Cell Reports
Resource

Cell PRESS
Open ACCESS

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

Georges Martin,[1,2] Andreas R. Gruber,[1,2] Walter Keller,[1,*] and Mihaela Zavolan[1,*]
[1]Computational and Systems Biology, Biozentrum, University of Basel, CH-4056 Basel, Switzerland
[2]These authors contributed equally to this work
*Correspondence: walter.keller@unibas.ch (W.K.), mihaela.zavolan@unibas.ch (M.Z.)

DOI 10.1016/j.celrep.2012.05.003

SUMMARY

Through alternative polyadenylation, human mRNAs acquire longer or shorter 3′ untranslated regions, the latter typically associated with higher transcript stability and increased protein production. To understand the dynamics of polyadenylation site usage, we performed transcriptome-wide mapping of both binding sites of 3′ end processing factors CPSF-160, CPSF-100, CPSF-73, CPSF-30, Fip1, CstF-64, CstF-64τ, CF I_m25, CF I_m59, and CF I_m68 and 3′ end processing sites in HEK293 cells. We found that although binding sites of these factors generally cluster around the poly(A) sites most frequently used in cleavage, CstF-64/CstF-64τ and CFI_m proteins have much higher positional specificity compared to CPSF components. Knockdown of CF I_m68 induced a systematic use of proximal polyadenylation sites, indicating that changes in relative abundance of a single 3′ end processing factor can modulate the length of 3′ untranslated regions across the transcriptome and suggesting a mechanism behind the previously observed increase in tumor cell invasiveness upon CF I_m68 knockdown.

INTRODUCTION

Expression of eukaryotic genes proceeds through numerous steps, including transcription, addition of a 7-methyl guanosine cap to the 5′-end (reviewed in Shatkin and Manley, 2000), splicing (Carrillo Oesterreich et al., 2011), selection of a cleavage site (Di Giammartino et al., 2011; Proudfoot, 2011), and in most cases, poly(A) tail addition. These processes typically involve multiple RNA-binding proteins and large ribonucleoprotein complexes that not only control the maturation of the mRNA in the nucleus, but also the stability, transport, editing, and finally the translation of mRNAs into proteins (Martin and Ephrussi, 2009; Moore, 2005; Sonenberg and Hinnebusch, 2009).

An affinity-purified mammalian 3′ end processing apparatus was found to contain approximately 16 core proteins, representing the cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), cleavage factors I and II (CF I_m and CF II_m), and poly(A) polymerase (PAP; for reviews see Mandel et al., 2008; Martin and Keller, 2007; Millevoi and Vagner, 2010), and ~70 more loosely bound polypeptides (Shi et al., 2009). CPSF consists of the six polypeptides CPSF-30, -73, -100, and -160, Fip1 (Kaufmann et al., 2004), and WDR33 (Shi et al., 2009). The polyadenylation signal, most frequently AAUAAA (Wickens, 1990), is recognized by CPSF-160, whereas CPSF-73 appears to be the nuclease responsible for the transcript cleavage (Mandel et al., 2006).

CF I_m is a tetramer (Yang et al., 2011) composed of two 25 kDa subunits, CF I_m25 (Coseno et al., 2008; Rüegsegger et al., 1996), that contact UGUA motifs in the pre-mRNAs (Brown and Gilmartin, 2003; Yang et al., 2011), and two larger polypeptides of either 59 or 68 kDa, CF I_m59 and CF I_m68, that can be modified by methylation (Martin et al., 2010). Although related in sequence, CF I_m68 and CF I_m59 probably differ in function because CF I_m59 but not CF I_m68 was found to interact with the splicing factor U2AF65 (Millevoi et al., 2006) whereas the RS-like domain of CF I_m68 interacts with SR proteins (Dettwiler et al., 2004). CF I_m68 was further found to shuttle between nucleus and cytoplasm during cell cycle (Cardinale et al., 2007) and to participate in mRNA export (Ruepp et al., 2009). Sequence-specific binding of the CF I_m68 subunit to motifs near the cleavage site of its own pre-mRNA has been reported to suppress cleavage in a regulatory loop (Brown and Gilmartin, 2003). Phosphorylation of the CF I_m/CF II_m subcomplex (the latter composed of the 48 kDa hClp1 and the 173 kDa hPcf11 [de Vries et al., 2000]) at serines or threonines is required to render the cleavage complex active (Ryan, 2007).

The CstF complex, consisting of polypeptides of 50, 64, and 77 kDa, has been implicated in the selection of poly(A) sites. The 64 kDa subunit contains an RNA recognition motif (RRM) and binds preferentially to U- or U/G-rich sequences downstream of the cleavage site (Beyer et al., 1997; Takagaki and Manley, 1997). Its overexpression in mouse primary B cells was reported to switch IgM heavy chain expression from the membrane-bound (μm) to the secreted form (μs) via the selection of an alternative poly(A) site (Takagaki et al., 1996), though these findings have been challenged by a subsequent study (Martincic et al., 1998). The upstream cleavage product finally acquires a tail of approximately 250 adenosine residues through the action of a PAP (reviewed in Martin and Keller, 2007).

Recently, it has been observed that in proliferating cells, hundreds of genes use upstream polyadenylation sites to

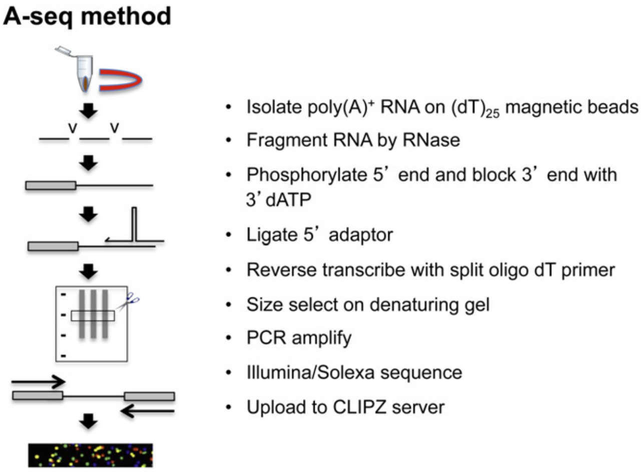
A-seq method
Isolate poly(A)+ RNA on (dT)25 magnetic beads
Fragment RNA by RNase
Phosphorylate 5' end and block 3' end with 3' dATP
Ligate 5' adaptor
Reverse transcribe with split oligo dT primer
Size select on denaturing gel
PCR amplify
Illumina/Solexa sequence
Upload to CLIPZ server

Figure 1. Outline of the A-seq Method Used to Map Binding Sites of 3′ End Cleavage Sites

Further details are provided in Experimental Procedures. See also Figure S1.

express mRNAs with shorter 3′ untranslated regions (UTRs) (Ji et al., 2009; Mayr and Bartel, 2009; Sandberg et al., 2008). Lacking regulatory elements such as microRNA binding sites, these transcripts presumably allow increased protein expression. The mechanism underlying these changes in poly(A) site use is still unknown. It has been suggested that 3′ end processing factors such as CstF-64 could be involved but that additional factors are certainly needed (Sandberg et al., 2008).

Because of the importance of polyadenylation for mRNA stability and function, we sought to determine which of the 3′ end processing factors is most predictive for the 3′ end processing site that is ultimately used in cleavage and may thus be at the root of global changes in polyadenylation that are observed in dividing and malignant cells. Toward this end, we mapped both the binding sites of core cleavage and polyadenylation factors as well as the 3′ end cleavage sites in the same system, the human embryonic kidney 293 (HEK293) cell line. Based on our analysis of the binding data and on previous reports of core 3′ end processing components that affect poly(A) site choice (Kim et al., 2010), we further investigated the effect that the siRNA-induced knockdowns of CF $\mathrm{I_m}$68 and CstF-64 have on the selection of cleavage sites.

Several high-throughput methods for mapping binding sites of RNA-binding proteins based on crosslinking and immunoprecipitation (CLIP) were recently introduced (Hafner et al., 2010; König et al., 2010; Licatalosi et al., 2008; Ule et al., 2005). Here we used a photoreactive nucleotide-based CLIP method (PAR-CLIP) in which we merged steps that were shown with other methods to improve accuracy. Similarly, several methods for determining sites of pre-mRNA cleavage and polyadenylation are currently in use (Jan et al., 2011; Mangone et al., 2010; Shepard et al., 2011). They start with either oligo dT priming (Jan et al., 2011; Mangone et al., 2010; Shepard et al., 2011) or splint ligation to the poly(A) tail followed by RNase H digestion (Jan et al., 2011). For our study, we developed a new method, which we called A-seq. Its main feature is that it enables sequencing of mRNA 3′ ends in the sense direction, avoiding sequencing through stretches of As or Ts at the 5′ end, thus only requiring standard base calling.

RESULTS

A-Seq Is an Efficient Method to Map 3′ End Processing Sites

A sketch of the A-seq method is given in Figure 1. It starts with selection of the poly(A)-containing RNAs on oligo-(dT)$_{25}$ Dynabeads, followed by partial fragmentation by RNase I. The 5′ ends are then phosphorylated, the 3′ ends blocked by 3′dATP, and an RNA primer is ligated to the 5′ end of the RNA fragments. Our reverse transcription (RT) primer consists of an anchor nucleotide (A, C, or G) designed to align to the first nucleotide upstream of the poly(A) tail, followed by six dTs, a stem-loop containing the 3′ adaptor sense strand (needed for priming the subsequent PCR reaction), its complement, and finally a stretch of 18 dTs, which together with the 6 dTs after the anchor nucleotide form a 24 nucleotide long contiguous oligo dT stretch that aligns to the poly(A) tail. The products of RT and PCR are therefore expected to have six As preceding the 3′ adaptor. These are removed by the annotation procedure prior to mapping the reads to the genome, but they allow us to accurately pinpoint the location of the pre-mRNA cleavage (see also Figure S1). The procedure can be completed in 2 days (see Extended Experimental Procedures for details).

For reasons that will become apparent in the following sections, we generated four A-seq libraries: one from cells grown in usual conditions, one from cells treated with a control siRNA, and two from cells treated with siRNAs directed against specific 3′ end processing factors. They were multiplexed and sequenced with an Illumina HiSeq-2000 deep sequencer generating 1.8–3.3 × 10^7 reads per library (for an overview, see Table S1). The data are deposited in the GEO database of NCBI (see Accession Numbers section) and can be further explored on our web server (http://www.clipz.unibas.ch). Analysis of these libraries as described in Experimental Procedures yielded a total of 31,906 cleavage sites (CSs) at a false discovery rate of 10%, estimated based on the presence of the polyadenylation signal. Of these, 17,669 sites were previously annotated in the polyA-DB database (Zhang et al., 2005) and an additional 1,962 correspond to 3′ ends of cDNAs in the GenBank database. We further found that, in order, 3,672 identified CSs map to 3′ UTRs but do not coincide with mRNA-documented 3′ ends, 321 map to coding exons, 2,388 to introns and 1,646 to the antisense strand of known genes. There were 4,248 not located within the loci of annotated genes. Overall, 23,996 of the 31,906 sites (75%) were located within or up to 1,000 nucleotides downstream of the transcription units of Entrez (http://www.ncbi.nlm.nih.gov/gene) genes.

A total of 7,588 of the cleavage sites from the untreated sample and 7,504 from the control siRNA-treated sample accumulated at least 90% of the 3′ end counts associated with the corresponding genes, and we therefore called them “dominant” sites. The vast majority of these sites, 7,314 (96% and 97% of the dominant sites identified in the two samples, respectively), were identified as dominant sites in both of these samples, indicating that the method has very good reproducibility. We used

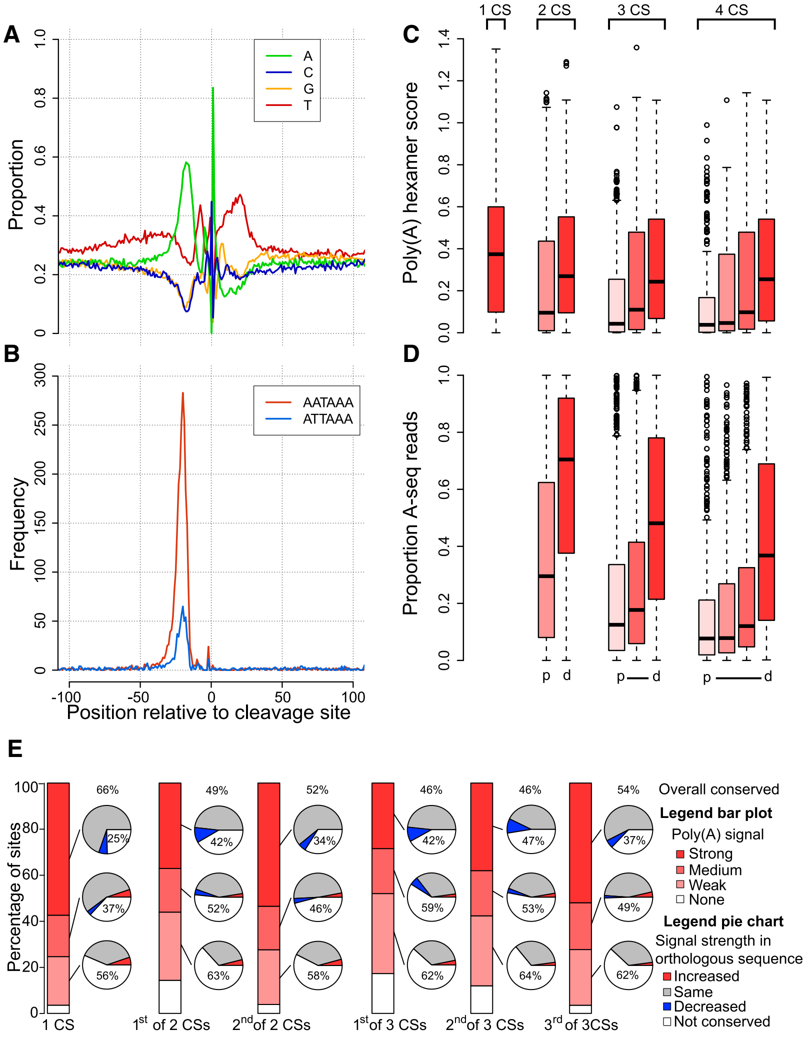

Figure 2. Relationship between Polyadenylation Motifs and Cleavage Sites

(A and B) Nucleotide composition (A) and frequency of the two most common poly(A) signals (B), as a function of distance relative to the dominant cleavage sites that are anchored at 0.

(C and D) Distributions of poly(A) hexamer scores (see Experimental Procedures) (C) and fraction of reads derived from the first, second, third, or fourth cleavage site (D) for genes with 1, 2, 3, and 4 identified cleavage sites. Cleavage sites are sorted from most proximal (left) to most distal (right). Distributions are summarized as box plots, with boxes indicating the interquartile range, the black horizontal the median, and the whiskers delimiting the interval of 1.5 times the interquartile range, centered at the median. Points outside of this interval are shown as circles.

(E) The type of poly(A) signal (strong: AAUAAA, medium: AUUAAA and AGUAAA, weak: all other motifs described in Beaudoing et al., 2000) found at alternative cleavage in genes with 1, 2, or 3 tandem cleavage sites (CSs) as well as their conservation in mouse, indicated by the type of poly(A) signal identified in the orthologous mouse regions (details in Experimental Procedures).

See also Figure S2.

the 3,000 most abundantly expressed of these CSs for our analyses. Nucleotide composition around the dominant cleavage sites closely resembles that observed in the nematode *Caenorhabditis elegans* (Jan et al., 2011), and the cleavage occurs preferentially at CA dinucleotides (Figure 2A). The canonical polyadenylation signal AAUAAA is strongly enriched in a window of ~40 nucleotides upstream of the dominant cleavage sites, peaking at −20 nucleotides. The most frequent variant polyadenylation signal, AUUAAA (Wickens, 1990), has a similar positional preference with respect to the dominant CS (Figure 2B).

Among genes that use multiple cleavage sites in HEK293 cells we found a preference for distal sites (Figure 2D). Northern blots performed for a few selected genes confirm the relative usage of alternative poly(A) sites that we inferred from A-seq and show that mainly distal cleavage sites are used, generating transcripts with long 3′ UTRs (Figure S2). To understand the mechanism behind this preference, we examined the "strength" of the alternative 3′ end processing sites, which we defined based on the relative frequency and positional preference of polyadenylation motifs, as described in Experimental Procedures and we called poly(A) hexamer score. We found that the poly(A) hexamer score increased progressively from the proximal to the distal site, in parallel with the frequency of A-seq reads at these alternative sites (Figures 2C and 2D). Furthermore, distal sites exhibit a higher degree of evolutionary conservation (Figure 2E).

Mapping of Binding Sites of RNA-Binding Proteins by a Modified PAR-CLIP Method

Based on previous work (Kishore et al., 2011) we mapped the binding sites of 3′ end processing factors with a method that combines steps from two recently published CLIP methods, PAR-CLIP (Hafner et al., 2010) and HITS-CLIP (Chi et al., 2009; Ule et al., 2005). We crosslinked 4-thio-uridine-containing RNA to proteins with 365 nm ultraviolet (UV) light (as in PAR-CLIP) to readily identify crosslinked positions based on the abundant crosslink-diagnostic mutations. We then employed the steps of nuclease digestion, primer ligation and blotting of immunoprecipitated complexes to nitrocellulose from the HITS-CLIP method, to minimize cloning of background RNA (König et al., 2010; Licatalosi et al., 2008; Ule et al., 2005). Immunoprecipitation (IP) was performed with protein-specific antibodies or with anti-FLAG antibodies when stably transformed cell lines were generated to produce FLAG-tagged proteins (see Experimental Procedures for details). Table S2 lists the 3′ end processing factors that we investigated. We obtained 1.2–2.4 × 10^7 reads per sample (Table S3), which we annotated on the CLIPZ web server (Khorshid et al., 2011) with a procedure described in further detail in the Extended Experimental Procedures.

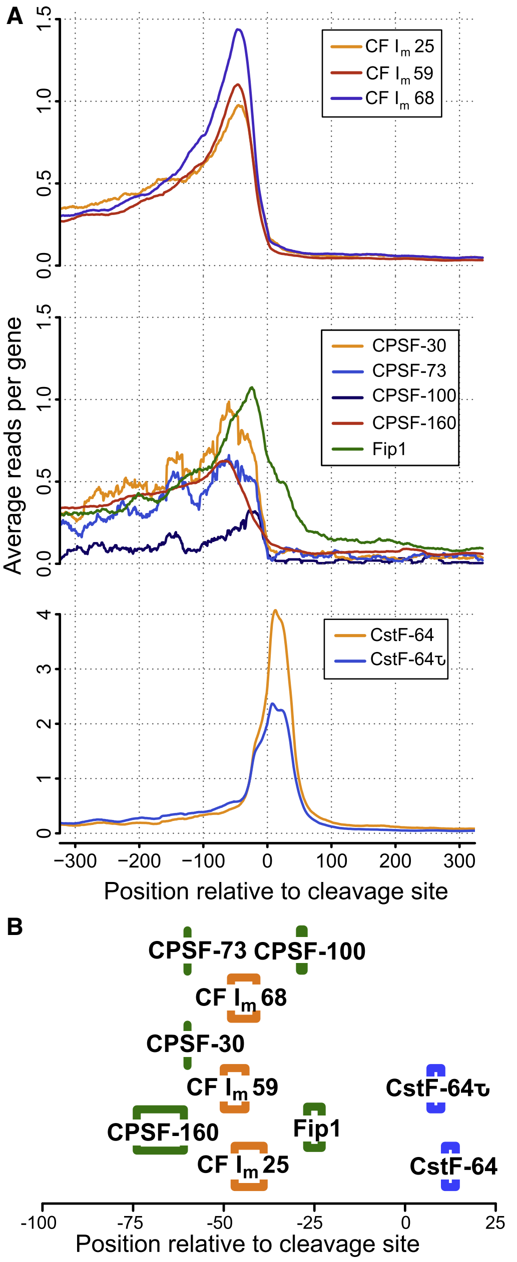

Figure 3. Positional Profiles of Binding of 3′ End Processing Factors Relative to Dominant Cleavage Sites

(A) Average density of reads from PAR-CLIP samples of CF I_m (top), CPSF (middle), and CstF (bottom) proteins in the vicinity of the 3,000 most abundant dominant cleavage sites.

(B) The span of the region in which the density of reads is within 1% of the density at the peak for each factor. Positions are indicated in nucleotides relative to the cleavage site, which is located at 0.

See also Figure S3.

3′ End Processing Factors Differ Widely in Their Specificity of Positioning Relative to the Cleavage Site

Based on the genes with a dominant cleavage site, we determined the sequence and positional preference of 3′ end processing factors relative to the CS. As the example in Figure S1A shows, cleavage sites are very well defined and colocalize with binding sites of multiple 3′ end processing factors. However, as indicated by the profiles in Figure 3A, only the cleavage stimulation factor subunits CstF-64 and CstF-64τ exhibit very high positional specificity (15–30-fold enrichment at the peak relative to adjacent regions of the gene), binding within 25 nucleotides downstream of the CS (Figure 3B). This may indicate that CstF-64 contacts the RNA at a late stage of 3′ end processing. The other proteins appear to bind more diffusely along transcription units with a preference for exons. The CF I_m subunits of 25, 59, and 68 kDa exhibit distinct peaks at 40–50 nt upstream of CSs (Figure 3A), with CF I_m25 showing the lowest positional specificity. The reason behind this characteristic positioning appears to be 2-fold. The CF I_m complex is known to bind UGUA motifs (Brown and Gilmartin, 2003) and indeed, half of the dominant sites have UGUA motifs whose frequency peaks at 40–50 nucleotides upstream of the cleavage site (Figure S3B, right panels). For the other half of the sites, with no UGUA motif within 100 nucleotides upstream of the cleavage site, CF I_m still has a peak, albeit smaller, at ∼40 nucleotides upstream of the cleavage site (Figure S3B, left panels). This may indicate that factors other than sequence-specific binding, for example interactions with other components of the 3′ end processing machinery, contribute to CF I_m positioning at the cleavage site.

CPSF-160, -30, and -73 were crosslinked with peaks upstream of the CF I_m components, whereas Fip1 and CPSF-100 crosslinked downstream of CF I_m (Figures 3A, middle panel, 3B, and S3A). It is especially surprising that CPSF-160, which has been previously shown to recognize the hexameric sequence of the poly(A) signal (Keller et al., 1991; Murthy and Manley, 1995) crosslinked along the entire transcription unit with a shallow peak some 70 nt upstream of the CSs (Figures 3A, middle panel, and 3B). This binding pattern was similar between replicates and even between samples prepared with 254 nm UV crosslinking or with 365 nm UV crosslinking after 4-thiouridine treatment (Figure S3A, right panel). A possible explanation is suggested by a previous study that showed that direct interaction of CPSF-160 with a sequence element 76 nt upstream of the AAUAA hexamer enhanced the efficiency of processing (Gilmartin et al., 1995). Such—frequently U-rich—upstream sequence elements, that seem to be essential for stabilizing the cleavage complex can be located at variable distances from the core poly(A) site and may contribute to the diffuse crosslinking pattern that we observed for CPSF-160. Another hypothesis is that CPSF binds diffusely across a loop created by the binding of CF I_m to two UGUA motifs (Yang et al., 2011). Finally, the fact that CPSF is recruited (together with other 3′ processing factors such as CF I_m and CstF) at the initial stages of transcription and interacts with the C-terminal domain of RNA polymerase II (Nag et al., 2007; Venkataraman et al., 2005) may also contribute to its diffuse crosslinking across the transcription unit.

Sequence Specificities of 3′ End Processing Factors

We used the MEME software (Bailey et al., 2009) to infer sequence motifs that are over-represented among the 500 most abundantly CLIPed sites from each of the samples and

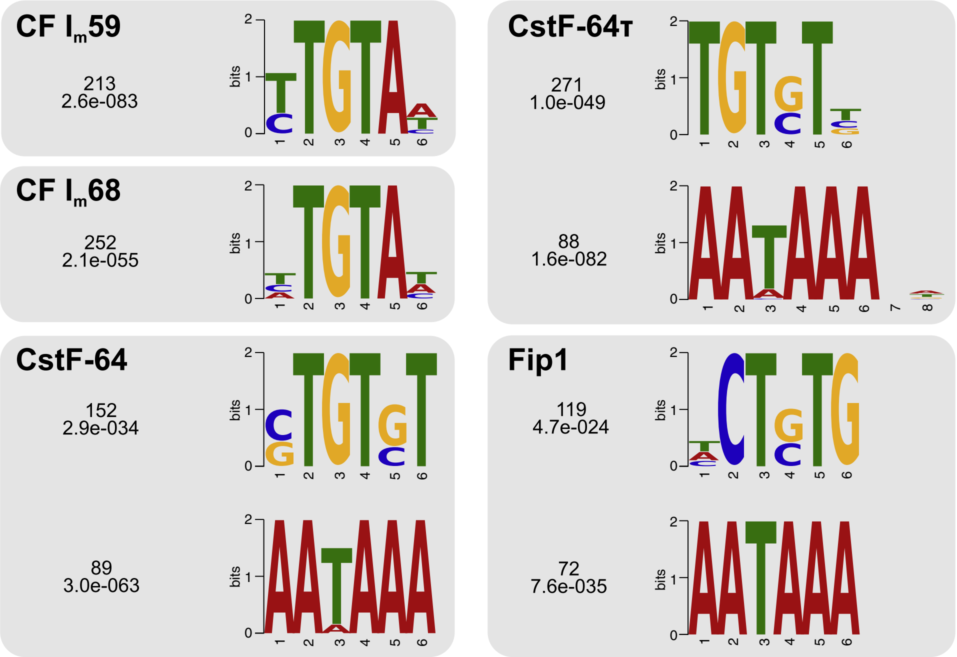

Figure 4. Sequence Motifs That Are Most Enriched in the Binding Sites of 3′ End Processing Factors

The MEME-identified motifs that were represented in at least 50 of the most abundantly isolated 500 sites of various 3′ end processing factors are shown. For each motif we indicated the number of sites among the top 500 that contained it and the E-value.
See also Figure S4.

thus may bind individual 3′ end processing factors (Figure 4, see also Experimental Procedures). CF I_m59 and CF I_m68 binding sites exhibit a strong and reproducible enrichment of the UGUA motif, consistent with the initial reporting of CF I_m binding at UGUA elements (Brown and Gilmartin, 2003). Recent structural studies (Li et al., 2011; Yang et al., 2011) showed that two CF I_m25 molecules of the tetrameric complex of the CF I_m68 RRM and CF I_m25 interact with two UGUA RNA molecules and RNA-binding experiments strongly suggest that residues of the CF I_m68 RRM domain also contact the RNA (Yang et al., 2011). Our findings that CF I_m59 and CF I_m68 crosslink more specifically around CSs (but not directly on the UGUA motifs, see Figure S4) compared to CF I_m25 may indicate that CF I_m25 interacts with the RNA more promiscuously, whereas CF I_m59 and CF I_m68 are within crosslinking distance only when the UGUA elements are specifically recognized by CF I_m25. Alternatively, the CF I_m complex may not bind efficiently to UGUA elements in which the U's are substituted by 4-thio-Us, and thus CF I_m25 is not crosslinked at these locations. A clear example of CF I_m binding at UGUA motifs is the first CS in the 3′ UTR of poly(A) polymerase-γ (PAPOLG) that has a noncanonical polyA signal and contains seven UGUA motifs upstream of the poly(A) site. All of these motifs crosslinked to CF I_m (Figure S1B).

Consistent with previous reports (Beyer et al., 1997; Pérez Cañadillas and Varani, 2003; Takagaki and Manley, 1997), we found that 153 and 272 of the 500 most abundantly CLIPed sites for CstF-64 and CstF-64τ, respectively are enriched in a G/U-rich motif (Figure 4). A second motif that is significantly enriched in the CstF-64 and CstF-64τ sites (89 and 88 occurrences, respectively, in the top 500 sites) resembles the canonical polyadenylation signal, likely reflecting the fact that these factors bind very close to the CS (Figure 3A). Of the CPSF proteins, only Fip1 yielded enriched sequence motifs that occurred in at least 50 of the top 500 sites. Both the canonical poly(A) signal (72 occurrences) as well as a G/U motif (119 occurrences) probably binding CstF-64 were identified by MEME in the top 500 sites of Fip1. The second Fip1 replicate, with similar but less biased positioning of Fip1 upstream of the CS (Figure S3A, left panel), did not reveal a significant enrichment of these motifs. As expected from the positional crosslinking profile, we were not able to reliably identify the polyadenylation signal among the top sites of CPSF-160, the protein that was previously shown to bind this element. The motif appeared in only 66 of the 500 most abundantly CLIPed sites from one of the CPSF-160 samples, but was not significantly enriched in the top sites from the second, smaller CPSF-160 sample. Crosslinking with 254 nm UV light in the absence of 4-thiouridine (Figure S3A, right panel) also did not identify this motif that together with the fact that CPSF components are crosslinked in a broad region upstream of the CS, with individual subunits exhibiting reproducible positional preferences, suggests that the appropriate recognition of the polyadenylation signal requires a specific conformation of the CPSF complex that is either very transient or difficult to capture by crosslinking.

The frequencies with which the above-mentioned enriched sequence motifs occur at different positions relative to the crosslinked sites of individual factors are shown in Figure S4. CF I_m25 crosslinks immediately downstream, as well as on the UGUA

Table 1. Proportion of Genes with Tandem Cleavage Sites in which the Indicated Factors Have the Maximum at the Dominant as Opposed to Alternative Cleavage Sites

Reads at the Dominant CS	>50%	>60%	>70%	>80%	>90%
Genes with 2 CS	1,354	1,160	964	761	473
Genes with 3 CS	532	418	311	213	122
Genes with 4 CS	166	128	88	62	31
Hexamer score	56.97	59.14	62.04	63.80	67.09
CF I$_m$68	54.53	57.33	60.57	64.29	67.89
CstF-64τ	57.07	59.55	61.67	63.03	66.29
CstF-64	52.92	55.80	57.86	59.65	63.42
CF I$_m$59	53.95	56.62	59.10	60.42	61.66
CF I$_m$25	44.88	46.54	49.63	52.12	52.72
CPSF-160	39.57	41.21	42.95	44.79	45.53
Fip1	31.97	33.70	34.88	36.58	39.62
CPSF-30	9.75	10.43	10.87	11.20	10.86
CPSF-73	4.53	4.63	4.99	5.60	5.27
CPSF-100	2.53	2.52	2.20	2.41	2.40

CS, cleavage sites.

motifs. The most frequent crosslinking position for CF I$_m$68 is immediately downstream of UGUA nucleotides, and both CF I$_m$68 and CF I$_m$59 crosslink in a broad region upstream of UGUA motifs. These results indicate that the UGUA motif is indeed specific for the CF I$_m$ complex. CstF-64 and CstF-64τ crosslink most frequently on the UGUSU motif (with S being C or G), but also within 20 nucleotides downstream of it, again indicating specific binding to this motif. Fip1 appears to be able to crosslink immediately downstream or even at the U of the AAUAAA poly(A) signal, in contrast to CstF-64, which crosslinks more than 10 nucleotides downstream of this motif. Thus, the enrichment of the polyadenylation signal in binding sites of Fip1 may be due to direct binding of Fip1 to this motif, whereas the enrichment in binding sites of CstF-64 is very likely due to the fact that the AAUAAA motif occurs in very close proximity to the CstF-64 motif.

Binding of CF I$_m$68 and CstF-64/CstF-64τ Are Most Predictive for the Location of the Cleavage Site

Selection of the 3′ end processing site is a complex process, involving many core 3′ end processing factors as well as modulatory interactions with other RNA-binding proteins. To uncover the factors that are most decisive in the selection of cleavage sites, we related the density of reads obtained through CLIP with 3′ end processing factors to the relative use of cleavage sites determined through A-seq. We asked two questions. The first was what proportion of genes with a clearly dominant cleavage site (accumulating at least 90% of the A-seq reads associated with the gene) has a peak in binding of a specific factor or subcomplex of factors in the immediate vicinity of this dominant site as opposed to somewhere else in the gene body. Table S4 summarizes the results. Strikingly, the binding of CstF-64/CstF-64τ alone "explains" over 50% of the dominant CSs, meaning that the highest density of reads from these factors occurs indeed in the immediate vicinity of the CS. Similarly, the CF I$_m$ 59 and 68 "explain" 39%–42% of the dominant CSs, and of the sites that do not have the peak of CstF-64/CstF-64τ binding in the immediate vicinity of the dominant CS, 15% have the peak of CF I$_m$59/68 binding in this region. Joint binding of CstF-64 and CF I$_m$59/68 explains over 60% of the dominant CSs, whereas 40% of the sites cannot be "explained" by the highest peak in factor/subcomplex binding around the dominant CS even when we use combinations of three factors. For comparison, 73% of the dominant CSs have the canonical poly(A) signal AAUAAA within 40 nucleotides upstream of the CS. Thus, we found that the core components and their associated signals explain a large fraction of the CS data at least for the dominant CSs.

The second question that we asked was whether the relative binding of 3′ end processing factors in the vicinity of alternative CSs explains the relative usage of alternative cleavage sites of genes. That is, is the highest peak in binding of a given factor occurring at the cleavage site that is predominantly used? For this purpose we extracted genes that had two, three or four tandem CSs in the same 3′ exon, one of these tandem sites being predominantly used. We performed the test with different cutoffs on the proportion of reads that the dominant site is required to accumulate (at least 50%, 60%, 70%, 80%, and 90% of the reads) relative to all cleavage sites considered for that gene. As shown in Table 1, we found that binding of CF I$_m$68, followed by CstF-64 and CstF-64τ is most indicative of the cleavage site that is selected among multiple alternatives for a given gene, suggesting that these factors may be important for poly(A) site selection.

Knockdown of CF I$_m$68 Induces the Use of Proximal Poly(A) Sites, Mimicking the Behavior Observed in Proliferating Cells

A tendency toward the use of proximal polyadenylation sites has been reported in dividing compared to resting cells, and in cancer cells relative to their normal counterparts (Mayr and Bartel, 2009; Sandberg et al., 2008), and it has been speculated that 3′ end processing factors may be involved. Recent studies also reported a proximal shift in poly(A) site usage in the TIMP2 and DHFR transcripts upon knockdown of CF I$_m$25 and CF I$_m$68, but not of CF I$_m$59 or CstF-64 in HeLa cells (Kim et al., 2010; Kubo et al., 2006). Based on the data that we presented above, indicating that the binding of CF I$_m$ and CstF-64 proteins are most predictive for the predominantly used cleavage site, we performed siRNA-mediated knockdowns of CF I$_m$68 and CstF-64 and found that the former results in a marked reduction in the amount of poly(A)$^+$ RNA per cell (Figure S5A). Furthermore, analysis of A-seq data obtained with siRNA-treated cells revealed that the knockdown of CF I$_m$68 but not that of CstF-64 leads to a systematic, transcriptome-wide shift toward proximal poly(A) sites. This is apparent in Figure 5A that shows that the majority of genes have a higher proximal/distal ratio of poly(A) site usage in the CF I$_m$68 knockdown compared to the control siRNA-treated sample. Consistent with the hypothesis that shortening of 3′ UTRs activates oncogenes (Mayr and Bartel, 2009), a gene that undergoes a dramatic shift in poly(A) site usage upon CF I$_m$68 knockdown is the ring finger protein 11

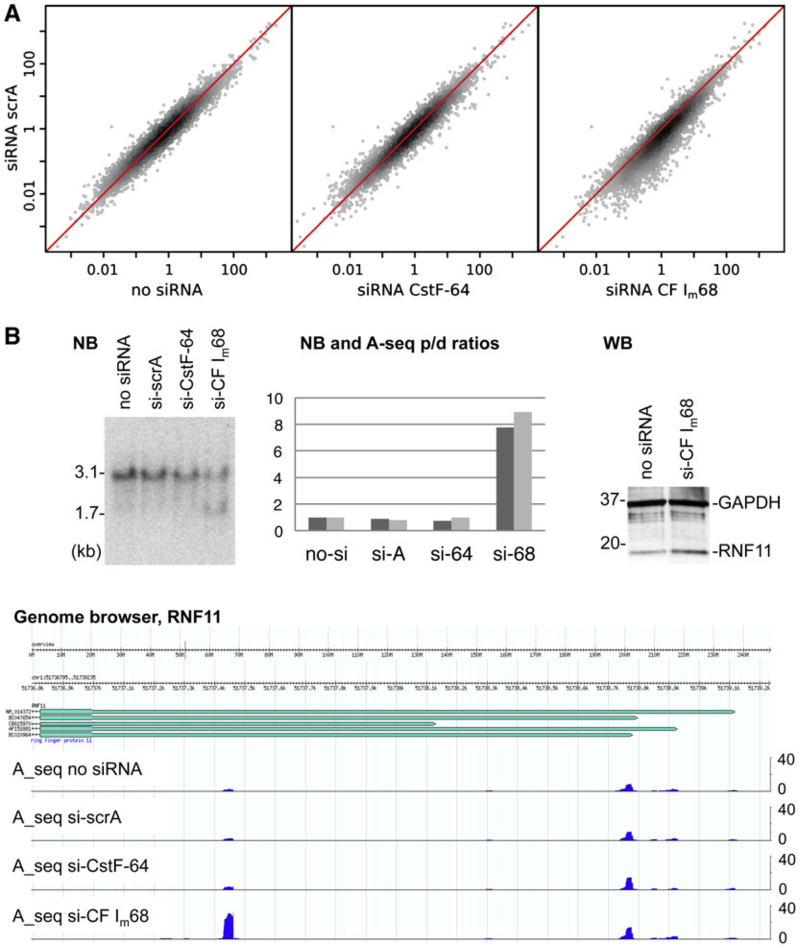

Figure 5. Scatter Plots of Proximal/Distal Poly(A) Site Usage Ratio in Pairs of A-seq Samples

(A) A-seq samples prepared from cells treated as indicated were used to infer poly(A) site usage. Each dot represents one gene that had more than one cleavage site in a terminal exon. The proximal/distal ratios were calculated as Σ(number of A-seq reads at all 3′ end processing regions except the distal one)/(number of A-seq reads at the most distal 3′ end processing region).

(B) Effects of CF I$_m$68 siRNA treatment on the poly(A) site choice (northern blot panel, NB and genome browser panel showing the A-seq results) and protein levels (WB panel) of RNF11. Comparison of proximal/distal site usage ratios between northern blots (dark gray columns, quantification done with the ImageJ software; http://rsb.info.nih.gov/ij/) and A-seq (light gray columns) are indicated in panel "NB and A-seq p/d ratios," where no-si indicates no siRNA treatment, si-A is siRNA scrambled control A, si-64 is siCstF-64 and si-68 is siCF I$_m$68 treatment.

See also Figure S5.

(RNF11), known to enhance signaling through the TGF-β pathway and to play a role in breast cancer progression (Yu et al., 2008). As shown in Figure 5B, shortening of RNF11 3′ UTR upon CF I$_m$68 knockdown is associated with a 2.4-fold increase in protein expression relative to GAPDH. Additional examples are shown in Figure S2, with northern blots confirming the A-seq results.

DISCUSSION

Like other mRNA biogenesis steps in mammalian cells, polyadenylation involves numerous core factors as well as modulator proteins that determine which of the typically multiple poly(A) sites that are encoded in a given gene is used to generate a particular transcript. Given the complexity of poly(A) site selection, it came as a surprise that specific conditions such as active cell division (Ji et al., 2009; Mayr and Bartel, 2009; Sandberg et al., 2008) are associated with a predominant use of proximal poly(A) sites. The factors that induce this change are still unknown. Knockdown of U1 snRNP has been previously shown to cause premature cleavage and polyadenylation, but typically at cryptic intronic sites (Kaida et al., 2010). Here we find that knockdown of CF I$_m$68 leads to preferential use of the more proximal among alternative polyadenylation sites that are all used to various extents under normal conditions. This change is not accompanied by a parallel induction of intronic site usage (Figure S5B). Whether CF I$_m$68 faithfully reproduces the changes in poly(A) site usage that are observed in dividing compared to resting cells remains to be determined. However, a previous study identified CF I$_m$68 as one of a handful of genes whose expression is very tightly regulated in cancer cells and whose downregulation increases the invasiveness of tumor cells that otherwise have poor ability to invade (Yu et al., 2008). Our study suggests a mechanism behind changes in polyadenylation that take place in tumor cells and affect their malignant properties. Consistently, we found that CF I$_m$68 knockdown induces a marked shift toward a proximal poly(A) site and increased expression of *RNF11*, a gene that modulates signaling through the TGF-β pathway that is frequently perturbed in cancers (Yu et al., 2008).

Genome-wide and transcriptome-wide maps of regulatory elements have substantially contributed to our understanding of the regulation of biological processes such as for example transcription during embryonic development (Göke et al., 2011). The intention of our study was to generate similar transcriptome-wide maps of binding sites of 3′ end processing factors to help unravel the rules of 3′ end processing. Our CLIP results indicate that binding of CF I$_m$ and CstF subcomplexes occurs at sites that are most frequently associated with cleavage events. Although differences in the inferred specificity of positioning of 3′ end processing factors relative to cleavage sites may be due to some factors being more easily crosslinked than others, the fact that we infer similar positional preference for multiple factors within one complex speaks against this hypothesis. Furthermore, the few cases that we also investigated with both 365 nm UV crosslinking after 4-thiouridine

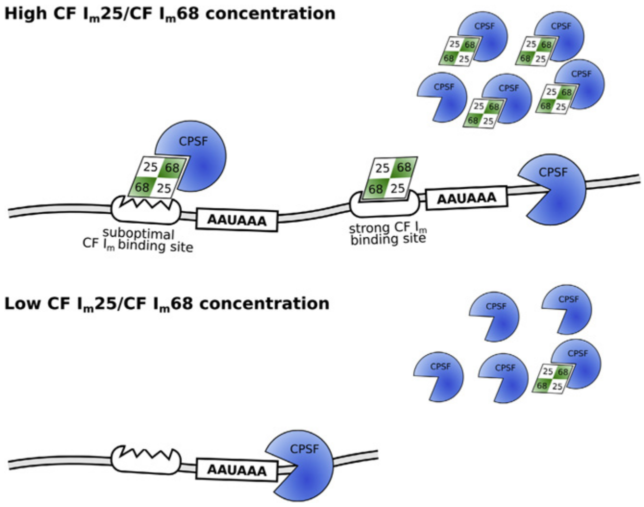
High CF I_m25/CF I_m68 concentration
CPSF
25 68
68 25
suboptimal CF I_m binding site
AAUAAA
strong CF I_m binding site
Low CF I_m25/CF I_m68 concentration

Figure 6. A Model of the Effect of CF I_m68 Concentration on the Choice of the Cleavage Site

Lack of deposition of CF I_m at proximal sites hinders the cleavage and stimulates transcription toward distal cleavage sites. Deposition of CF I_m at optimal, distal sites releases the block on the CPSF cleavage activity, allowing formation of the 3′ end. Absence of CF I_m from some 3′ end complexes when its concentration is low results in no inhibition being sensed at the proximal site, where CPSF can cleave to produce a mature 3′ end. For simplicity, additional factors and the RNA polymerase are not depicted.

treatment and 254 nm showed that the inferred specificities were not simply due to the crosslinking method. Nonetheless, it is surprising that the CPSF-160 protein, which has been previously shown to bind the AAUAAA hexamer, the most specific signal for polyadenylation that is present in the large majority of the 3′ end processing sites that we obtained, was not specifically crosslinked to such motifs. Although the reason is presently unclear, the fact that other CPSF subunits such as the CPSF-73 and CPSF-30 were also crosslinked relatively distantly from the cleavage site suggests that the CPSF complex in its entirety was crosslinked to RNA either as a pre- or postcleavage complex (Zarkower and Wickens, 1987). If the actual cleavage reaction is very fast, the complex may be in a standby position at a distance from its specific site most of the time. Nonetheless, it is also possible that the RNA binding pocket of CPSF-160 does not contain amino acids that can be crosslinked to the very specific AAUAAA motif.

In spite of its binding preference being very predictive for the cleavage site, the knockdown of CstF-64 did not induce dramatic changes in poly(A) site choice. This could be due to CstF-64 contacting the RNA at a relatively late stage of 3′ end processing, enhancing the efficiency of cleavage rather than contributing to cleavage site selection. Alternatively, CstF-64τ may be sufficient for the function of the CstF complex when CstF-64 is downregulated. Indeed, CstF-64τ has been previously reported to be essential for spermatogenesis when CstF-64 expression is silenced (Dass et al., 2007) and can even form heterodimers with CstF-64 (Shi et al., 2009). Here we found that the targets of CstF-64τ expressed in HEK293 cells from a FLAG-tagged construct are similar to those of CstF-64 (Table S4), suggesting that the two proteins have redundant functions, with CstF-64τ taking over when CstF-64 is not expressed.

Finally, although the fact that CF I_m68 was more frequently crosslinked at the most frequently used cleavage sites may suggest that CF I_m68 directly and specifically selects the poly(A) site, knockdown of CF I_m68 resulted in a systematic shift toward proximal cleavage sites, which are skipped in cells that expressed normal levels of CF I_m68. A model that could explain these observations is that stable binding of CF I_m68 at high-affinity (predominantly distal) sites is required for the proper function of CPSF and efficient 3′ end processing. Low-affinity or sub-optimally positioned CF I_m68 binding sites located around proximal cleavage sites may reduce the binding of the CPSF complex under normal conditions and stimulate the RNA polymerase to continue transcription toward distal cleavage sites. Alternatively, CF I_m68 may establish interactions with UGUA motifs located upstream of two alternative polyadenylation sites, looping out the proximal cleavage site and promoting cleavage at the distal site (Yang et al., 2011). In either case, the absence of CF I_m68 in the knockdown may allow proximal sites to interact with CPSF, leading to cleavage at these sites (Figure 6).

To conclude, our results reveal a global tendency of proximal poly(A) site use when the level of the CF I_m68 3′ end processing factor subunit is reduced. Moreover, they provide a very extensive map of both the binding sites of 3′ end processing factors as well as the 3′ end processing sites in a mammalian cell line. Because binding sites of many other regulators such as Argonaute 2 and HuR (Kishore et al., 2011) have also been mapped in the HEK293 cell line our data should help reveal additional crosstalks in the regulation of 3′ end processing.

EXPERIMENTAL PROCEDURES

Antibodies

Mouse monoclonal antibody M2 (Sigma, F1804) was used to immunoprecipitate FLAG-tagged proteins for the PAR-CLIP procedure. Commercial antibodies from Santa Cruz Biotechnology included: sc-81109 (CF I_m25), sc-16473 (CstF-64), sc-28872 (CPSF-160), sc-81232 (CPSF-30), sc-26661 (CPSF-73), and sc-32233 (GAPDH). Antibodies against CPSF-160 (A301-580A), CPSF-100 (A301-581A), CPSF-73 (A301-581A), Fip1 (A301-462A), CstF-64tau (A301-486A), and CF I_m59 (A301-359A-1) were from Bethyl. Rnf11 (65-154) antibody was from Abnova. We also used antibody #1005 (Kaufmann et al., 2004) for IP with Fip1 for sample MCLIP_FIP1_AR_173-2 (Kaufmann et al., 2004) and anti-CPSF-100 antiserum mAb-J1/27 (Jenny et al., 1994) for sample MCLIP_CPSF100_AT_187-1.

RNAi

For RNAi, HEK293 cells were seeded at a density of 20% in 6-well plates and all subsequent steps were done according to the “forward method” from the RNAiMAX protocol (Invitrogen). The next day, double stranded siRNAs (starting from 30 pmol, from Dharmacon or Santa Cruz, see below) were incubated with Lipofectamine RNAiMAX (Invitrogen) and added to the wells. After 3 days, cells were harvested and used for confirmation of the siRNA treatment by western blot and for A-seq.

The siRNAs were, for control siRNA scrambled-A, sc-37007 (scrambled-A); CF I_m68, 5′-NNGACCGAGA UUACAUGGAUA-3′ dsRNA oligo from Dharmacon; CstF-64, 5′-NNCCUGAAUG GGCGCGAAUUC-3′ dsRNA oligo from Dharmacon. The levels of CstF-64 and CF I_m68 were reduced to 5.2%–7.1% and 3.2%–4.4%, respectively in different experiments.

Inference of 3′ End Cleavage Sites from A-seq Data

A-seq reads were preprocessed to remove the six diagnostic adenosines derived from the poly(A) tail as well as the 3′ adaptor sequence and then were mapped to the human genome (hg19) and annotated using the CLIPZ server (Khorshid et al., 2011). We selected reads that mapped uniquely to genomic regions and whose annotation was "mRNA," "repeat," "miscRNA," or "unknown" and based on the precise mapping of their 3′ ends we computed putative cleavage sites with their associated abundance. To minimize the frequency of internal priming sites in our data set, we discarded those that in the eight-nucleotide-region immediately downstream of the putative cleavage site had at least seven A nucleotides. These amounted to 11.2% inferred internal priming sites across our A-seq libraries. We then scaled the library size to 1,000,000 for all samples to obtain a normalized expression value (tags per million [TPM]) for each processing site. The 3′ end cleavage appears to have some degree of imprecision, major cleavage sites being usually flanked by sites with lower abundance. We therefore inferred "3′ end processing regions" by applying single-linkage clustering with a distance threshold of 10 nucleotides to the pooled set of sites from all four A-seq libraries. We retained only 3′ end processing regions with at least 1.9 TPM in at least one A-seq library and calculated a false discovery rate in this set of 10%, based on the occurrence of polyadenylation signals upstream of the resulting sites (Shepard et al., 2011).

The representative cleavage site for a 3′ end processing region was chosen by ranking individual sites by their expression value in each A-seq library and then determining the overall top ranked site (majority vote over all A-seq libraries). In cases when multiple sites had the same rank, the most 5′ site was chosen. These sites were used in subsequent analyses.

Associating Cleavage Sites with Genes

Gene and transcript data were obtained from NCBI (accessed on: 13.07.2011). First, we assigned transcript exons to cleavage sites if the 3′ end processing regions in which the cleavage sites were located overlapped with the exon. Next, we used the Entrez Gene data to associate transcript IDs to genes, allowing us to associate 23,996 cleavage sites with genes.

Extraction of Cleavage Sites

Dominant cleavage sites were selected as those that accumulated at least 90% of all reads associated with cleavage sites of the respective gene in both the HEK293 wt and the siRNA control samples. This procedure resulted in the extraction of 7,314 dominant cleavage sites. Sites were then ranked by expression in the HEK293 wt and the top 3,000 sites were used for subsequent analyses. To extract tandem cleavage sites we determined the terminal exon from all transcripts associated with a gene that had the highest number of CSs.

Calculation of the Poly(A) Hexamer Score

We estimated the frequencies $w(s)$ of occurrence of the canonical poly(A) hexamer AAUAAA and its 11 single nucleotide variants (Beaudoing et al., 2000) in the 40 nt upstream of the cleavage site of genes with a unique cleavage site. We also estimated the frequency $\pi(i)$ with which the canonical hexamer occurred at each position i relative to unique cleavage sites, as opposed to any other position within the 40-nt long window upstream of these sites. The poly(A) hexamer score for a given cleavage site was then computed as $\Sigma\, w(s[i\ldots i+5])\pi(i)$, with i between -39 and -5. $s[i\ldots i+5]$ was the motif found at position i upstream of the CS. For the conservation analysis, we identified poly(A) signals hexamers within 40 nt upstream of cleavage sites. We then extracted pairwise human/mouse alignments of 10 nts-long windows containing the hexamer sequence via the UCSC/Galaxy interface at http://main.g2.bx.psu.edu. A polyadenylation hexamer was considered conserved if one of the 12 polyadenylation hexamers was detected in the mouse sequence.

Extraction of Binding Sites of 3′ Processing Factor Subunits

The reads obtained from PAR-CLIP experiments were mapped to the human genome and annotated with the CLIPZ server as described in Extended Experimental Procedures. Reads that appeared to be due to spurious mappings of truncated 3′ adaptor sequences (showing a perfect 10-mer match to the 3′ adaptor sequence) were discarded. For subsequent analysis we then extracted uniquely mapped reads with annotations mRNA, unknown or repeat. For the motif analysis, we constructed 40-nucleotide long, nonoverlapping binding sites ordered by their read coverage by reads, as described in Kishore et al. (2011).

Identification of Binding Motifs of 3′ End Processing Factors

Sites that were covered over at least 50% of their length by an annotated repeat element (according to the repeat annotation from the genome browser at the University of California, Santa Cruz) were excluded from subsequent analysis. We subjected the top 500 sites of each protein to the MEME motif discovery tool (Bailey et al., 2009). Requiring that a motif had an E-value of 10^{-15} or less and that it was present in at least 10% of the top 500 sites of a given protein resulted in identification of enriched motifs for CF $\mathrm{I_m}$59, CF $\mathrm{I_m}$68, Fip1, CstF-64, and CstF-64τ.

Defining the Relationship between Complexes of 3′ Processing Factors and Cleavage Sites

To determine which factors or sub-complexes are most decisive in selecting the poly(A) site we first analyzed the 3,000 most highly expressed genes (according to the number of A-seq reads in our data) that had a strongly dominant CS (that accumulated at least 90% of the A-seq reads associated with the gene). We determined the loci of these genes by taking the union of loci inferred from the genomic mapping of all RefSeq transcripts (http://www.ncbi.nlm.nih.gov/RefSeq/) that are associated with each of the genes in the Entrez database and we extended each locus by 1,000 nt at the 3′ end. We then split these loci into windows of 120 nt with the reference (0th) window being located −80 to +40 around the dominant CS, region that should contain most of the core signals that are necessary for 3′ end processing (see positional binding profiles of individual 3′ end processing factors in Figure 3). We then estimated the probability that a given factor binds in a given window as the proportion of reads associated with the locus in a given CLIP experiment that map precisely in the respective window. Similarly, we estimated the probability that a "complex" of multiple factors binds in a given window by the product of the probabilities that the individual components bind in that window. We considered that a factor or complex was "predictive" for 3′ end processing if its probability of binding was highest in the windows −1 to +1 around the experimentally determined CS. We further investigated how predictive the number of reads of a particular 3′end factor is for the choice of a cleavage site as compared to other cleavage sites located in the same exon. Only genes with cleavage sites that where spaced by at least 500 nt were considered. We determined read counts in the 120 nt window upstream for CPSF and CF $\mathrm{I_m}$ factors, and in the 120 nt window downstream of the cleavage site for CstF factors. A factor was considered to "explain" a site when the read count at that site was highest compared to all other competing cleavage sites.

ACCESSION NUMBERS

The Gene Expression Omnibus (GEO) accession number for the A-seq and CLIP data is GSE37401.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, four figures, and five tables and can be found with this article online at doi:10.1016/j.celrep.2012.05.003.

LICENSING INFORMATION

ACKNOWLEDGMENTS

We acknowledge the excellent technical assistance of Béatrice Dimitriades and the constructive comments from members of the Zavolan group, particularly Shivendra Kishore, on the manuscript. We also thank Christian Beisel, Ina Nissen, and Manuel Kohler from the DBSSE of the ETH for the support to accomplish the Illumina sequencing. The work was supported by the University of Basel, the Louis-Jeantet Foundation for Medicine and the Swiss National Science Foundation (grant 31003A_133145 to W.K.).

Received: February 24, 2012
Revised: April 10, 2012
Accepted: May 7, 2012
Published online: June 7, 2012

REFERENCES

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37* (Web Server issue), W202-8.

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

Beyer, K., Dandekar, T., and Keller, W. (1997). RNA ligands selected by cleavage stimulation factor contain distinct sequence motifs that function as downstream elements in 3′-end processing of pre-mRNA. J. Biol. Chem. *272*, 26769–26779.

Brown, K.M., and Gilmartin, G.M. (2003). A mechanism for the regulation of pre-mRNA 3′ processing by human cleavage factor I_m. Mol. Cell *12*, 1467–1476.

Cardinale, S., Cisterna, B., Bonetti, P., Aringhieri, C., Biggiogera, M., and Barabino, S.M. (2007). Subnuclear localization and dynamics of the Pre-mRNA 3′ end processing factor mammalian cleavage factor I 68-kDa subunit. Mol. Biol. Cell *18*, 1282–1292.

Carrillo Oesterreich, F.C., Bieberstein, N., and Neugebauer, K.M. (2011). Pause locally, splice globally. Trends Cell Biol. *21*, 328–335.

Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature *460*, 479–486.

Coseno, M., Martin, G., Berger, C., Gilmartin, G., Keller, W., and Doublié, S. (2008). Crystal structure of the 25 kDa subunit of human cleavage factor Im. Nucleic Acids Res. *36*, 3474–3483.

Dass, B., Tardif, S., Park, J.Y., Tian, B., Weitlauf, H.M., Hess, R.A., Carnes, K., Griswold, M.D., Small, C.L., and Macdonald, C.C. (2007). Loss of polyadenylation protein tauCstF-64 causes spermatogenic defects and male infertility. Proc. Natl. Acad. Sci. USA *104*, 20374–20379.

de Vries, H., Rüegsegger, U., Hübner, W., Friedlein, A., Langen, H., and 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.

Dettwiler, S., Aringhieri, C., Cardinale, S., Keller, W., and Barabino, S.M. (2004). Distinct sequence motifs within the 68-kDa subunit of cleavage factor I_m mediate RNA binding, protein-protein interactions, and subcellular localization. J. Biol. Chem. *279*, 35788–35797.

Di Giammartino, D.C., Nishida, K., and Manley, J.L. (2011). Mechanisms and consequences of alternative polyadenylation. Mol. Cell *43*, 853–866.

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

Göke, J., Jung, M., Behrens, S., Chavez, L., O'Keeffe, S., Timmermann, B., Lehrach, H., Adjaye, J., and Vingron, M. (2011). Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development. PLoS Comput. Biol. *7*, e1002304.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129–141.

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

Jenny, A., Hauri, H.P., and Keller, W. (1994). Characterization of cleavage and polyadenylation specificity factor and cloning of its 100-kilodalton subunit. Mol. Cell. Biol. *14*, 8183–8190.

Ji, Z., Lee, J.Y., Pan, Z., Jiang, B., and Tian, B. (2009). Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc. Natl. Acad. Sci. USA *106*, 7028–7033.

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

Kaufmann, I., Martin, G., Friedlein, A., Langen, H., and 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.

Keller, W., Bienroth, S., Lang, K.M., and Christofori, G. (1991). Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3′ processing signal AAUAAA. EMBO J. *10*, 4241–4249.

Khorshid, M., Rodak, C., and Zavolan, M. (2011). CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. Nucleic Acids Res. *39* (Database issue), D245–D252.

Kim, S., Yamamoto, J., Chen, Y., Aida, M., Wada, T., Handa, H., and Yamaguchi, Y. (2010). Evidence that cleavage factor Im is a heterotetrameric protein complex controlling alternative polyadenylation. Genes Cells *15*, 1003–1013.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat. Methods *8*, 559–564.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. *17*, 909–915.

Kubo, T., Wada, T., Yamaguchi, Y., Shimizu, A., and Handa, H. (2006). Knock-down of 25 kDa subunit of cleavage factor Im in HeLa cells alters alternative polyadenylation within 3′-UTRs. Nucleic Acids Res. *34*, 6264–6271.

Li, H., Tong, S., Li, X., Shi, H., Ying, Z., Gao, Y., Ge, H., Niu, L., and Teng, M. (2011). Structural basis of pre-mRNA recognition by the human cleavage factor Im complex. Cell Res. *21*, 1039–1051.

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

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

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

Mangone, M., Manoharan, A.P., Thierry-Mieg, D., Thierry-Mieg, J., Han, T., Mackowiak, S.D., Mis, E., Zegar, C., Gutwein, M.R., Khivansara, V., et al. (2010). The landscape of C. elegans 3′UTRs. Science *329*, 432–435.

Martin, G., and Keller, W. (2007). RNA-specific ribonucleotidyl transferases. RNA *13*, 1834–1849.

Martin, G., Ostareck-Lederer, A., Chari, A., Neuenkirchen, N., Dettwiler, S., Blank, D., Rüegsegger, U., Fischer, U., and Keller, W. (2010). Arginine methylation in subunits of mammalian pre-mRNA cleavage factor I. RNA *16*, 1646–1659.

Martin, K.C., and Ephrussi, A. (2009). mRNA localization: gene expression in the spatial dimension. Cell *136*, 719–730.

Martincic, K., Campbell, R., Edwalds-Gilbert, G., Souan, L., Lotze, M.T., and Milcarek, C. (1998). Increase in the 64-kDa subunit of the polyadenylation/cleavage stimulatory factor during the G0 to S phase transition. Proc. Natl. Acad. Sci. USA *95*, 11095–11100.

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

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

Millevoi, S., and Vagner, S. (2010). Molecular mechanisms of eukaryotic pre-mRNA 3′ end processing regulation. Nucleic Acids Res. *38*, 2757–2774.

Moore, M.J. (2005). From birth to death: the complex lives of eukaryotic mRNAs. Science *309*, 1514–1518.

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

Nag, A., Narsinh, K., and Martinson, H.G. (2007). The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. Nat. Struct. Mol. Biol. *14*, 662–669.

Pérez Cañadillas, J.M., and Varani, G. (2003). Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. EMBO J. *22*, 2821–2830.

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

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

Ruepp, M.D., Aringhieri, C., Vivarelli, S., Cardinale, S., Paro, S., Schümperli, D., and Barabino, S.M. (2009). Mammalian pre-mRNA 3′ end processing factor CF I m 68 functions in mRNA export. Mol. Biol. Cell *20*, 5211–5223.

Ryan, K. (2007). Pre-mRNA 3′ cleavage is reversibly inhibited in vitro by cleavage factor dephosphorylation. RNA Biol. *4*, 26–33.

Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A., and Burge, C.B. (2008). Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. Science *320*, 1643–1647.

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

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

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

Sonenberg, N., and Hinnebusch, A.G. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell *136*, 731–745.

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

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

Ule, J., Jensen, K., Mele, A., and Darnell, R.B. (2005). CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods *37*, 376–386.

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

Wickens, M. (1990). How the messenger got its tail: addition of poly(A) in the nucleus. Trends Biochem. Sci. *15*, 277–281.

Yang, Q., Coseno, M., Gilmartin, G.M., and Doublié, 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.

Yu, K., Ganesan, K., Tan, L.K., Laban, M., Wu, J., Zhao, X.D., Li, H., Leung, C.H., Zhu, Y., Wei, C.L., et al. (2008). A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers. PLoS Genet. *4*, e1000129.

Zarkower, D., and Wickens, M. (1987). Specific pre-cleavage and post-cleavage complexes involved in the formation of SV40 late mRNA 3′ termini in vitro. EMBO J. *6*, 4185–4192.

Zhang, H., Hu, J., Recce, M., and Tian, B. (2005). PolyA_DB: a database for mammalian mRNA polyadenylation. Nucleic Acids Res. *33*(Database issue), D116–D120.