* **29**: 82-86.
- 838 Sun Y, Zhang Y, Hamilton K, Manley JL, Shi Y, Walz T, Tong L. 2018. Molecular basis for the recognition of
839 the human AAUAAA polyadenylation signal. *Proc Natl Acad Sci U S A* **115**: E1419-E1428.
- 840 Takagaki Y, Manley JL. 2000. Complex protein interactions within the human polyadenylation machinery
841 identify a novel component. *Mol Cell Biol* **20**: 1515-1525.
- 842 Tian B, Graber JH. 2012. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip Rev RNA* **3**:
843 385-396.
- 844 Tian B, Manley JL. 2017. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18**: 18-30.
- 845 Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C,
846 Bordoli L et al. 2018. SWISS-MODEL: homology modelling of protein structures and complexes.
847 *Nucleic Acids Res* **46**: W296-W303.
- 848 West SM, Mecnas D, Gutwein M, Aristizabal-Corrales D, Piano F, Gunsalus KC. 2018. Developmental
849 dynamics of gene expression and alternative polyadenylation in the *Caenorhabditis elegans*
850 germline. *Genome Biol* **19**: 8.
- 851 Yang Q, Coseno M, Gilmartin GM, Doublet S. 2011. Crystal structure of a human cleavage factor
852 CFI(m)25/CFI(m)68/RNA complex provides an insight into poly(A) site recognition and RNA looping.
853 *Structure* **19**: 368-377.
- 854 Yang Q, Gilmartin GM, Doublet S. 2010. Structural basis of UGUA recognition by the Nudix protein CFI(m)25
855 and implications for a regulatory role in mRNA 3' processing. *Proc Natl Acad Sci U S A* **107**: 10062-
856 10067.
- 857 Yang W, Hsu PL, Yang F, Song JE, Varani G. 2018. Reconstitution of the CstF complex unveils a regulatory
858 role for CstF-50 in recognition of 3'-end processing signals. *Nucleic Acids Res* **46**: 493-503.
- 859 Yang XC, Sullivan KD, Marzluff WF, Dominski Z. 2009. Studies of the 5' exonuclease and endonuclease
860 activities of CPSF-73 in histone pre-mRNA processing. *Mol Cell Biol* **29**: 31-42.
- 861 Zhelkovsky A, Helmling S, Moore C. 1998. Processivity of the *Saccharomyces cerevisiae* poly(A) polymerase
862 requires interactions at the carboxyl-terminal RNA binding domain. *Mol Cell Biol* **18**: 5942-5951.
- 863 Zhu Y, Wang X, Forouzmand E, Jeong J, Qiao F, Sowd GA, Engelman AN, Xie X, Hertel KJ, Shi Y. 2018.
864 Molecular Mechanisms for CFI_m-Mediated Regulation of mRNA Alternative Polyadenylation. *Mol*
865 *Cell* **69**: 62-74 e64.

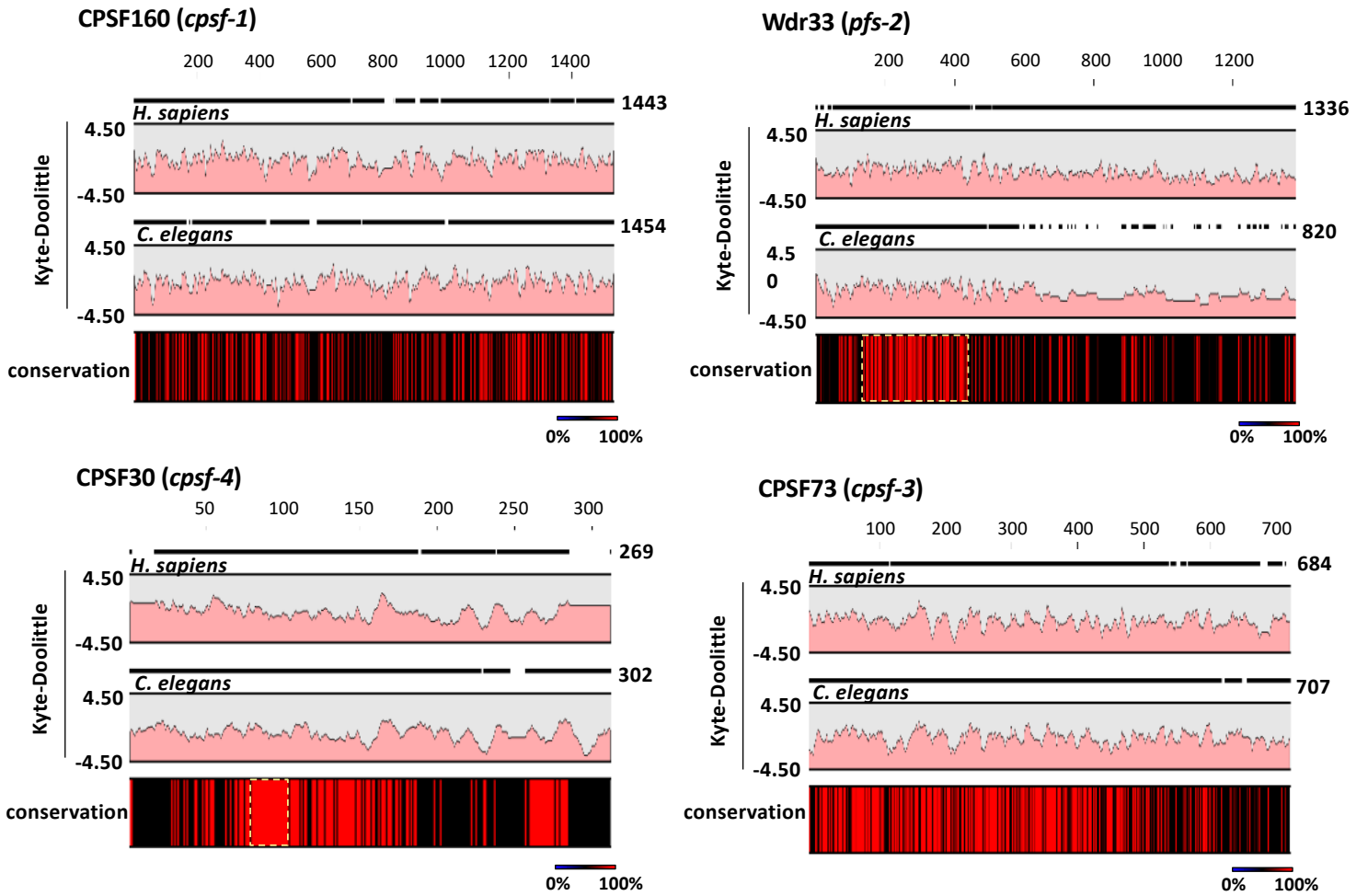
866

Figure 1

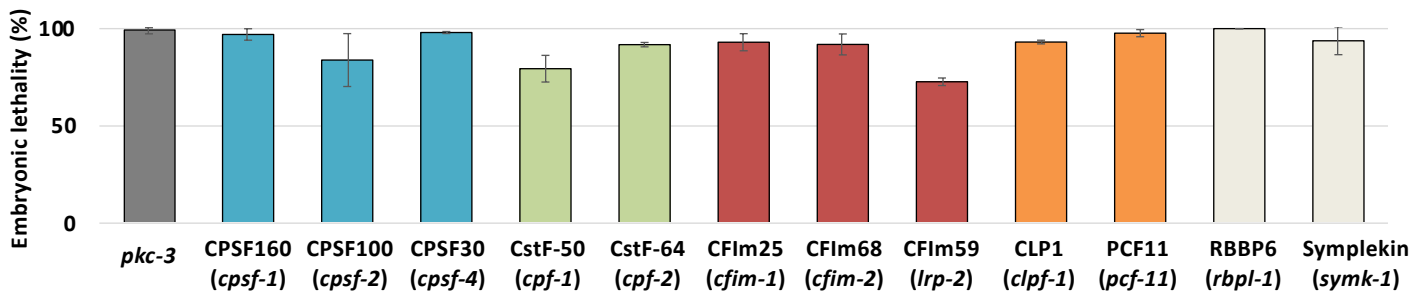
A



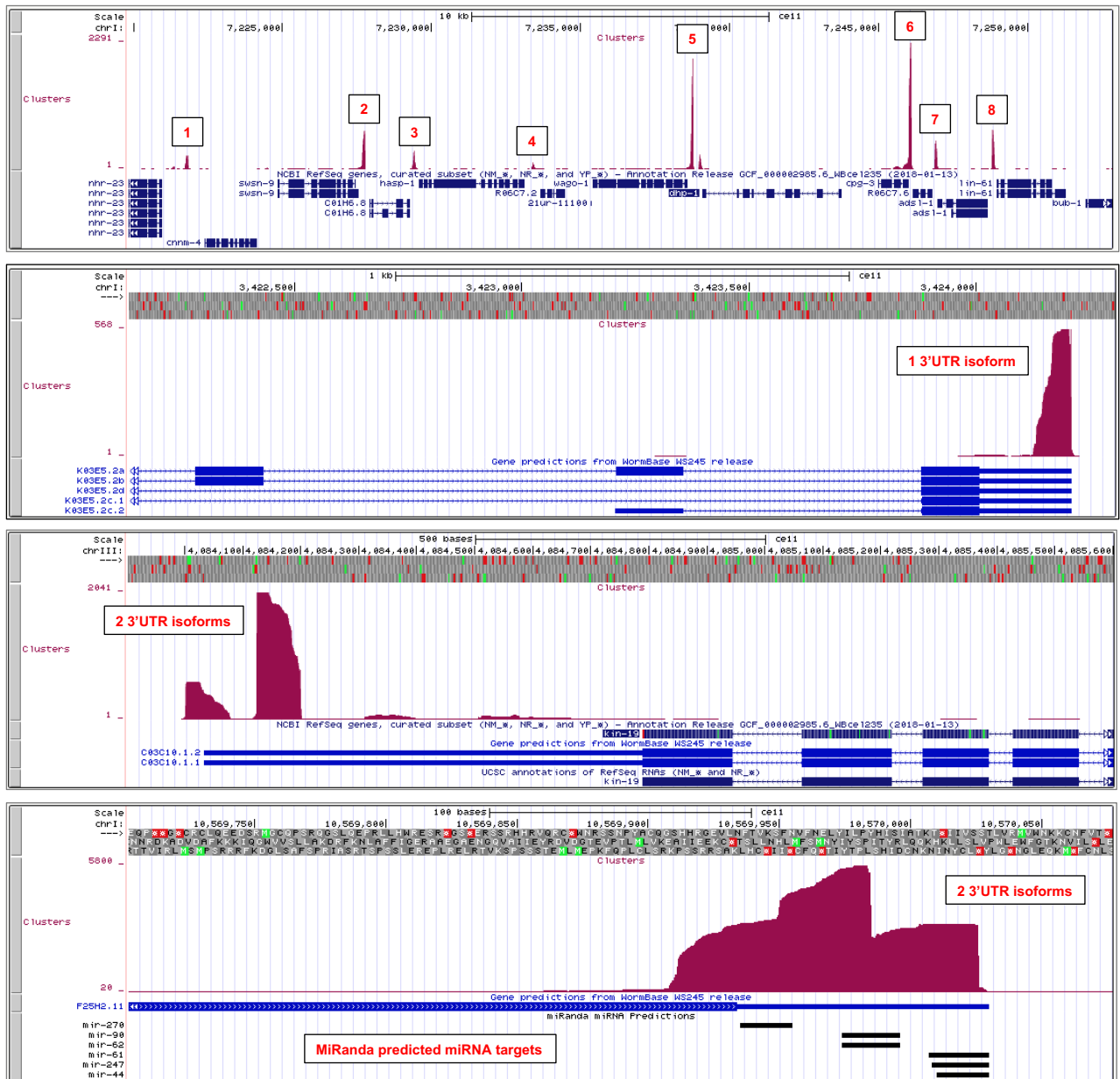
B



C



A

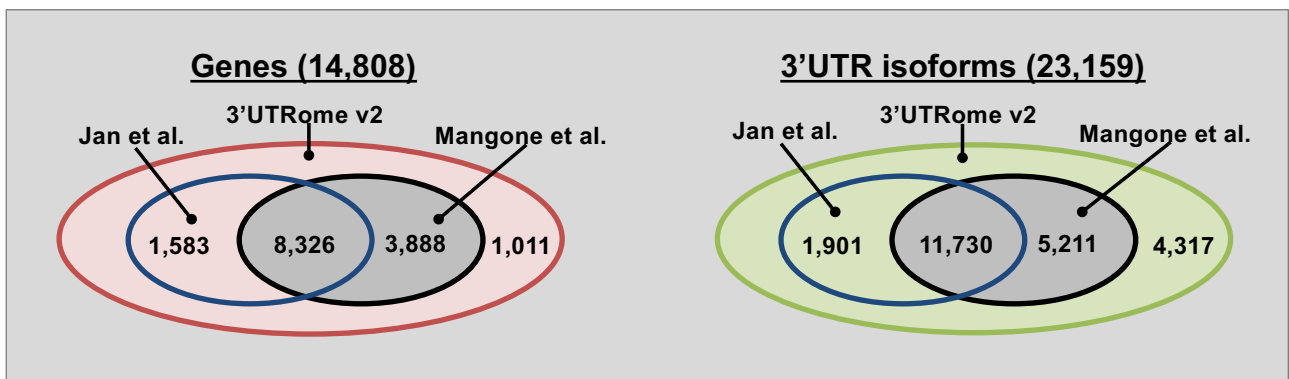


B

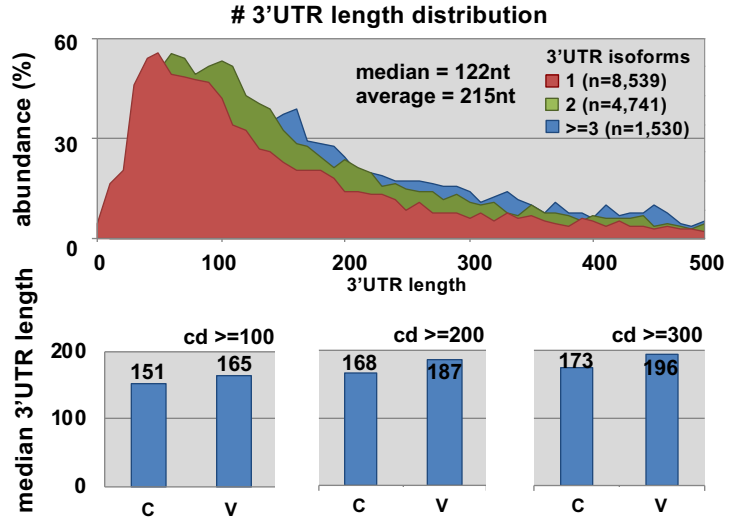
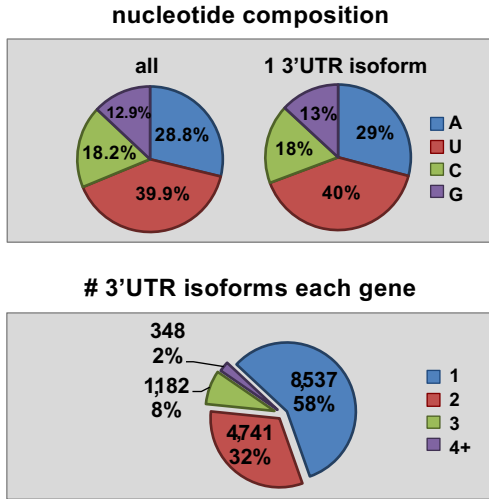
# 3'UTRs in genes (isoforms)			# reads used	
total	in 3'UTRome v1	new	mapped	in each cluster
14,808 (23,159)	11,377 (16,941)	4,638 (6,218)	4,838,905	220 (median)

clusters >=10%

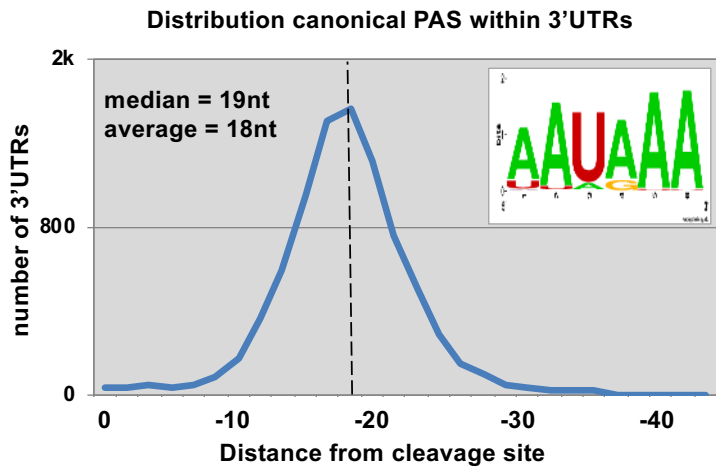
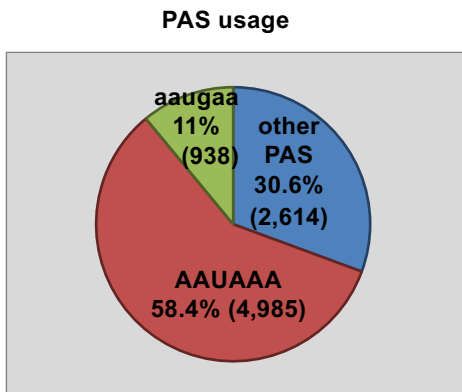
C



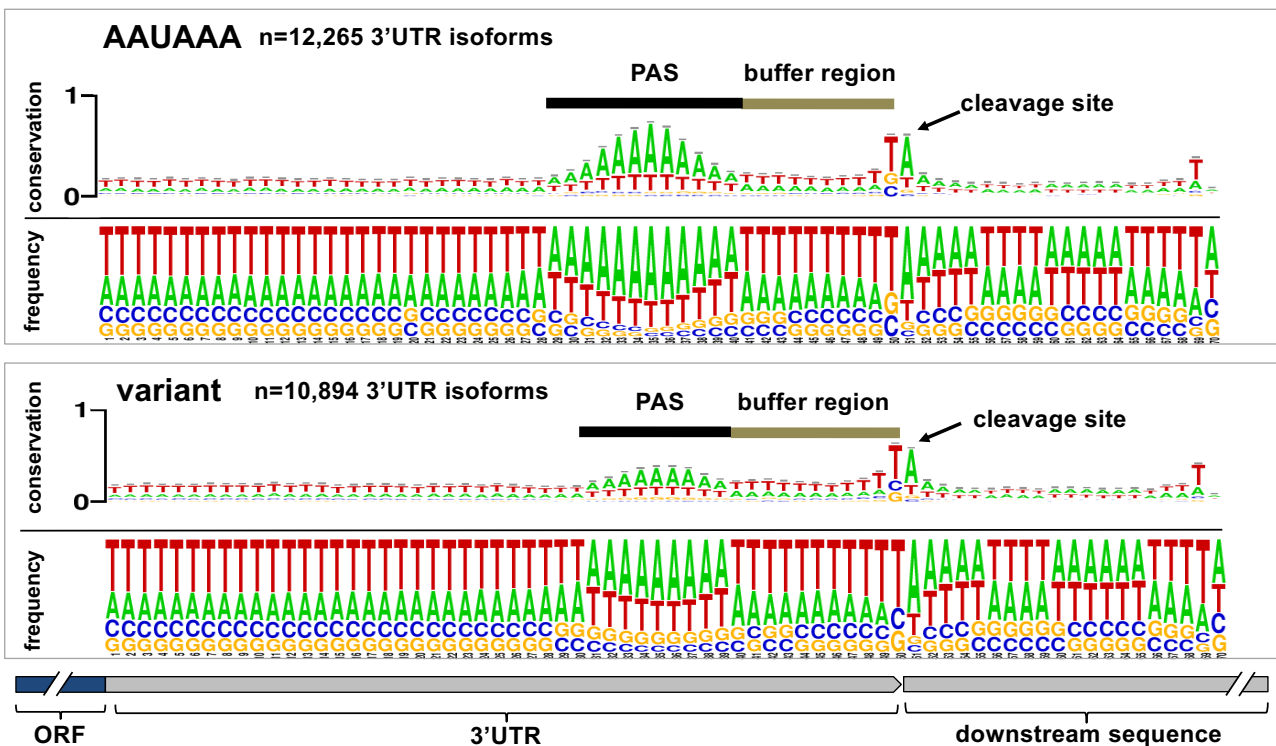
A



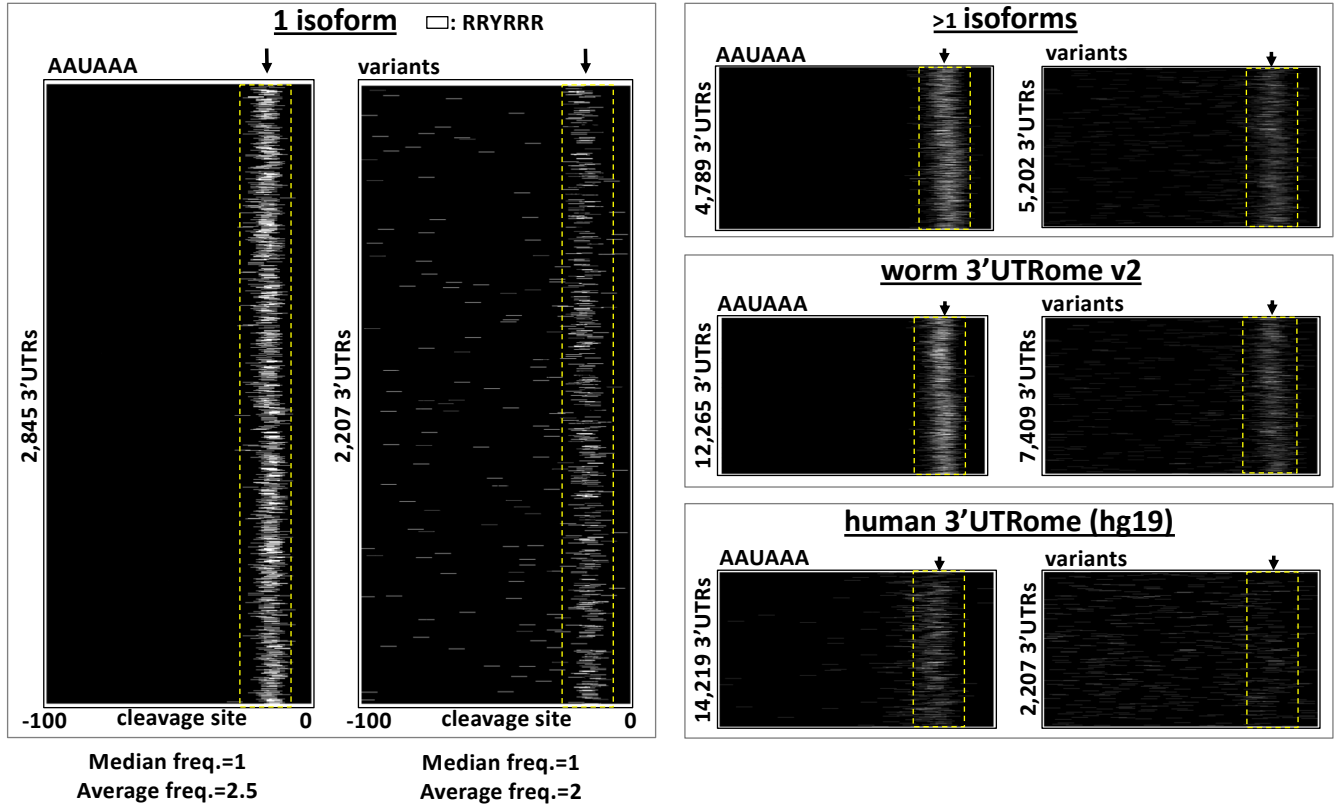
B



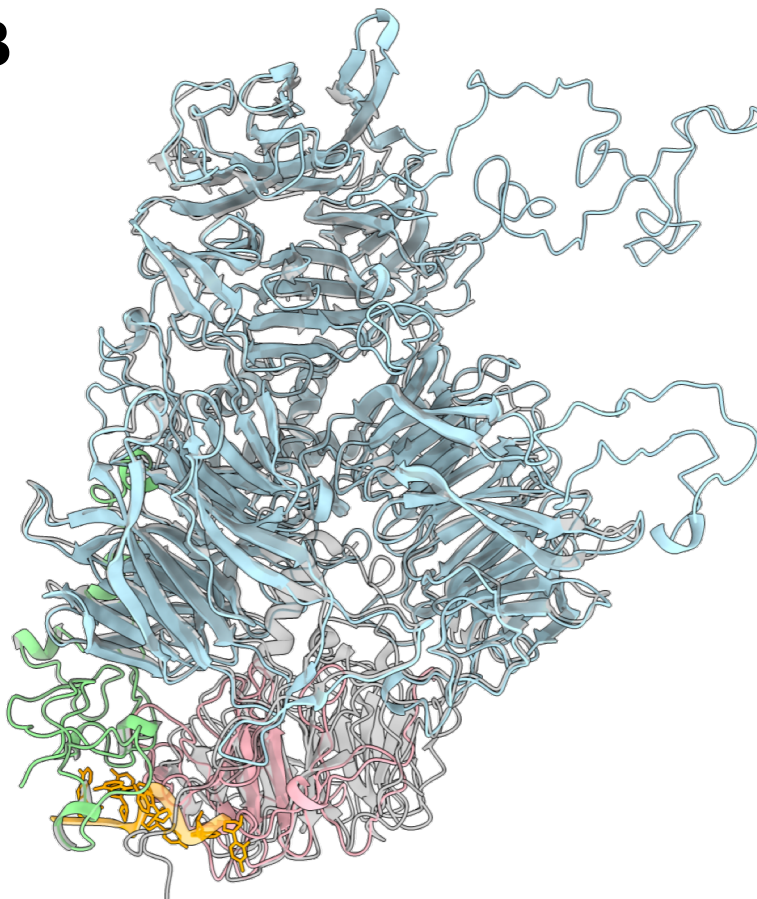
C



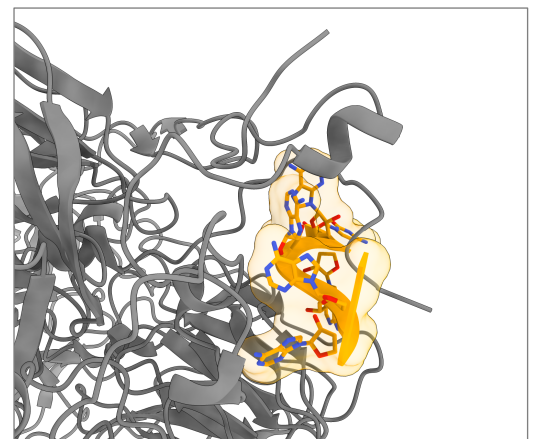
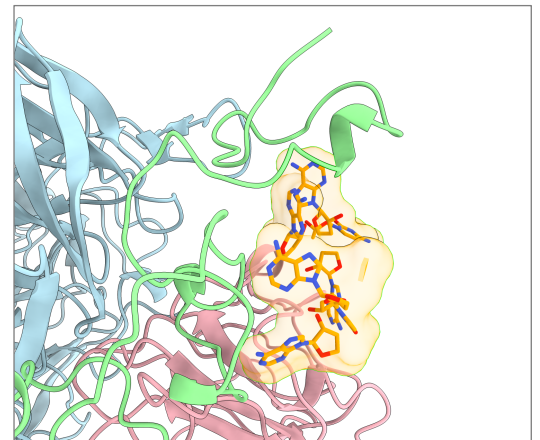
A



B



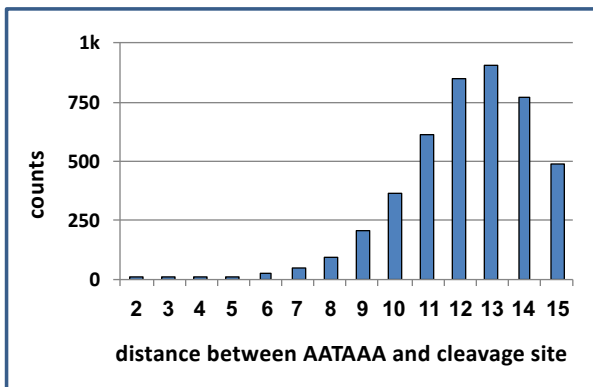
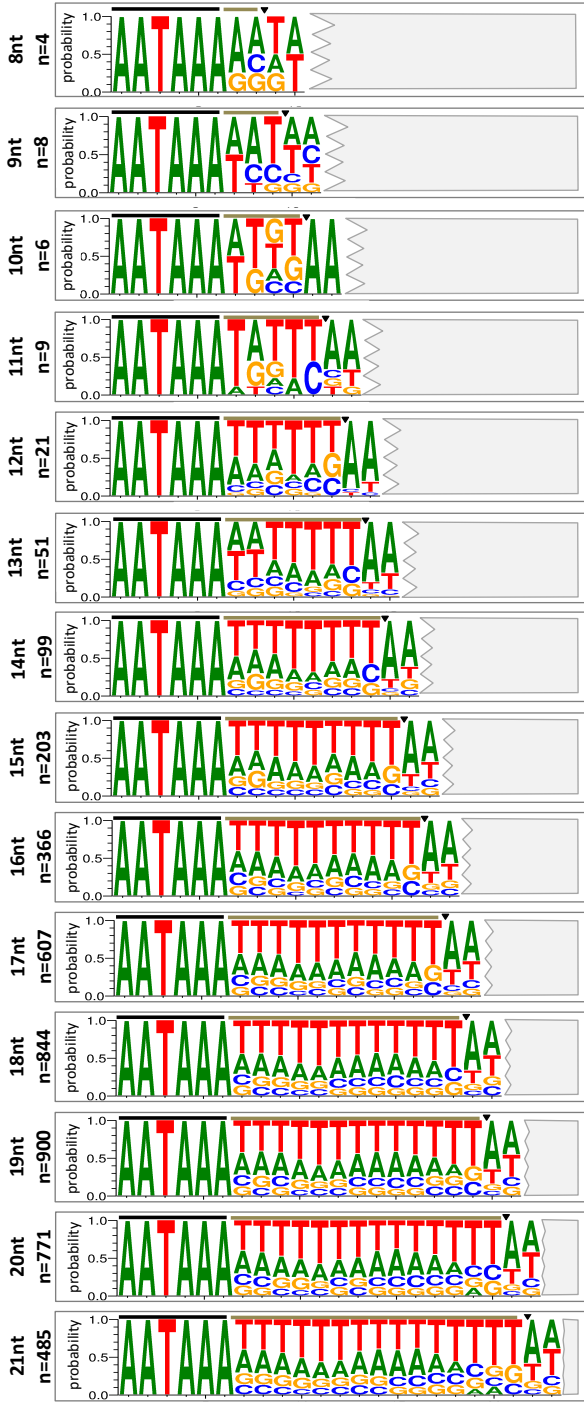
Homology model (*C. elegans*)



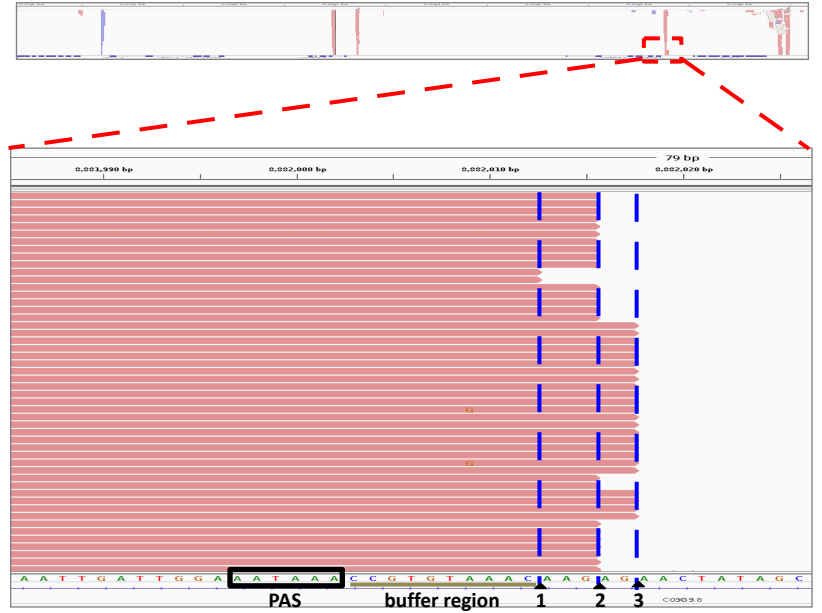
Cryo-EM structure (*Homo sapiens*)

A

genes w/ unique 3'UTRs with canonical PAS=4,374

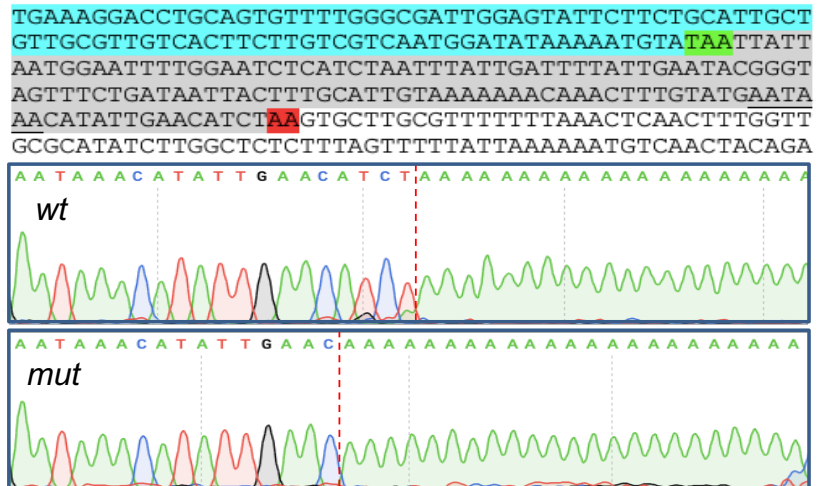


B



C

M03A1.3



Y106G6H.9

